中国科学技术大学
University of Science and Technology of China

机器学习中的数学基础

顾言午

中国科学技术大学，大数据学院

2022 年 9 月 3 日

1 线性代数

2 多元微积分

3 概率论与数理统计

4 其他

- 向量、矩阵与张量
- 范数、距离

向量

$$\vec{a} = \begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix}^T$$

or

$$\boldsymbol{a} = \begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix}^T$$

矩阵

$$\boldsymbol{A} = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ 1 & x_3 & x_3^2 & x_3^3 \\ 1 & x_4 & x_4^2 & x_4^3 \end{bmatrix}$$

张量

$$\boldsymbol{A} = \begin{bmatrix} [0,0,0] & [256,256,0] \\ [0,256,256] & [256,256,256] \end{bmatrix}$$

向量范数:
正则化，防止过拟合

- L1 范数 $\|x\|_1 = \sum_{k=1}^{n} |x_k|$
- L2 范数 $\|x\|_2 = \sqrt{\sum_{k=1}^{n} x_k^2}$
- 无穷范数 $\|x\|_\infty = \max |x_k|$

矩阵范数:
防止过拟合

$$\|A\| = \max \frac{\|Ax\|}{\|x\|}, \|x\| \neq 0$$

- L1 范数 $\|A\|_1 = \max_j \sum_{i=1}^{m} |x_{ij}|$
- L2 范数 $\|A\|_2 = \max \lambda_i(A^H A)$
- 无穷范数 $\|A\|_1 = \max_i \sum_{j=1}^{n} |x_{ij}|$

距离：
验证算法效果

- 曼哈顿距离 $d = \sum_{k=1}^{n} |x_k - y_k|$
- 欧氏距离 $d = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$
- 闵可夫斯基距离 $d = \sqrt[p]{\sum_{k=1}^{n} (x_k - y_k)^p}$
- 切比雪夫距离 $d = \max(|x_k - y_k|)$
- 夹角余弦 $d = \frac{\sum_{k=1}^{n} x_k y_k}{\sqrt{\sum_{k=1}^{n} x_k^2} \sqrt{\sum_{k=1}^{n} y_k^2}}$
- 汉明距离，两串字符串中不同位数的数目

- 特征值分解
- LU 分解
- SVD 分解
- Moore-Penrose 伪逆

通过旋转变换，将矩阵的主要信息转化到对角线上，主成分分析(PCA)

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Delta}\boldsymbol{P}^{-1}$$

幂方法

L: 下三角矩阵，U: 上三角矩阵，便于求矩阵的逆，从而计算

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$

Cholesky 分解与 Doolitle 分解

奇异值

$$\boldsymbol{A}_{m\times n} = \boldsymbol{U}_{m\times m}\boldsymbol{\Sigma}_{m\times n}\boldsymbol{V}^T_{n\times n}$$

其中

$$\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}, \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$$

$\boldsymbol{\Sigma}$ 为对角元为 $\sigma_i(A)$ 的对角矩阵,$\sigma_i(A) = \lambda_i(A^TA)$
QR 分解

$$A = U\Sigma V^T$$

$$A^+ = V\Sigma^{-1}U^T$$

1 线性代数

2 多元微积分

- 基本概念
- 向量函数、方向梯度与海森矩阵
- 矩阵函数
- 一些优化方法
- 凸集和凸函数
- 线性规划问题
- 非线性优化问题
- KKT 条件
- 梯度下降法

3 概率论与数理统计

4 其他

1 线性代数

2 多元微积分

- 基本概念
- 向量函数、方向梯度与海森矩阵
- 矩阵函数
- 一些优化方法
- 凸集和凸函数
- 线性规划问题
- 非线性优化问题
- KKT 条件
- 梯度下降法

3 概率论与数理统计

4 其他

1 线性代数

2 多元微积分

- 基本概念
- 向量函数、方向梯度与海森矩阵
- 矩阵函数
- 一些优化方法
- 凸集和凸函数
- 线性规划问题
- 非线性优化问题
- KKT 条件
- 梯度下降法

3 概率论与数理统计

4 其他

假设我们有一个以向量为自变量的函数

$$f(\boldsymbol{x}) = f(x_1, x_2, ..., x_n)$$

那么

$$\begin{aligned} df &= \frac{\partial f}{\partial x_1} dx_1 + \frac{\partial f}{\partial x_2} dx_2 + ... + \frac{\partial f}{\partial x_n} dx_n \\ &= \begin{pmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \end{pmatrix} \begin{pmatrix} dx_1 \\ dx_2 \\ \vdots \\ dx_n \end{pmatrix} \end{aligned}$$

记

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \end{pmatrix}$$

记

$$\nabla^2 f = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial x_1 \partial^2 f}{\partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{pmatrix}$$

试试对 $f(\boldsymbol{x})$ 进行泰勒展开？

大纲

中国科学技术大学
University of Science and Technology of China

1 线性代数

2 多元微积分

- 基本概念
- 向量函数、方向梯度与海森矩阵
- 矩阵函数
- 一些优化方法
- 凸集和凸函数
- 线性规划问题
- 非线性优化问题
- KKT 条件
- 梯度下降法

3 概率论与数理统计

4 其他

- $\frac{\partial f(\boldsymbol{A})}{\partial \boldsymbol{A}} = (\frac{\partial f(\boldsymbol{A})}{\partial A_{ij}})_{ij}$
- $\frac{\partial \boldsymbol{A}(x)}{\partial x} = (\frac{\partial A_{ij}}{\partial x})_{ij}$
- 请注意，求导的链式法则仍然满足

我们来推导 $\frac{\partial A^{-1}}{\partial x} = -A^{-1}\frac{\partial A}{\partial x}$, $\frac{\partial \ln det(A)}{\partial A} = A^{-T}$

设 $a, \boldsymbol{a}, \boldsymbol{A}$ 均与 $x, \boldsymbol{x}$ 无关， u, v 均有关, f, u, v 可导

- $\frac{\partial a\boldsymbol{u}}{\partial x} = a\frac{\partial \boldsymbol{u}}{\partial x}$
- $\frac{\partial \boldsymbol{A}\boldsymbol{u}}{\partial x} = \frac{\partial \boldsymbol{u}}{\partial x}\boldsymbol{A}^T$
- $\frac{\partial \boldsymbol{u}^T}{\partial x} = (\frac{\partial \boldsymbol{u}}{\partial x})^T$
- $\frac{\partial f(\boldsymbol{u})}{\partial x} = \frac{\partial \boldsymbol{u}}{\partial x}\frac{\partial f(\boldsymbol{u})}{\partial \boldsymbol{u}}$

设 $a, \boldsymbol{a}, \boldsymbol{A}$ 均与 $x, \boldsymbol{x}$ 无关，$f(\boldsymbol{u}), \boldsymbol{u}(\boldsymbol{x}), \boldsymbol{v}(\boldsymbol{x})$ 可导

- $\frac{\partial \boldsymbol{u}^T \boldsymbol{v}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} \boldsymbol{v} + \frac{\partial \boldsymbol{v}}{\partial \boldsymbol{x}} \boldsymbol{u}$
- $\frac{\partial \boldsymbol{A}\boldsymbol{u}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} A^T$
- $\frac{\partial \boldsymbol{x}^T \boldsymbol{A}}{\partial \boldsymbol{x}} = A$
- $\frac{\partial f(\boldsymbol{u})}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} \frac{\partial f(\boldsymbol{u})}{\partial \boldsymbol{u}}$
- $\frac{\partial \boldsymbol{x}^T \boldsymbol{x}}{\partial \boldsymbol{x}} = 2\boldsymbol{x}$

尝试计算

$$\frac{\partial \boldsymbol{u}^T \boldsymbol{A} \boldsymbol{v}}{\partial \boldsymbol{x}}, \frac{\partial \boldsymbol{a}\boldsymbol{x}\boldsymbol{x}^T \boldsymbol{b}}{\partial \boldsymbol{x}}$$

设 $a, \boldsymbol{a}, \boldsymbol{A}$ 均与 $x, \boldsymbol{X}$ 无关，$f(\boldsymbol{u}), \boldsymbol{u}(\boldsymbol{X}), \boldsymbol{v}(\boldsymbol{X})$ 可导

- $\frac{\partial \boldsymbol{u}^T \boldsymbol{v}}{\partial \boldsymbol{X}} = \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{X}} \boldsymbol{v} + \frac{\partial \boldsymbol{v}}{\partial \boldsymbol{X}} \boldsymbol{u}$
- $\frac{\partial f(\boldsymbol{u})}{\partial \boldsymbol{X}} = \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{X}} \frac{\partial f(\boldsymbol{u})}{\partial \boldsymbol{u}}$
- $\frac{\partial \boldsymbol{a}^T \boldsymbol{X} \boldsymbol{b}}{\partial \boldsymbol{X}} = \boldsymbol{a}\boldsymbol{b}^T$

尝试计算

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{b}}{\partial \boldsymbol{X}}$$

1 线性代数

2 多元微积分

基本概念 ■ 向量函数、方向梯度与海森矩阵 ■ 矩阵函数 ■ 一些优化方法 ▪ 凸集和凸函数

- 线性规划问题
- 非线性优化问题
- KKT 条件
- 梯度下降法

3 概率论与数理统计

4 其他

凸集：给定集合 $C \subseteq \mathbb{R}^n$. 若 $\forall x, y \in C$ 满足

$$\forall t \in (0,1), tx + (1-t)y \in C$$

那么集合 C 为凸集

凸函数：给定一个函数 $f: \mathbb{R}^n \mapsto R$. 如果满足 $dom(f)$ 是凸集而且 $\forall x, y \subseteq dom(f)$,

$$\forall t \subset [0,1], f(tx + (1-t)y) \leq tf(x) + \ 1-t)f(y)$$

那么函数 f 是凸函数

一阶条件：假设函数 f 可微，那么 f 是凸函数当且仅当 $\forall x, y \in dom(f)$,

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

二阶条件：假设函数 f 二阶可微，那么 f 是凸函数当且仅当 $\forall x \in dom(f)$

$$\nabla^2 f(x) \succcurlyeq 0,$$

即海森矩阵半正定

Thm: 假设函数 f 可微凸函数，那么 x 是 f 的全局最优当且仅当

$$\nabla f(x) = 0$$

$$\begin{aligned} \min \quad & c^T x \\ s.t. \quad & A_e x_e = b_e \\ & A_i x_i \leq b_i \end{aligned}$$

可基于单纯形法或对偶问题求解

$$\begin{aligned} \min \quad & f(x) \\ s.t. \quad & g(x) = 0 \end{aligned}$$

转化为考虑

$$\min \quad f(x) + \lambda g(x)$$

，其中 λ 可以为任意值. 可以直接求导

$$\begin{aligned} \min \quad & c^T x \\ s.t. \quad & g_i(x) \geq 0, i = 1, ..., m \\ & h_i(x) = 0, i = 1, ..., l \end{aligned}$$

Kuhn-Tucker 条件:
设 $\bar{x}$ 为约束问题 (67) 的可行点,f 和 $g_i, i \in \mathcal{I}(\bar{x})$ 在点 $\bar{x}$ 可微,$g_i, i \notin \mathcal{I}(\bar{x})$ 在点 $\bar{x}$ 连续, h_j 在点

$\bar{x}$ 连续可微, 向量集 $\{\nabla g_i(\bar{x}), i \in \mathcal{I}(\bar{x}); \nabla h_j(\bar{x}), j = 1, ..., l\}$ 线性无关. 如果 $\bar{x}$ 是局部最优解, 则存在数 $\lambda_i \geq 0$ 和 μ_j 使得

$$\lambda_0 \nabla f(\bar{x}) - \sum_{i \in \mathcal{I}(\bar{x})} \lambda_i \nabla g_i(\bar{x}) - \sum_{j=1}^{l} \mu_j \nabla h_j(\bar{x}) = 0 \tag{82}$$

定义 Lagrange 函数 $L(x, \lambda, \mu) = f(x) - \sum_{i=1}^{m} \lambda_i g_i(x) - \sum_{j=1}^{l} \mu_j h_j(x)$.
若 $\bar{x}$ 为问题局部最优解, 则存在乘子向量 $\bar{\lambda} \geq 0, \bar{\mu}$ 使得

$$\nabla_x L(\bar{x}, \bar{\lambda}, \bar{\mu}) = 0.$$

此时, 一阶必要条件可表达为

$$(K-T) \begin{cases} \nabla_x L(x, \lambda, \mu) = 0 \\ g_i(x) \geq 0, i = 1, ..., m \\ \lambda_i g_i(x) = 0, i = 1, ..., m \\ \lambda_i \geq 0, i = 1, ..., m \\ h_j(x) = 0, j = 1, ..., l \end{cases} \tag{83}$$

$$
\begin{aligned}
f(x) &= f(x_0) + f'(x_0)(x - x_0) \\
0 &= f(x_0) + f'(x_0)(x - x_0) \\
x &= x_0 - \frac{f(x_0)}{f'(x_0)}
\end{aligned}
$$

$$
\begin{aligned}
f(\boldsymbol{x}) &= f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) \\
0 &= f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) \\
x &= x_0 - \frac{f(\boldsymbol{x}_0)}{\nabla f(\boldsymbol{x}_0)}
\end{aligned}
$$

$$f(x^{(k)}+s) \approx f(x^{(k)}) + g^{(k)T}s + \frac{1}{2}s^T G_k s$$

$$g^{(k)} = \nabla f(x^{(k)}), G_k = \nabla^2 f(x^{(k)})$$

$$\hat{s} = -G_k^{-1} g^{(k)}$$

1 线性代数

2 多元微积分
- 凸集和凸函数
- 线性规划问题
- 非线性优化问题
- KKT 条件
- 梯度下降法

3 概率论与数理统计

■ 组合、概率规则和公理 ■ 期望与方差 ■ 分布 ■ 贝叶斯公式、先验与后验 ■ 最大似然估计和最大后验估计

4 其他

1 线性代数

2 多元微积分
- 凸集和凸函数
- 线性规划问题
- 非线性优化问题
- KKT 条件
- 梯度下降法

3 概率论与数理统计

■ 组合、概率规则和公理 ■ 期望与方差 ■ 分布 ■ 贝叶斯公式、先验与后验 ■ 最大似然估计和最大后验估计

4 其他

请自主复习概率论与数理统计相关知识
大数定律是机器学习的基础

1 线性代数

2 多元微积分
- 凸集和凸函数
- 线性规划问题
- 非线性优化问题
- KKT 条件
- 梯度下降法

3 概率论与数理统计

组合、概率规则和公理 ■ **期望与方差** ■ 分布 ■ 贝叶斯公式、先验与后验 ■ 最大似然估计和最大后验估计

4 其他

请自主复习概率论与数理统计相关知识
理解二者不可兼得

1 线性代数

2 多元微积分
- 凸集和凸函数
- 线性规划问题
- 非线性优化问题
- KKT 条件
- 梯度下降法

3 概率论与数理统计

组合、概率规则和公理 ■ 期望与方差 ■ 分布 ■ 贝叶斯公式、先验与后验 ■ 最大似然估计和最大后验估计

4 其他

机器学习中比较重要的分布：

- 0-1 分布
- 几何分布
- 二项分布
- (多元) 高斯 (正态) 分布
- 指数分布
- 泊松分布
- 伽玛分布
- 贝塔分布
- 迪利克雷分布

1 线性代数

2 多元微积分
- 凸集和凸函数
- 线性规划问题
- 非线性优化问题
- KKT 条件
- 梯度下降法

3 概率论与数理统计

组合、概率规则和公理 期望与方差 分布 贝叶斯公式、先验与后验 最大似然估计和最大后验估计

4 其他

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)}$$

or

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{\int_{\mathbb{R}} f_{Y|X}(y|u)f_X(u)du}$$

先验：在考虑实验之前，我们首先通过经验给出参数的一个分布
后验：结合先验分布和实验数据，更新我们对先验分布的认知

一个例子

假设我们的观测值 x 服从关于 θ 的二项分布，
$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, x = 0, 1, ..., n$
我们有先验知识，θ 服从参数为 α, β 的贝塔分布
$\pi(\alpha, \beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, 0 \leq \theta \leq 1$
如果我们观测到了一个值 x，那么 y 应该服从什么分布？

一个例子

1 线性代数

2 多元微积分
- 凸集和凸函数
- 线性规划问题
- 非线性优化问题
- KKT 条件
- 梯度下降法

3 概率论与数理统计

组合、概率规则和公理 ■ 期望与方差 ■ 分布 ■ 贝叶斯公式、先验与后验 ■ 最大似然估计和最大后验估计

4 其他

极大似然估计的核心思想是：认为当前发生的事件是概率最大的事件。因此就可以给定的数据集，使得该数据集发生的概率最大来求得模型中的参数。

$$L(\theta) = \prod_{i=1}^{n} P(X_i|\theta), \theta = \arg\max_{\theta} L(\theta)$$

极大后验估计的核心思想是：允许引入参数的先验分布

$$\theta = \arg\max_{\theta} P(\theta|X) = \arg\max_{\theta} \frac{P(X|\theta)P(\theta)}{P(X)}$$

$$= \arg\max_{\theta} P(X|\theta)P(\theta) = \arg\max_{\theta} L(\theta)P(\theta)$$

假设我们对 n 个观测点 x_i 进行观测得到结果 y_i, 且 $y \sim N(w^T x, \sigma^2)$, 试通过 MLE 和 MAP 去计算 $\hat{w}$.

一个例子

见 https://zhuanlan.zhihu.com/p/86009986

给出信息熵的公式

$$H(X) = -\sum_{i=1}^{n} p(x_i) log(p(x_i))$$

信息熵 H 作为对随机实验不确定程度的度量，满足三个规则：

- H 是 p 的连续函数；
- 对于等概结果为 n 的随机实验，H 是 n 的单调递增函数；
- 组合可加性

$$\begin{aligned} H_n(p_1, p_2, ..., p_n) =& H_{n-1}(p_1 + p_2, p_3, ..., p_n) \\ &+ (p_1 + p_2) H_2(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}) \end{aligned}$$

假设现在有一个样本集中两个概率分布 p, q，其中 p 为真实分布，q 为非真实分布。假如，按照真实分布 p 来衡量识别一个样本所需要的编码长度的期望即位信息熵 $-\sum_{i=1}^{n} p(x_i) log(p(x_i))$.
如果采用错误的分布 q 来表示来自真实分布 p 的平均编码长度，则应该是交叉熵：

$$H(p, q) = -\sum_{i=1}^{n} p(x_i) log(q(x_i))$$

KL 散度公式为:

$$D(p\|q) = \sum_{i=1}^{n} p(x_i) log(\frac{p(x_i)}{q(x_i)})$$

- 不对称性
- 非负性

谢谢!

HW1 & HW2 Reference

HW1

1. Calculate $\frac{\partial \ln\det(\mathbf{A})}{\partial x}$

Solve: 在我们的课程中，我们已经证明了

$$\frac{\partial \ln\det(\mathbf{A})}{\partial \mathbf{A}} = \mathbf{A}^{-T}$$

右边的矩阵代表当 A_{ij} 变动时，$\ln\det(\mathbf{A})$ 将会变动 $(\mathbf{A}^{-T})_{ij} \cdot d\mathbf{A}_{ij}$.

我们又知道

$$\frac{\partial \mathbf{A}}{\partial x}$$

也是一个 $i \times j$ 的矩阵，当 x 变动时，$\mathbf{A}_{ij}$ 将变动 $(\frac{\partial \mathbf{A}}{\partial x})_{ij} \cdot dx$

综上，$\frac{\partial \ln\det(\mathbf{A})}{\partial x}$ 为以下矩阵的元素和

$$\mathbf{A}^{-T} * \frac{\partial \mathbf{A}}{\partial x}$$

其中 $*$ 代表 Hadamard Product，即两个矩阵对应元素乘积

因此最终结果可以表述为（这里可以）

$$\frac{\partial \ln\det(\mathbf{A})}{\partial x} = tr(\mathbf{A}^{-1} \cdot \frac{\partial \mathbf{A}}{\partial x})$$

另外，永停姐姐提供了一个证法，利用等式 $\ln(\det(A)) = tr(\ln(A))$。易得

$$\frac{\partial \ln\det(\mathbf{A})}{\partial x} = \frac{\partial tr(\ln(\mathbf{A}))}{\partial x} = tr(\frac{\partial \ln(\mathbf{A})}{\partial x}) = tr(\mathbf{A}^{-1} \cdot \frac{\partial \mathbf{A}}{\partial x})$$

其中 $\ln \mathbf{A}$ 满足 $\exp(\ln \mathbf{A}) = \mathbf{A}, \exp(\mathbf{M}) = \mathbf{I} + \mathbf{M} + \frac{1}{2}\mathbf{M}^2 \dots$

2. 书习题1.2

Solve: 略，言之有理即可

3. 已知随机变量 $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2] \sim \mathcal{N}(\mu, \Sigma)$,计算 $P(\mathbf{x}_1), P(\mathbf{x}_1|\mathbf{x}_2)$

Solve: 我们先假设 $\mathbf{x} \in \mathbb{R}^n, \mathbf{x}_1 \in \mathbb{R}^{n_1}, \mathbf{x}_2 \in \mathbb{R}^{n_2}, n_1 + n_2 = n, \Sigma \in \mathbb{R}^{n \times n}, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix}, \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$

由于 Σ 正定，顺序主子式 $|\Sigma_{11}| > 0$, 因此 Σ_{11} 可逆，我们对 Σ 进行分解

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} I_{n_1} & 0 \\ \Sigma_{12}^T\Sigma_{11}^{-1} & I_{n_2} \end{pmatrix} \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12} \end{pmatrix} \begin{pmatrix} I_{n_1} & \Sigma_{11}^{-1}\Sigma_{12} \\ 0 & I_{n_2} \end{pmatrix}$$

取逆得到

$$\Sigma^{-1}=\begin{pmatrix} I_{n_1} & -\Sigma_{11}^{-1}\Sigma_{12} \\ 0 & I_{n_2} \end{pmatrix}\begin{pmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & (\Sigma_{22}-\Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12})^{-1} \end{pmatrix}\begin{pmatrix} I_{n_1} & 0 \\ -\Sigma_{12}^T\Sigma_{11}^{-1} & I_{n_2} \end{pmatrix}$$

记 $\Sigma_{22}-\Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12}=\Sigma_*$

写出概率密度函数

$$f(x)=\frac{1}{(\sqrt{2\pi})^n\det(\Sigma)}\exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

对于指数部分，我们可以得到

$$\begin{aligned}(x-\mu)^T\Sigma^{-1}(x-\mu)=&(x_1-\mu_1)^T\Sigma_{11}^{-1}(x_1-\mu_1)\\&+[(x_2-\mu_2)-\Sigma_{12}^T\Sigma_{11}^{-1}(x_1-\mu_1)]^T\Sigma_*^{-1}[(x_2-\mu_2)-\Sigma_{12}^T\Sigma_{11}^{-T}(x_1-\mu_1)]\end{aligned}$$

当我们求 $P(x_1)$ 时候，实际是对 $x_2\in\mathbb{R}^{n_2}$ 求积分，由于 x_1 不变，可以认为求了一个 $\mathcal{N}(\mu_2+\Sigma_{12}^T\Sigma_{11}^{-1}(x_1-\mu_2),\Sigma_*)$ 的积分，可以求得对含 x_2 的指数部分求积分结果为 $|\Sigma|/|\Sigma_{11}|$, 也可以分析得出

$$x_1\sim\mathcal{N}(\mu_1,\Sigma_{11}).$$

当计算条件分布的时候，我们利用公式

$$P(x_1|x_2)=\frac{P(x_1,x_2)}{P(x_2)}$$

再利用之前得到的结果(注意需要将 Σ^{-1} 对于 Σ_{22} 分解)，可以得到结果

$$\begin{aligned}P(x_1|x_2)=\sqrt{\frac{|\Sigma_{22}|}{(2\pi)^{n_2}|\Sigma|}}\exp(-\frac{1}{2}([(x_1-\mu_1)^T-(x_2-\mu_2)^T\Sigma_{22}^{-1}\Sigma_{12}^T](\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T)^{-1}\\[(x_1-\mu_1)-\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2)])\end{aligned}$$

4. 证明 $\|x\|_p$ 是凸函数

   一个说明，$p<1,p\neq1$ 时，不能满足向量范数要求的三角不等式，因此不能算范数，我们只需考虑 $p>1$ 的情况. ($p=0,p=1$ 显然)

Proof. 对于 $\forall t\in(0,1),u,v\in\mathbb{R}^n$, 有

$$\|tu+(1-t)v\|_p\leq\|tu\|_p+\|(1-t)v\|_p=t\|u\|_p+(1-t)\|v\|_p$$

其中用到了向量范数必须满足的三角不等式和正齐次性

5. 证明判定凸函数的0阶和1阶条件相互等价

Proof. 充分性：设满足 $\forall t\in[0,1],f(tx+(1-t)y)\leq tf(x)+(1-t)f(y)$

令 $g(t)=tf(x)+(1-t)f(y)-f(tx+(1-t)y)$, 显然 $g(0)=0,g(t)\geq0,t\in[0,1]$. 易得 $\lim_{t\to1-}g'(t)\leq0$

而

$$g'(t) = f(x) - f(y) - \nabla f(tx + (1-t)y)^T (x - y)$$
$$\lim_{t \to 1-} g'(t) = f(x) - f(y) - \nabla f(x)^T (x - y) \leq 0$$
$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

必要性：设满足 $f(y) \geq f(x) + \nabla f(x)^T (y - x)$. 易得 $\lim_{t \to 0+} g'(t) \geq 0, \lim_{t \to 1-} \leq 0$. 倘若我们可以证明 $g''(t) \leq 0$, 0阶条件得证.

$$g''(t) = -(x - y)^T \nabla^2 f(tx + (1-t)y)(x - y)$$

只需证明 ∇f 是正定的即可. 而对于任意向量 x, y

$$f(y) - f(x) - \nabla f(x)^T (y - x) = (y - x)^T \nabla^2 f(x)(y - x) + O(\|y - x\|^3) \geq 0$$

当 $\|y - x\| \to 0$ 时，可以发现 $(y - x)^T \nabla^2 f(x)(y - x) \geq 0$, 由于 x, y 任意取，$\nabla^2 f$ 正定，证毕

HW2

1. 习题2.2

Solve：

10折交叉验证要求我们保证每个子集尽可能保持数据分布的一致性，即正反例数量相同，所以最终只会判断为正确/错误（或者认为50%随机选择），错误率期望为50%

留一法的训练集中数量较多的 label 必然不是留出的样本的 label，必定预测错误，错误率为100%

2. 习题2.4

Solve：

回顾混淆矩阵 confusion matrix

	预测为正	预测为反
真实为正	TP	FN
真实为反	FP	TN

真假：对于预测的准确性而言，正反：对于预测的结果而言

真正例率：$TPR = \frac{TP}{TP + FN}$, 假正例率：$FPR = \frac{FP}{FP + FN}$

查准率：$R = \frac{TP}{TP + FN}$, 查全率：$P = \frac{TP}{TP + FP}$

就数值而言，$TPR = R$, 其他没有直接的数值关系

3. 习题2.5

Solve：

先回顾 ROC 曲线。我们不妨将 ROC 曲线的横坐标扩大 m^- 倍，纵坐标扩大 m^+ 倍，这样绘制 ROC 曲线图时每一步都走一个单位长度，方便说明.

Step 1. 用分类器对所有数据分类，得到结果为一个 $[0,1]$ 的值，值越大说明越容易被判定为正.

Step 2. 将所有数据按预测结果降序排列

Step 3. 从最大预测结果开始，如果实际为真，向上走一步，如果实际为假，向右走一步

Step 4. 如果有若干样本预测结果相同，先同时向右、上走相应的步数，将起点终点直接相连

下证明 l_{rank} 对应 ROC 曲线上的面积。

对于每一个反样本（即向右走），我们假设没有样本预测结果与之相同。在正样本中：比其预测结果大的，已经在其之前绘制（即已经向上走过了），在 ROC 曲线中表现为曲线之下的部分；反之，比其预测结果小的，还未绘制，且终将在其之后绘制，在 ROC 曲线中表现为曲线之上的部分。曲线之上的部分对应 $\sum_{x^+}\mathbb{I}(f(x^+)<f(x^-))$

假如有 p,q 个正、反样本预测结果相同，除了 $\mathbb{I}(f(x^+)<f(x^-))$ 部分，这对应的是 ROC 曲线纵向方向上对应的矩形，我们还需要计算起点终点直接相连的三角形部分 $\frac{1}{2}pq$. 在 loss 中对应为 $\sum_{x^+}\sum_{x^-}\frac{1}{2}\mathbb{I}(f(x^+)=f(x^-))$.

只需对应缩放，即得到我们的证明。

4. 习题2.9

Solve：

Step 1. 提出原假设和备择假设，离散情况下 $H_0:P(X=x_i)=p_i,i=1,2,\ldots,$ 连续情况下 $H_0:X\sim F(x)$

Step 2. 将 X 的取值范围划分为 k 个互不相交的子区间 $A_1,\ldots,A_k,$ 按照习惯，$k\geq 5$.

Step 3. 记落入 A_i 区间的样本个数为 $n_i,\sum_{i=1}^k n_i=n$

Step 4. 记随机样本落入 A_i 区间的概率为 q_i

Step 5. 计算

$$\chi^2=\sum_{i=1}^k\frac{(n_i-nq_i)^2}{nq_i},\chi^2\sim\chi^2_{k-1}$$

HW3 & HW4 Reference

注意：方法不唯一，言之成理即可！

0.1 [课本习题 3.2] 试证明，对于参数 w，对率回归的目标函数 (3.18) 是非凸的，但其对数似然函数 (3.27) 是凸的。

$$y = \frac{1}{1 + \mathrm{e}^{-\boldsymbol{w}^\top \boldsymbol{x} + b}} \tag{3.18}$$

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m} (-y_i \boldsymbol{\beta}^\top \hat{\boldsymbol{x}}_i + \ln(1 + e^{\boldsymbol{\beta}^\top \hat{\boldsymbol{x}}_i})) \tag{3.27}$$

证明.

- 需要注意，标量函数先对向量变量的转置求导，再对向量变量求导，得到的是矩阵；标量函数先对向量变量求导，再对向量变量的转置求导，得到的是标量。很多同学混淆使用。《矩阵分析与应用（第 2 版）》张贤达
- 方法不唯一，但是需要注意符号书写清晰

$$\frac{\partial y}{\partial \boldsymbol{w}} = \frac{\boldsymbol{x} e^{-\left(\boldsymbol{w}^\top \boldsymbol{x} + b\right)}}{\left(1 + e^{-\left(\boldsymbol{w}^\top \boldsymbol{x} + b\right)}\right)^2} = \boldsymbol{x} y (1 - y)$$

$$\frac{\partial^2 y}{\partial \boldsymbol{w}^\top \partial \boldsymbol{w}} = \frac{\partial}{\partial \boldsymbol{w}^\top} \frac{\partial y}{\partial \boldsymbol{w}} = \frac{\partial y}{\partial \boldsymbol{w}^\top} \boldsymbol{x}(1 - y) - \frac{\partial y}{\partial \boldsymbol{w}^\top} \boldsymbol{x} y = \boldsymbol{x}^\top \boldsymbol{x} y (1 - 2y)(1 - y)$$

$\boldsymbol{x}^\top \boldsymbol{x} \geq 0$ 恒成立，当 $0.5 < y < 1$ 时，$y(1 - 2y)(1 - y) < 0$，此时 $\frac{\partial^2 y}{\partial \boldsymbol{w}^\top \partial \boldsymbol{w}} < 0$，因此函数 (3.18) 非凸。

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{m} \left(-y_i \hat{\boldsymbol{x}}_i + \frac{1}{1 + e^{\boldsymbol{\beta}^\top \hat{\boldsymbol{x}}_i}} \hat{\boldsymbol{x}}_i e^{\boldsymbol{\beta}^\top \hat{\boldsymbol{x}}_i} \right)$$

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}^\top \partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}^\top} \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}^\top} \sum_{i=1}^{m} \left(-y_i \hat{\boldsymbol{x}}_i + \frac{1}{1 + e^{\boldsymbol{\beta}^\top \hat{\boldsymbol{x}}_i}} \hat{\boldsymbol{x}}_i e^{\boldsymbol{\beta}^\top \hat{\boldsymbol{x}}_i} \right) = \sum_{i=1}^{m} \frac{e^{\boldsymbol{\beta}^\top \hat{\boldsymbol{x}}_i}}{\left(1 + e^{\boldsymbol{\beta}^\top \hat{\boldsymbol{x}}_i}\right)^2} \hat{\boldsymbol{x}}_i{}^\top \hat{\boldsymbol{x}}_i$$

由于 $\hat{\boldsymbol{x}}_i{}^\top \hat{\boldsymbol{x}}_i \geq 0$ 且 $\frac{e^{\boldsymbol{\beta}^\top \hat{\boldsymbol{x}}_i}}{\left(1 + e^{\boldsymbol{\beta}^\top \hat{\boldsymbol{x}}_i}\right)^2} \geq 0$，因此函数 (3.27) 为凸函数。

□

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9
C_1	1	1	1	1	1	1	1	x	x
C_2	0	0	0	0	1	1	1	x	x
C_3	0	0	1	1	0	0	1	x	x
C_4	0	1	0	1	0	1	0	x	x

0.2 [课本习题 3.7] 令码长为 9，类别数为 4，试给出海明距离意义下理论最优的 ECOC 二元码并证明之。

解.
- 任意两个类别间的海明距离足够大
- 任意两个分类器的输出应该尽量独立
- 最优定义 1: 任意两个编码之间的最小距离最大
- 最优定义 2: 任意两个类别之间的海明距离要大，并且任意两个类别间反码的距离最大
- 最优定义 3: 编码间距离和最大

(构造方法以及相关证明可参考文章“Solving Multiclass Learning Problems via Error-Correcting Output Codes”)

采用“Exhaustive Codes”方法构造。当类别为 4，可行编码有 7 种（即 $f_1 - f_7$），f_7 后的任意编码都是之前编码的反码, 因此 f_8、f_9 可以为任意编码。此时，类别间最小海明距离为 4。

□

0.3 在 LDA 多分类情形下，试计算类间散度矩阵 S_b 的秩，并证明

解. 令

$$\boldsymbol{S}_b = \sum_c m_c \left(\boldsymbol{\mu}_c - \boldsymbol{\mu}\right)\left(\boldsymbol{\mu}_c - \boldsymbol{\mu}\right)^\top = \boldsymbol{A}\boldsymbol{M}\boldsymbol{A}^\top$$

其中

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{\mu}_1 - \boldsymbol{\mu} & \boldsymbol{\mu}_2 - \boldsymbol{\mu} & \cdots & \boldsymbol{\mu}_N - \boldsymbol{\mu} \end{pmatrix}, \quad \boldsymbol{M} = \operatorname{diag}\left(m_1, m_2, \cdots, m_N\right).$$

接着，可以得到

$$\operatorname{rank} \boldsymbol{S}_b = \operatorname{rank}\left(\boldsymbol{A}\boldsymbol{M}\boldsymbol{A}^\top\right) = \operatorname{rank}\left(\left(\boldsymbol{A}\boldsymbol{M}^{\frac{1}{2}}\right)\left(\boldsymbol{A}\boldsymbol{M}^{\frac{1}{2}}\right)^\top\right) = \operatorname{rank}\left(\boldsymbol{A}\boldsymbol{M}^{\frac{1}{2}}\right) = \operatorname{rank} \boldsymbol{A}$$

因为 $\sum_c m_i(\boldsymbol{\mu}_i - \boldsymbol{\mu}) = \mathbf{0}$，所以 $\operatorname{rank} \boldsymbol{S}_b = \operatorname{rank} \boldsymbol{A} \leq N - 1$.

□

0.4 给出公式 (3.45) 的推导公式。

$$\mathbf{S}_b\mathbf{W} = \lambda\mathbf{S}_w\mathbf{W} \tag{3.45}$$

解. 问题:

$$\max_{\mathbf{W}} \frac{\operatorname{tr}\left(\mathbf{W}^\top\mathbf{S}_b\mathbf{W}\right)}{\operatorname{tr}\left(\mathbf{W}^\top\mathbf{S}_w\mathbf{W}\right)} \tag{3.44}$$

令上式分母为 1，则可上述问题转化为

$$\begin{aligned} \min_{\mathbf{W}} \quad & -\operatorname{tr}\left(\mathbf{W}^\top \mathbf{S}_b \mathbf{W}\right) \\ \text{s.t.} \quad & \operatorname{tr}\left(\mathbf{W}^\top \mathbf{S}_w \mathbf{W}\right) = 1 \end{aligned} \tag{1}$$

引入拉格朗日乘子法：

$$L(\mathbf{W}, \lambda) = -\operatorname{tr}\left(\mathbf{W}^\top \mathbf{S}_b \mathbf{W}\right) + \lambda\left(\operatorname{tr}\left(\mathbf{W}^\top \mathbf{S}_w \mathbf{W}\right) - 1\right) \tag{2}$$

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{W}} &= -\frac{\partial \operatorname{tr}\left(\mathbf{W}^\top \mathbf{S}_b \mathbf{W}\right)}{\partial \mathbf{W}} + \lambda \frac{\partial \operatorname{tr}\left(\mathbf{W}^\top \mathbf{S}_w \mathbf{W}\right)}{\partial \mathbf{W}} \\ &= -\left(\mathbf{S}_b + \mathbf{S}_b^\top\right)\mathbf{W} + \lambda\left(\mathbf{S}_w + \mathbf{S}_w^\top\right)\mathbf{W} \\ &= -2\mathbf{S}_b\mathbf{W} + 2\lambda\mathbf{S}_w\mathbf{W} \end{aligned} \tag{3}$$

令 $\frac{\partial L}{\partial \mathbf{W}} = 0$，即可得到公式 (3.45)。 □

0.5 证明 $X(X^\top X)^{-1}X^\top$ 是投影矩阵，并对线性回归模型从投影角度解释。

证明. 令 $P = X(X^\top X)^{-1}X^\top$，那么

$$P^\top = \left(X(X^\top X)^{-1}X^\top\right)^\top = X\left(X(X^\top X)^{-1}\right)^\top = X(X^\top X)^{-1}X^\top = P$$

因此 P 是一个对称矩阵，又因为

$$P^2 = X(X^\top X)^{-1}X^\top X(X^\top X)^{-1}X^\top = X(X^\top X)^{-1}(X^\top X)(X^\top X)^{-1}X^\top = X(X^\top X)^{-1}X^\top = P$$

因此 P 是一个幂等矩阵，所以 P 是一个投影矩阵。 □

解释. 线性回归模型：$\hat{\boldsymbol{y}} = X^\top(X^\top X)^{-1}X^\top \boldsymbol{y}$。可以发现，$\hat{\boldsymbol{y}}$ 其实是 $\boldsymbol{y}$ 在线性空间的投影。 □

1 HW4

1.1 [课本习题 4.1] 试证明对于不含冲突数据（即特征向量完全相同但标记不同）的训练集，必存在与训练集一致（即训练误差为 0）的决策树。

证明. (反证法) 假设不存在与训练集一致的决策树，那么训练集训练得到的决策树必然含有冲突数据，这与假设矛盾，因此必然存在与训练集一致决策树。 □

1.2 [课本习题 4.9] 试将 4.4.2 节对缺失值的处理机制推广到基尼指数的计算中去。

解.

$$\begin{aligned} \operatorname{Gini}(D, a) &= \rho \times \operatorname{Gini\_index}(\tilde{D}, a) \\ &= \rho \times \sum_{v=1}^{V} \tilde{r}_v \operatorname{Gini}(\tilde{D}^v) \\ &= \rho \times \sum_{v=1}^{V} \tilde{r}_v \left(1 - \sum_{i=1}^{k} \tilde{p}_k^2\right) \end{aligned}$$

□

1.3 **假设离散随机变量 $X \in \{1, \ldots, K\}$，其取值为 k 的概率 $P(X = k) = p_k$，其熵为 $H(p) = -\sum_k p_k \log_2 p_k$，试用拉格朗日乘子法证明熵最大分布为均匀分布。**

证明.

$$L(p, \lambda) = -\sum_{i=1}^{k} p_i \log_2 p_i + \lambda \left(\sum_{i=1}^{k} p_i - 1 \right)$$

$$\frac{\partial L}{\partial p_i} = -\log_2 p_i - \frac{1}{\ln 2} + \lambda = 0 \Longrightarrow p_1 = p_2 = \cdots = p_k = 2^{\lambda - \frac{1}{\ln 2}}$$

$$\frac{\partial L}{\partial \lambda} = \sum_{i=1}^{k} p_i - 1 = 0 \Longrightarrow p_1 = p_2 = \cdots = p_k = \frac{1}{k}$$

□

1.4 **下表表示的二分类数据集，具有三个属性 A、B、C，样本标记为两类“+”, “-”。请运用学过的知识完成如下问题：**

实例	A	B	C	类别
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-
10	F	F	2.0	+

1.4.1 整个训练样本关于类属性的熵是多少?

解. 类别 $+$ 的概率为 $p^+ = \frac{5}{10}$，类别 $-$ 的概率为 $p^- = \frac{5}{10}$，因此熵为

$$\mathrm{Ent}(D) = -p^+ \log_2 p^+ - p^- \log_2 p^- = 1.$$

□

1.4.2 数据集中 A、B 两个属性的信息增益各是多少?

解.

$$\begin{aligned}
\mathrm{Gain}(D, A) &= \mathrm{Ent}(D) - \sum_v \frac{|D^v|}{|D|} \mathrm{Ent}\left(D^v\right) \\
&= 1 - \left(\frac{4}{10} \left(-\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} \right) + \frac{6}{10} \left(-\frac{4}{6} \log_2 \frac{4}{6} - \frac{2}{6} \log_2 \frac{2}{6} \right) \right) \\
&= 0.125
\end{aligned}$$

$$
\begin{aligned}
\operatorname{Gain}(D, B) &= \operatorname{Ent}(D) - \sum_{v} \frac{|D^v|}{|D|} \operatorname{Ent}\left(D^v\right) \\
&= 1 - \left(\frac{5}{10}\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) + \frac{5}{10}\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right)\right) \\
&= 0.029
\end{aligned}
$$

□

1.4.3 对于属性 C，计算所有可能划分的信息增益?

解. 可取划分点为 $\{1.5, 2.5, 3.5, 4.5, 5.5, 6.5, 7.5\}$，然后对应划分信息增益为:

$$
\begin{aligned}
&\operatorname{Gain}(D, C, 1.5) = 1 - \tfrac{9}{10}\left(-\tfrac{4}{9}\log_2\tfrac{4}{9} - \tfrac{5}{9}\log_2\tfrac{5}{9}\right) \approx 0.108 \\
&\operatorname{Gain}(D, C, 2.5) = 1 - \tfrac{8}{10}\left(-\tfrac{3}{8}\log_2\tfrac{3}{8} - \tfrac{5}{8}\log_2\tfrac{5}{8}\right) \approx 0.236 \\
&\operatorname{Gain}(D, C, 3.5) = 1 - \left[\tfrac{3}{10}\left(-\tfrac{2}{3}\log_2\tfrac{2}{3} - \tfrac{1}{3}\log_2\tfrac{1}{3}\right) + \tfrac{7}{10}\left(-\tfrac{3}{7}\log_2\tfrac{3}{7} - \tfrac{4}{7}\log_2\tfrac{4}{7}\right)\right] \approx 0.035 \\
&\operatorname{Gain}(D, C, 4.5) = 1 - \left[\tfrac{4}{10}\left(-\tfrac{3}{4}\log_2\tfrac{3}{4} - \tfrac{1}{4}\log_2\tfrac{1}{4}\right) + \tfrac{6}{10}\left(-\tfrac{2}{6}\log_2\tfrac{2}{6} - \tfrac{4}{6}\log_2\tfrac{4}{6}\right)\right] \approx 0.125 \\
&\operatorname{Gain}(D, C, 5.5) = 1 - \left[\tfrac{6}{10}\left(-\tfrac{3}{6}\log_2\tfrac{3}{6} - \tfrac{3}{6}\log_2\tfrac{3}{6}\right) + \tfrac{4}{10}\left(-\tfrac{2}{4}\log_2\tfrac{2}{4} - \tfrac{2}{4}\log_2\tfrac{2}{4}\right)\right] = 0 \\
&\operatorname{Gain}(D, C, 6.5) = 1 - \left[\tfrac{7}{10}\left(-\tfrac{4}{7}\log_2\tfrac{4}{7} - \tfrac{3}{7}\log_2\tfrac{3}{7}\right) + \tfrac{3}{10}\left(-\tfrac{1}{3}\log_2\tfrac{1}{3} - \tfrac{2}{3}\log_2\tfrac{2}{3}\right)\right] \approx 0.035 \\
&\operatorname{Gain}(D, C, 7.5) = 1 - \tfrac{9}{10}\left(-\tfrac{5}{9}\log_2\tfrac{5}{9} - \tfrac{4}{9}\log_2\tfrac{4}{9}\right) \approx 0.108
\end{aligned}
$$

□

1.4.4 根据 Gini 指数，A 和 B 两个属性哪个是最优划分?

解.

$$
\begin{aligned}
\operatorname{Gini\_index}(D, A) &= \frac{4}{10}\left(1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2\right) + \frac{6}{10}\left(1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2\right) \\
&= 0.417 \\
\operatorname{Gini\_index}(D, B) &= \frac{5}{10}\left(1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2\right) + \frac{5}{10}\left(1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2\right) \\
&= 0.48
\end{aligned} \tag{4}
$$

因此，A 是最优划分。

□

1.4.5 采用算法 C4.5，构造决策树。

解.

- 指标：信息增益率，不是信息增益
- 构造方法和构造结果不唯一，建议大家构造前简述自己的构造方法，可以拿一半的过程分。

初始：$D = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$

第一层划分:

$$
\operatorname{Gain\_ratio}(D, A) = \frac{\operatorname{Gain}(D, A)}{\operatorname{IV}(D, A)} = \frac{0.125}{0.971} = 0.129
$$

$$\text{Gain\_ratio}(D, B) = \frac{\text{Gain}(D, B)}{\text{IV}(D, B)} = \frac{0.029}{1} = 0.029$$

$$\text{Gain\_ratio}(D, C, 2.5) = \frac{\text{Gain}(D, C, 2.5)}{\text{IV}(D, C, 2.5)} = \frac{0.236}{0.722} = 0.326$$

选择 $(C, 2.5)$ 作为 D 的划分，得到 $D^1_{c\leq 2.5} = \{1, 10\}$，$D^1_{c>2.5} = \{2, 3, 4, 5, 7, 8, 9\}$。

第二层划分: 因为 $D^1_{c\leq 2.5}$ 元素类别一致，不再划分，主要针对 $D^1_{c>2.5}$ 继续划分。

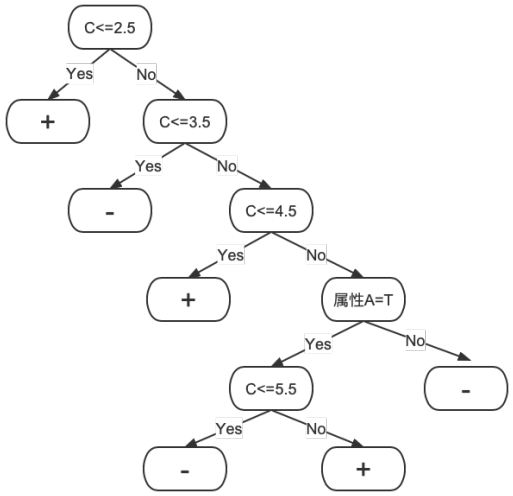
$$\text{Gain\_ratio}(D^1_{c>2.5}, A) = \frac{\text{Gain}(D^1_{c>2.5}, A)}{\text{IV}(D^1_{c>2.5}, A)} = \frac{0.159}{0.954} = 0.167$$

$$\text{Gain\_ratio}(D^1_{c>2.5}, B) = \frac{\text{Gain}(D^1_{c>2.5}, B)}{\text{IV}(D^1_{c>2.5}, B)} = \frac{0.049}{1} = 0.049$$

$$\text{Gain\_ratio}(D^1_{c>2.5}, C, 3.5) = \frac{\text{Gain}(D^1_{c>2.5}, C, 3.5)}{\text{IV}(D^1_{c>2.5}, C, 3.5)} = \frac{0.092}{0.544} = 0.169$$

选择 $(C, 3.5)$ 作为 $D^1_{c>2.5}$ 的划分，得到 $D^2_{2.5<c\leq 3.5} = \{6\}$，$D^2_{c\geq 3.5} = \{2, 3, 4, 5, 7, 8, 9\}$。

第三层划分: 以此类推...



□

HW5 and HW6 Reference

1 HW5

1.1 试述将线性函数$f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$用作神经元激活函数的缺陷

（言之有理即可）

解：当单元层和隐藏层激活函数为线性函数$f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$时，每一层输出都是上层输入的线性函数，无论神经网络有多少层，输出都是输入的线性组合，神经网络实际仍为原始的感知机；当输出层激活函数也为线性函数时，相当于整体的线性回归。此时的网络无法处理非线性问题。使用非线性激活函数增加了神经网络模型的非线性因素，使得神经网络可以任意逼近任何非线性函数，这样神经网络就可以应用到众多的非线性模型中。

1.2 讨论$\frac{e^{x_i}}{\sum_{j=1}^{C} e^{x_j}}$ 和$log\frac{e^{x_i}}{\sum_{j=1}^{C} e^{x_j}}$的数值溢出问题

解：实数在计算机中以二进制表示，计算时非精确值。当数值过小会取0（下溢出）或数值过大导致上溢出。对于softmax函数，当$x_i \to -\infty, i = 1, 2, \cdots, C$时，$\sum_{i=1}^{C} e^{x_i} \to 0$，分母计算可能四舍五入为0，发生下溢出。当$x_i \to +\infty$时，$e^{x_i} \to +\infty$，分子计算可能出现上溢出。

解决方法：

对于softmax函数，令$M = max(x_i), i = 1, 2, \cdots, C$，将计算$f(x_i)$改为计算$f(x_i - M)$即可。

说明如下：

$$\frac{e^{x_i}}{\sum_{j=1}^{C} e^{x_j}} = \frac{\frac{e^{x_i}}{e^M}}{\sum_{j=1}^{C} \frac{e^{x_j}}{e^M}} = \frac{e^{x_i - M}}{\sum_{j=1}^{C} e^{x_j - M}}$$

$e^{x_i - M} \leq e^{M-M} = 1$，分子不会发生上溢；

$\sum_{j=1}^{C} e^{x_j - M} \geq e^{M-M} = 1$，分母不会发生下溢。

对于log softmax函数，同上，

$$log\frac{e^{x_i}}{\sum_{j=1}^{C} e^{x_j}} = log\frac{\frac{e^{x_i}}{e^M}}{\sum_{j=1}^{C} \frac{e^{x_j}}{e^M}} = log\frac{e^{x_i - M}}{\sum_{j=1}^{C} e^{x_j - M}} = x_i - M - log\sum_{j=1}^{C} e^{x_j - M}$$

$1 = e^{M-M} \leq \sum_{j=1}^{C} e^{x_j - M} \leq \sum_{j=1}^{C} e^{M-M} = C$，所以分子不会发生上溢,分母不会发生下溢。

1.3 计算$\frac{e^{x_i}}{\sum_{j=1}^{C} e^{x_j}}$ 和$log\frac{e^{x_i}}{\sum_{j=1}^{C} e^{x_j}}$关于向量$\boldsymbol{x} = [x_1, \cdots, x_C]$的梯度

（计算对向量梯度需分别计算对每个分量的梯度，计算结果为向量形式）

解：令$f(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{C} e^{x_j}}, g(x_i) = logf(x_i), i, k = 1, 2, \cdots, C$

$k \neq i$时，

$$\frac{\partial f(x_i)}{\partial x_k} = -\frac{e^{x_i+x_k}}{(\sum_{j=1}^{C} e^{x_j})^2}$$

$k = i$时，

$$\frac{\partial f(x_i)}{\partial x_k} = \frac{e^{x_i}\sum_{m=1,m\neq i}^{C} e^{x_m}}{(\sum_{j=1}^{C} e^{x_j})^2}$$

所以

$$\begin{aligned}\frac{\partial f(x_i)}{\partial \boldsymbol{x}} &= \frac{e^{x_i}}{(\sum_{j=1}^{C} e^{x_j})^2}[-e^{x_1}, \cdots, -e^{x_{i-1}}, \sum_{m=1,m\neq i}^{C} e^{x_m}, \cdots, -e^{x_C}] \\ &= \frac{f(x_i)}{\sum_{j=1}^{C} e^{x_j}}[-e^{x_1}, \cdots, -e^{x_{i-1}}, \sum_{m=1,m\neq i}^{C} e^{x_m}, \cdots, -e^{x_C}]\end{aligned}$$

同理可得，

$$\begin{aligned}\frac{\partial g(x_i)}{\partial \boldsymbol{x}} &= \frac{1}{\sum_{j=1}^{C} e^{x_j}}[-e^{x_1}, \cdots, -e^{x_{i-1}}, \sum_{m=1,m\neq i}^{C} e^{x_m}, \cdots, -e^{x_C}] \\ &= \frac{f(x_i)}{e^{x_i}}[-e^{x_1}, \cdots, -e^{x_{i-1}}, \sum_{m=1,m\neq i}^{C} e^{x_m}, \cdots, -e^{x_C}]\end{aligned}$$

1.4 考虑如下简单网络,假设激活函数为ReLU ,用平方损失$\frac{1}{2}(y-\hat{y})^2$计算误差，请用BP算法更新一次所有参数 (学习率为1)，给出更新后的参数值(给出详细计算过程)，并计算给定输入值$\boldsymbol{x} = (0.2, 0.3)$时初始时和更新后的输出值,检查参数更新是否降低了平方损失值.

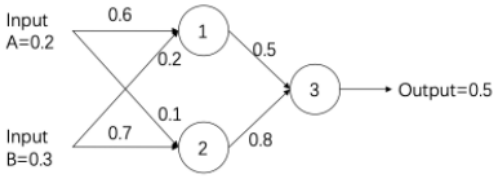


Figure 1: 简单网络

（注意题目中BP算法使用的激活函数为ReLU函数以及链式求导计算梯度）

解：

$$ReLU'(x) = \begin{cases} 1 & x > 0 \\ 0 & x < 0 \end{cases} \tag{1}$$

$v_{11} = 0.6, v_{12} = 0.1, v_{21} = 0.2, v_{22} = 0.7, w_1 = 0.5, w_2 = 0.8, \alpha_1, \alpha_2, \gamma$为节点1,2,3的输入值，$\beta_1, \beta_2, \hat{\gamma}$为节点1,2,3的输出值

第一次正向传播：

$\alpha_1 = v_{11}x_1 + v_{21}x_2 = 0.6 \times 0.2 + 0.2 \times 0.3 = 0.18$

$\beta_1 = max(0, \alpha_1) = 0.18$

$\alpha_2 = v_{12}x_1 + v_{22}x_2 = 0.1 \times 0.2 + 0.7 \times 0.3 = 0.23$

$\beta_2 = max(0, \alpha_2) = 0.23$

$\gamma = w_1\beta_1 + w_2\beta_2 = 0.5 \times 0.18 + 0.8 \times 0.23 = 0.274$

$\hat{\gamma} = max(0, \gamma) = 0.274$

误差$E = \frac{1}{2}(\gamma - \hat{\gamma})^2 = 0.5 \times (0.5 - 0.274)^2 = 0.025538$

误差逆传播计算梯度：

$\frac{\partial E}{\partial v_{11}} = \frac{\partial E}{\partial \hat{\gamma}} \frac{\partial \hat{\gamma}}{\partial \gamma} \frac{\partial \gamma}{\partial \beta_1} \frac{\partial \beta_1}{\partial \alpha_1} \frac{\partial \alpha_1}{\partial v_{11}} = (\hat{\gamma} - \gamma) ReLU'(\gamma) w_1 ReLU'(\alpha_1) x_1 = -0.0226$

$\frac{\partial E}{\partial v_{12}} = \frac{\partial E}{\partial \hat{\gamma}} \frac{\partial \hat{\gamma}}{\partial \gamma} \frac{\partial \gamma}{\partial \beta_2} \frac{\partial \beta_2}{\partial \alpha_2} \frac{\partial \alpha_2}{\partial v_{12}} = (\hat{\gamma} - \gamma) ReLU'(\gamma) w_2 ReLU'(\alpha_1) x_1 = -0.03616$

$\frac{\partial E}{\partial v_{21}} = \frac{\partial E}{\partial \hat{\gamma}} \frac{\partial \hat{\gamma}}{\partial \gamma} \frac{\partial \gamma}{\partial \beta_1} \frac{\partial \beta_1}{\partial \alpha_1} \frac{\partial \alpha_1}{\partial v_{21}} = (\hat{\gamma} - \gamma) ReLU'(\gamma) w_1 ReLU'(\alpha_1) x_2 = -0.0339$

$\frac{\partial E}{\partial v_{22}} = \frac{\partial E}{\partial \hat{\gamma}} \frac{\partial \hat{\gamma}}{\partial \gamma} \frac{\partial \gamma}{\partial \beta_2} \frac{\partial \beta_2}{\partial \alpha_2} \frac{\partial \alpha_2}{\partial v_{22}} = (\hat{\gamma} - \gamma) ReLU'(\gamma) w_2 ReLU'(\alpha_1) x_2 = -0.05424$

$\frac{\partial E}{\partial w_1} = \frac{\partial E}{\partial \hat{\gamma}} \frac{\partial \hat{\gamma}}{\partial \gamma} \frac{\partial \gamma}{\partial \beta_1} = (\hat{\gamma} - \gamma) ReLU'(\gamma) \beta_1 = -0.04068$

$\frac{\partial E}{\partial w_2} = \frac{\partial E}{\partial \hat{\gamma}} \frac{\partial \hat{\gamma}}{\partial \gamma} \frac{\partial \gamma}{\partial \beta_2} = (\hat{\gamma} - \gamma) ReLU'(\gamma) \beta_1 = -0.05198$

学习率$\eta = 1$，更新权重

$v'_{11} = v_{11} - \eta \frac{\partial E}{\partial v_{11}} = 0.6226$

$v'_{12} = v_{12} - \eta \frac{\partial E}{\partial v_{12}} = 0.13616$

$v'_{21} = v_{21} - \eta \frac{\partial E}{\partial v_{21}} = 0.2339$

$v'_{22} = v_{22} - \eta \frac{\partial E}{\partial v_{22}} = 0.75424$

$w'_1 = w_1 - \eta \frac{\partial E}{\partial w_1} = 0.54068$

$w'_2 = w_2 - \eta \frac{\partial E}{\partial w_2} = 0.85198$

第二次正向传播更新输出使用新权重，计算过程与第一次一致，计算结果如下：

$\alpha'_1 = 0.19469, \alpha'_2 = 0.253504, \beta'_1 = 0.19469, \beta'_2 = 0.253504, \gamma' = 0.3212, \hat{\gamma}' = 0.32125, E' = 0.015976$

$0.015976 \leq 0.025538$，可知参数更新降低了平方损失。

2 HW6

2.1 试讨论线性判别分析与线性核支持向量机在何种条件下等价

（言之有理即可）

解：线性判别分析能够解决 n 分类问题，而线性核 SVM 只能解决二分类问题。当线性判别分析的投影向量和线性核SVM的超平面向量垂直的时候，SVM 的最大间隔就是线性判别分析所要求的异类投影点间距，同时在这种情况下，线性判别分析的同类样例的投影点也会被这个超平面所划分在一起，使其间隔较小。所以（1）线性判别分析求解出来的投影向量和线性核 SVM 求解出来的超平面向量垂直，（2）数据集只有两类，（3）数据集线性可分时，SVM 和 LDA等价。

2.2 试析SVM对噪声敏感的原因

（言之有理即可）

解：（1）SVM 的基本形态是一个硬间隔分类器，它要求所有样本都满足硬间隔约束，因此噪

声很容易影响 SVM 的学习。（2）存在噪声时，SVM 容易受噪声信息的影响，将训练得到的超平面向两个类间靠拢，导致训练的泛化能力降低，尤其是当噪声成为支持向量时，会直接影响整个超平面。（3）当 SVM 推广到到使用核函数时，会得到一个更复杂的模型，此时噪声也会一并被映射到更高维的特征，可能会对训练造成更意想不到的结果。综上， SVM 对噪声敏感。

2.3 试使用核技巧推广对率回归产生“核对率回归”

（使用对率回归函数推导或使用对率损失函数代替0/1损失函数推导也可）

解：对率回归的L2正则化目标函数

$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m}(-y_i\boldsymbol{\beta}^T\hat{x}_i + ln(1+e^{\boldsymbol{\beta}^T\hat{x}_i}))$

$F = \ell(\beta) + \dfrac{1}{2} \parallel \beta \parallel$

设ϕ为$x->F$的映射，在F中作对率回归，即$h(\hat{x}) = \beta^T\phi(\hat{x})$

由表示定理可得，$h(\hat{x}) = \sum_{i=1}^{m}\alpha_i\kappa(\hat{x},\hat{x_i})$，所以 $\boldsymbol{\beta} = \sum_{i=1}^{m}\alpha_i\phi(\boldsymbol{x}_i)$，带入上式可得：

$$
\begin{aligned}
F &= \sum_{i=1}^{m}(-y_i\boldsymbol{\beta}^T\phi(\hat{\boldsymbol{x}}_{\boldsymbol{i}}) + ln(1+e^{\boldsymbol{\beta}^T\phi(\hat{\boldsymbol{x}}_{\boldsymbol{i}})})) + \frac{\boldsymbol{\lambda}}{2} \parallel \boldsymbol{\beta} \parallel^2 \\
&= \sum_{i=1}^{m}(-y_i\sum_{j=1}^{m}\alpha_j\phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_j) + ln(1+e^{\sum_{j=1}^{m}\alpha_j\phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_j)})) + \frac{\boldsymbol{\lambda}}{2} \parallel \sum_{j=1}^{m}\alpha_j\phi(\boldsymbol{x}_j) \parallel^2 \\
&= \sum_{i=1}^{m}(-y_i\sum_{j=1}^{m}\alpha_j\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) + ln(1+e^{\sum_{j=1}^{m}\alpha_j\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)})) + \frac{\boldsymbol{\lambda}}{2} \parallel \sum_{j=1}^{m}\alpha_i\alpha_j\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)
\end{aligned}
$$

目标函数为$\min\limits_{\alpha} F$，得到L2正则化下的核对率回归

2.4 支持向量回归的对偶问题如下，

$$
\max_{\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}}} g(\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}}) = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i-\hat{\alpha}_i)(\alpha_j-\hat{\alpha}_j)\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)+\sum_{i=1}^{m}(y_i(\hat{\alpha}_i-\alpha_i)-\epsilon(\hat{\alpha}_i-\alpha_i))
$$

$$
s.t. \quad C \geq \boldsymbol{\alpha},\hat{\boldsymbol{\alpha}} \geq 0 \quad and \quad \sum_{i=1}^{m}(\alpha_i-\hat{\alpha}_i)
$$

请将该问题转化为类似于如下标准型的形式（$\boldsymbol{u},\boldsymbol{v},\boldsymbol{k}$均已知），

$$
\max_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^T\boldsymbol{v} - \frac{1}{2}\boldsymbol{\alpha}^T\boldsymbol{K}\boldsymbol{\alpha}
$$

$$
s.t. \quad C \geq \boldsymbol{\alpha} \geq 0 \quad and \quad \boldsymbol{\alpha}^T\boldsymbol{u} = 0
$$

例如在软间隔SVM中，$\boldsymbol{v}=1,\boldsymbol{u}=\boldsymbol{y},\boldsymbol{K}[i,j]=y_iy_j\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$，

若$\kappa(x_i,x_j)=\phi(x_i)^T\phi(x_j)=(x_i^Tx_j)^2$，求$\phi(x_i)$表达式。

（注意题目有两个小问）

解：

1.

令$\boldsymbol{\alpha}=[\alpha_1,\alpha_2,\cdots,\alpha_m]^T,\boldsymbol{y}=[y_1,y_2,\cdots,y_m]^T,\boldsymbol{K}_{ij}=\kappa(\boldsymbol{x};\boldsymbol{x}_j),\boldsymbol{\epsilon}^*=[\epsilon,\epsilon,\cdots,\epsilon]^T$

$\boldsymbol{\alpha}^* = [\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}]^T, v = [-\boldsymbol{y} - \boldsymbol{\epsilon}^*, \boldsymbol{y} - \boldsymbol{\epsilon}^*]^T, \boldsymbol{K}^* = \begin{bmatrix} \boldsymbol{K} & -\boldsymbol{K} \\ -\boldsymbol{K} & \boldsymbol{K} \end{bmatrix}$

则

$$\begin{aligned}\sum_{i=1}^{m}(y_i(\hat{\alpha}_i - \alpha_i) - \epsilon(\hat{\alpha}_i + \alpha_i)) &= \sum_{i=1}^{m}(\alpha_i(-y_i - \epsilon)) + \hat{\alpha}_i(y_i - \epsilon)) \\ &= \boldsymbol{\alpha}^T(-\boldsymbol{y} - \boldsymbol{\epsilon}^*) + \hat{\boldsymbol{\alpha}}^T(\boldsymbol{y} - \boldsymbol{\epsilon}^*) \\ &= \boldsymbol{\alpha}^{*T}\boldsymbol{v}\end{aligned}$$

$$\begin{aligned}-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i - \hat{\alpha}_i)(\alpha_j - \hat{\alpha}_j)\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) &= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)\alpha_j - \alpha_i\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)\hat{\alpha}_j \\ &\quad - \hat{\alpha}_i\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)\alpha_j + \hat{\alpha}_i\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)\hat{\alpha}_j) \\ &= -\frac{1}{2}(\boldsymbol{\alpha}^T\boldsymbol{K}\boldsymbol{\alpha} - \boldsymbol{\alpha}^T\boldsymbol{K}\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}^T\boldsymbol{K}\boldsymbol{\alpha} + \hat{\boldsymbol{\alpha}}^T\boldsymbol{K}\hat{\boldsymbol{\alpha}}) \\ &= -\frac{1}{2}\boldsymbol{\alpha}^{*T}\boldsymbol{K}^*\boldsymbol{\alpha}^*\end{aligned}$$

其中$\alpha_i^* = \alpha_i$或 $\hat{\alpha}_i$，所以$0 \le \alpha^* \le C.$。

综上，SVM的对偶问题可转化为标准型。

2.

设$\boldsymbol{x}$是m维向量，则

$$\begin{aligned}\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^T\boldsymbol{x}_j)^2 &= (\\ &= (\sum_{u=1}^{m} x_{ui}x_{uj})(\sum_{v=1}^{m} x_{vi}x_{vj}) \\ &= \sum_{1\le u,v\le m} x_{iu}x_{iv}x_{ju}x_{jv} \\ &= [x_1x_1, x_1x_2, \cdots, x_1x_m, \cdots, x_2x_m, \cdots, x_mx_m]^T \times [x_1x_1, x_1x_2, \cdots, x_1x_m, \cdots, x_2x_m, \cdots, x_mx_m]^T\end{aligned}$$

则$\phi(x_i)$为一个m^2维向量，每个分量为$x_{iu}x_{iv}, 1 \le u, v \le m$，

$\phi(x_i) = [x_1x_1, x_1x_2, \cdots, x_1x_m, \cdots, x_2x_m, \cdots, x_mx_m]^T$，各分量互不相同。

HW7 and HW14 Reference

（作业证明题证明思路不唯一，表明思路与关键步骤即可）

1 HW7

1.1 实践中使用式(7.15) 决定分类类别时，若数据的维数非常高，则概率连乘$\prod_i^d P(x_i|c)$的结果通常会非常接近于0从而导致下溢.试述防止下溢的可能方案.

（言之有理即可）

解：通常采用取对数的方法将连乘变为连加：$\Pi_{i=1}^d P(x_i|c) \rightarrow log[\Pi_{i=1}^d P(x_i|c)] = \Sigma_{i=1}^d logP(x_i|c)$

1.2 试证明:二分类任务中两类数据满足高斯分布且方差相同时，线性判别分析产生贝叶斯最优分类器.

（思路与关键步骤正确即可）

解：假设数据满足高斯分布：$P(x|c) \sim N(\mu_c, \Sigma)$，模型中需要确定的参数有均值$\mu_1$和$\mu_0$，以及共同的方差$\Sigma$，$\Sigma$为对称正定矩阵。数据集的对数似然为：

$$\begin{aligned} LL(\mu_1, \mu_0, \Sigma) &= \sum_i logP(x_i|c_i) \\ &= \sum_i log\{\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp[-\frac{1}{2}(x_i - \mu_{c_i})^T \Sigma^{-1}(x_i - \mu_{c_i})]\} \end{aligned}$$

通过最大化对数似然可以求得参数估计：

$$\nabla_{\mu_c} LL = 0 \Rightarrow \mu_c = \frac{1}{|D_c|} \sum_{x_i \in D_c} x_i$$

$$\nabla_{\Sigma^{-1}} LL = 0 \Rightarrow \Sigma = \frac{1}{m} \sum_i (x_i - \mu_{c_i})(x_i - \mu_{c_i})^T$$

上式中求取Σ^{-1}梯度时应用了关系$\nabla_A |A| = |A|(A^{-1})^T$。那么，该贝叶斯分类器的决策函数为：

$$h_{Bayes}(x) = arg \max_c P(c)P(x|c)$$

对于二分类任务，这等价于：

$$
\begin{aligned}
h_{Bayes}(x) &= sign[P(1)P(x|1) - P(0)P(x|0)] \\
&= sign\{exp[-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)] - exp[-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)]\} \\
&= sign[\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1 + 2(\mu_1-\mu_0)^T\Sigma^{-1}x]
\end{aligned}
$$

上式中第二行采取了同先验假设，亦即P(0)=P(1)=1/2. 在3.4节线性判别分析(LDA)中，关于μ_1, μ_0的定义与上面求得的μ_c完全相同，而根据(3.33)式可知，$S_w = m\Sigma$。在3.4节中求得最优投影直线方向为$w = S_w^{-1}(\mu_0 - \mu_1)$， LDA对于新数据的分类是根据投影点距离两个投影中心的距离远近决定的，可以将其表达为:

$$
\begin{aligned}
h_{LDA}(x) &= sign[(w^Tx - w^T\mu_0)^2 - (w^Tx - w^T\mu_1)^2] \\
&= sign\{[2w^Tx - w^T(\mu_1+\mu_0)][w^T(\mu_1-\mu_0)]\}
\end{aligned}
$$

注意到上式第二行右边项$w^T(\mu_1-\mu_0) = -(\mu_1-\mu_0)^TS_w^{-1}(\mu_1-\mu_0)$，而$S_w$为对称正定矩阵，因此该项恒为负，因此，上式可以进一步化简为：

$$
\begin{aligned}
h_{LDA}(x) &= sign[w^T(\mu_1+\mu_0) - 2w^Tx] \\
&= sign[(\mu_0-\mu_1)^T\Sigma^{-1}(\mu_1+\mu_0-2x)] \\
&= sign[\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1 + 2(\mu_1-\mu_0)^T\Sigma^{-1}x]
\end{aligned}
$$

对比可知，决策函数$h_{Bayes}(x)$和$h_{LDA}(x)$完全相同，因此可以说LDA产生了最优Bayes分类。

1.3 证明EM算法的收敛性

（证明单调性即可，也可使用Jensen不等式）

解：证明EM算法的收敛性，即证明EM算法每次迭代得到的Θ^t满足：

$$
P(X|\Theta^{t+1}) \geq P(X|\Theta^t)
$$

因为$lnP(X|\Theta) = lnP(X,Z|\Theta) - lnP(Z|X,\Theta)$，两边取关于$Z|X,\Theta^t$的期望有：

$$
\mathbb{E}_{Z|X,\Theta^t}lnP(X|\Theta) = \mathbb{E}_{Z|X,\Theta^t}lnP(X,Z|\Theta) - \mathbb{E}_{Z|X,\Theta^t}lnP(Z|X,\Theta)
$$

因为$lnP(X|\Theta)$与Z无关，所以：

$$
\mathbb{E}_{Z|X,\Theta^t}lnP(X|\Theta) = \int_Z P(Z|X,\Theta^t)lnP(X|\Theta)dZ = lnP(X|\Theta)
$$

同时有$\mathbb{E}_{Z|X,\Theta^t}lnP(X,Z|\Theta) = Q(\Theta,\Theta^t)$，记$H(\Theta,\Theta^t) = \mathbb{E}_{Z|X,\Theta^t}lnP(Z|X,\Theta)$

1)

因为$\Theta^{t+1} = arg\ \underset{\Theta}{max}\, Q(\Theta,\Theta^t)$,

所以$Q(\Theta^{t+1},\Theta^t) \geq Q(\Theta,\Theta^t)$,

令$\Theta = \Theta^t$，则 $Q(\Theta^{t+1},\Theta^t) \geq Q(\Theta^t,\Theta^t)$,

2)

$H(\Theta^{t+1},\Theta^t) - H(\Theta^t,\Theta^t) = -D_{KL}[P(Z|X,\Theta^t)||P(Z|X,\Theta^{t+1})]$

根据KL散度的性质可知，上式小于等于0。

综上可知，$P(X|\Theta^n)$单调递增，上界为1，所以EM算法是收敛的。

1.4 在HMM中，求解概率$P(x_{n+1}|x_1, x_2, \cdots, x_n)$

（注意题目要求的是关于观测序列的条件概率，不是联合概率$P(x_1, y_1, \cdots, x_n, y_n)$）
解：在前向算法中，

$$P(x_1, x_2, \cdots, x_n|\lambda) = \sum_{i=1}^{N} \alpha_n(i)$$

所以

$$\begin{aligned} P(x_{n+1}|x_1, x_2, \cdots, x_n) &= \frac{P(x_1, x_2, \cdots, x_{n+1}|\lambda)}{P(x_1, x_2, \cdots, x_n|\lambda)} \\ &= \frac{\sum_{i=1}^{N} \alpha_{n+1}(i)}{\sum_{i=1}^{N} \alpha_n(i)} \\ &= \frac{\sum_{i=1}^{N} \sum j = 1^N \alpha_n(j) a_{j,i} b_{i,x_n}}{\sum_{i=1}^{N} \alpha_n(i)} \end{aligned}$$

求解概率$P(x_{n+1}|x_1, x_2, \cdots, x_n)$的步骤为：
（1）初值：$\alpha_1(i) = \pi_i b_{i,x_1}$
（2）递推：$\alpha_t(i) = \sum_{j=1}^{N} \alpha_{t-1}(j) a_{j,i} b_{i,x_t}$
（3）终止：$P(x_{n+1}|x_1, x_2, \cdots, x_n) = \frac{\sum_{i=1}^{N} \sum j=1^N \alpha_n(j) a_{j,i} b_{i,x_n}}{\sum_{i=1}^{N} \alpha_n(i)}$

2 HW14

2.1 假设数据集$D = x_1, x_2, \cdots, x_m$，任意$x_i$是从均值为$\mu$、方差$\lambda^{-1}$的正态分布$N(\mu, \lambda^{-1})$中独立采样而得到。假设$\mu$和$\lambda$的先验分布为$p(\mu, \lambda) = N(\mu|\mu_0, (\kappa 0\lambda)^{-1}) Gam(\lambda|a_0, b_0)$，其中$Gam(\lambda|a_0, b_0) = \frac{1}{\Gamma(a_0)} {b_0}^{a_0} \lambda^{a_0 - 1} exp(-b_0\lambda)$

（1）请写出联合概率分布$p(D, \mu, \lambda)$

（2）请写出证据下界（即变分推断的优化目标），并证明其为观测数据边际似然$\sum_{i=1}^{m} logp(x_i)$的下界

（3）请用变分推断法近似推断后验概率$p(\mu, \lambda|D)$

（第一问注意采样，第二问主要依据KL与证据下界关联和KL性质证明，第三问通过求导来计算分布）
解：（1）
因为x_i从正态分布$N(\mu, \lambda^{-1})$中采样而来，所以有$p(x_i|\mu, \lambda) = \sqrt{\frac{\lambda}{2\pi}} exp(-\frac{\lambda(x_i - \mu)^2}{2})$。
所以

$$\begin{aligned} p(D, \mu, \lambda) &= p(D|\mu, \lambda) p(\mu, \lambda) \\ &= \Pi_{i=1}^{m} p(x_i|\mu, \lambda) p(\mu, \lambda) \\ &= \Pi_{i=1}^{m} \sqrt{\frac{\lambda}{2\pi}} exp(-\frac{\lambda(x_i - \mu)^2}{2}) \sqrt{\frac{\kappa_0}{2\pi}} \frac{1}{\Gamma(a_0)} {b_0}^{a_0} \lambda^{a_0 - \frac{1}{2}} exp(-\frac{\kappa_0 \lambda(\mu - \mu_0)^2}{2} - b_0\lambda) \\ &= (\frac{1}{2\pi})^{\frac{m+1}{2}} \frac{\sqrt{\kappa_0}}{\Gamma(a_0)} {b_0}^{a_0} \lambda^{a_0 + \frac{m-1}{2}} exp(-\sum_{i=1}^{m} \frac{\lambda(x_i - \mu)^2}{2} - \frac{\kappa_0 \lambda(\mu - \mu_0)^2}{2} - b_0\lambda) \end{aligned}$$

（2）证据下界为：

$$L = E_q[log\ p(x,z)] - E_q[log\ q(z)] = E_q[p(x|\mu,\lambda)] + E_q[log\ p(\lambda)] - E_q[q(\mu)] - E_q[log\ q(\mu)]$$

变分目标为找到

$$q^*(z) = \underset{q(z)}{arg\ min}\, KL(q(z)||p(z|x))$$

即需要找到$q^*(z) \approx p(z|x)$来近似得到$p(z|x)$，又：

$$KL(q(z)||p(z|x)) = E_q[log\ q(z)] - E_q[log\ p(x,z)] + log\ p(x) \geq 0$$

所以:

$$\sum_{i=1}^{m} log\ p(x_i) = log\ p(x) \geq E_q[log\ p(x,z)] - E_q[log\ q(z)] = L$$

可知，证据下界即为$\sum_{i=1}^{m} logp(x_i)$的下界，得证

（3）

通过最大化L来最小化$KL(q(z)||p(z|x))$

令

$$\frac{\partial L}{\partial q_\lambda(\mu)} = E_\lambda(log\ p(\mu|\lambda)) + E_\lambda(log\ p(D|\mu,\lambda)) - log\ q(\mu) = 0$$

有

$$\begin{aligned} log\ q^*(\mu) &= -\frac{1}{2}E(\lambda\kappa_0)(\mu-\mu_0)^2 - \frac{1}{2}E(\lambda)\sum_{i=1}^{m}(x_i-\mu)^2 \\ &= -\frac{1}{2}E(\lambda)[(\kappa_0+m)\mu^2 + \sum_{i=1}^{m}{x_i}^2 - 2\mu(\kappa_0\mu_0+m\bar{x})] \\ &= -\frac{1}{2}E(\lambda)[(\kappa_0+m)(\mu - \frac{\kappa_0\mu_0+m\bar{x}}{\kappa_0+m})^2 + \sum_{i=1}^{m}{x_i}^2 - \frac{(\kappa_0\mu_0+m\bar{x})^2}{\kappa_0+m}] \\ &\sim N(\mu|\frac{\kappa_0\mu_0+m\bar{x}}{\kappa_0+m}, [(\kappa_0+m)E(\lambda)^{-1}]) \end{aligned}$$

令

$$\frac{\partial L}{\partial q_\mu(\lambda)} = E_\mu(log\ p(D|\mu|,\lambda)) + E_\mu(log\ p(\lambda)) - log\ q(\lambda) = 0$$

有

$$\begin{aligned} log\ q^*(\lambda) &= -\frac{1}{2}\lambda E_\mu(\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^{m}(x_i-\mu)^2) + (a_0-1)log\ \lambda - b_0\lambda + \frac{m+1}{2}log\ \lambda \\ &= (a_0 + \frac{m-1}{2})log\ \lambda - \lambda[b_0 + \frac{1}{2}E_\mu(\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^{m}(x_i-\mu)^2)] \\ &\sim Gam(\lambda|a_0+\frac{m-1}{2}, b_0 + \frac{1}{2}E_\mu(\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^{m}(x_i-\mu)^2)) \end{aligned}$$

所以

$$p(\mu,\lambda|D) \sim N(\mu|\frac{\kappa_0\mu_0+m\bar{x}}{\kappa_0+m}, [(\kappa_0+m)E(\lambda)^{-1}])Gam(\lambda|a_0+\frac{m-1}{2}, b_0+\frac{1}{2}E_\mu(\kappa_0(\mu-\mu_0)^2+\sum_{i=1}^{m}(x_i-\mu)^2))$$

2.2 给出CRF的预测问题的解法

（CRF预测问题可通过维特比算法求解）
解：预测问题，即寻找序列 $y=(y_1,y_2,\cdots,y_n)$, 使得 $P(y|x)$ 最大

$$\begin{aligned} y^* &= arg\ \underset{y}{max}\, P_w(y|x) \\ &= arg\ \underset{y}{max}\, \frac{expw\cdot F(y,x)}{Z_w(x)} \\ &= arg\ \underset{y}{max}\, expw\cdot F(y,x) \\ &= arg\ \underset{y}{max}\, w\cdot F(y,x) \end{aligned}$$

可以看出，CRF的预测问题变为求非规范路径概率最大化的最优路径问题

$$arg\ \underset{y}{max}\, w\cdot F(y,x)$$

此时只需计算非规范化概率，不必计算概率。为了求解最优路径，将优化目标写成如下形式：

$$\underset{y}{max}\sum_{i=1}^{n} w\cdot F_i(y_{i-1},y_i,x)$$

其中，

$$F_i(y_{i-1},y_i,x)=f_1(y_{i-1},y_i,x),f_2(y_{i-1},y_i,x),\cdots,f_K(y_{i-1},y_i,x)^T$$

为局部特征向量。上述最优路径问题可使用维特比算法求解，算法如下：
输入：模型特征向量$F(y,x)$，和权值向量w，观测序列$x=(x_1,x_2,\cdots,x_n)$；
输出：最优路径$y^*=(y_1^*,y_2^*,\cdots,y_n^*)$.
（1）初始化

$$\delta_i(j)=w\cdot F_1(y_0=start,y_1=j,x),j=1,2,\cdots,m$$

（2）递推。对$i=1,2,\cdots,n$

$$\delta_i(l)=\underset{1\le j\le m}{max}\,\delta_{i-1}(j)+w\cdot F_i(y_{i-1}=j,y_i=l,x),l=1,2,\cdots,m$$

$$\Psi_i(l)=arg\ \underset{1\le j\le m}{max}\,\delta_{i-1}(j)+w\cdot F_i(y_{i-1}=j,y_i=l,x),l=1,2,\cdots,m$$

（3）终止

$$\underset{y}{max}\, w\cdot F(y,x)=arg\ \underset{1\le j\le m}{max}\,\delta_n(j)$$

$$y_n^*=arg\ \underset{1\le j\le m}{max}\,\delta_n(j)$$

（4）返回路径

$$y_i^*=\Psi_{i+1}(y_{i+1}^*),i=n-1,n-2,\cdots,1$$

求得最优路径$y^*=(y_1^*,y_2^*,\cdots,y_n^*)$。

Boosting and Clustering

zyt

ustc

2022 年 11 月 12 日

Content

8.2

对于 $0/1$ 损失函数来说, 指数损失函数并非仅有的一致替代函数. 考虑式 (8.5), 试证明: 任意损失函数 $\ell(-f(\boldsymbol{x})H(\boldsymbol{x}))$, 若对于 $H(\boldsymbol{x})$ 在区间 $[-\infty, \delta](\delta > 0)$ 上单调递减, 则 ℓ 是 $0/1$ 损失函数的一致替代函数.
Hint:

$$\begin{aligned} L(H \mid \mathcal{D}) &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\ell(-yH(\boldsymbol{x}))] \\ &= P(y=1 \mid \boldsymbol{x})\ell(-H(\boldsymbol{x})) + P(y=-1 \mid \boldsymbol{x})\ell(H(\boldsymbol{x})) \end{aligned}$$

$$\frac{\partial L(H \mid \mathcal{D})}{\partial H(\boldsymbol{x})} = P(\boldsymbol{x})(-P(y=1 \mid \boldsymbol{x})\ell(-H(\boldsymbol{x})) + P(y=-1 \mid \boldsymbol{x})\ell(H(\boldsymbol{x}))) = 0$$

We can get

$$P(y=1 \mid \boldsymbol{x})\ell(-H(\boldsymbol{x})) = P(y=-1 \mid \boldsymbol{x})\ell(H(\boldsymbol{x}))$$

Suppose $P(y=1 \mid \boldsymbol{x}) > P(y=-1 \mid \boldsymbol{x})$, get $\text{sign}(H(x)) = 1$

给定任意的两个相同长度向量 x, y ，其余弦距离为 $1-\frac{x^{\top} y}{|x| y|}$ ，证明余弦距离不满足传递性，而余弦夹角 $\arccos\left(\frac{x^{\top} y}{|x||y|}\right)$ 满足

Hint: transitivity: $d(x, y)+d(y, z) \geq d(x, z)$

For cosine distance, this equals to $(x-y)^{\mathrm{T}}(x+z) \geq 0$, which is easy to validate.

For arccosine, WLOG we can suppose x, y, z both are unit vectors, then this euqal to $\cos(\arccos x^T z) \geq \cos(\arccos x^T y + \arccos y^T z)$

$$\Leftrightarrow x^T z \geq (x^T y)(y^T z) - \sqrt{(1-x^T y)^2(1-y^T z)^2}$$

Some errors: directly from geometric/ in above using $(x^T y)(y^T z) = x^T (yy^T) z$, only consider 3d situation.

kmeans 算法收敛性
Hint:
We show that the loss function is guaranteed to decrease monotonically in each iteration.
EM convergence. See PRML P425 for details.
Some errors: Only prove the corroctness of center choosen.

在k-means算法中替换欧式距离为其他任意的度量, 请问 “聚类中心如何计算?

Hint: Object function: $L = \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in C_i} \operatorname{dist}(\boldsymbol{x}, \boldsymbol{\mu}_i)$

If L is idfferentiable, then we let $\frac{\partial E}{\partial \boldsymbol{\mu}_i} = 0 (i = 1, \ldots, k)$

ch8 Ensemble

1. Are all ensemble make model better?
2. Boosting, why adaboost is useful? How the parameters change in each iteration in adaboost? The procedure of GBDT.
3. Bagging, why bagging is useful? The procedure of stacking.

ch9 Clustering

1. Metrics: Jaccard, FM, Rand, DBI, DI
2. Distance: Minkowski, VDM
3. Methods: Kmeans(convergence proof, choosen of initial parameters), LVQ, GMM, DBSCAN, AGNES(different linkage) and the difference of this methods(i.e. in special situation, which one is better)
4. Theoretical Analysis: convergence of kmeans, convergence of GMM,

8.2

2022 年 11 月 13 日

修改题目为

对于 0/1 损失函数来说, 指数损失函数并非仅有的一致替代函数. 考虑式 (8.5), 试证明: 任意损失函数 $\ell(-f(\boldsymbol{x})H(\boldsymbol{x}))$, 若对于 $\boldsymbol{f}H(\boldsymbol{x})$ 在区间 $[-\infty,\delta](\delta>0)$ 上单调递减, 则 ℓ 是 0/1 损失函数的一致替代函数.

$$\begin{aligned} L(H\mid\mathcal{D}) &= E_{\boldsymbol{x}\sim\mathcal{D}}[\ell(-yH(\boldsymbol{x}))] \\ &= P(y=1\mid\boldsymbol{x})\ell(-H(\boldsymbol{x}))+P(y=-1\mid\boldsymbol{x})\ell(H(\boldsymbol{x})) \end{aligned}$$

(L(−U))函数的特点是: 在 $[-\infty,\delta]$ 区间是单调递减函数(无论其凹凸性如何), 在 $[\delta,+\infty]$ 区间, 可以是任意形状曲线, 无论其单调性如何。对该损失函数进行最小化时，所对应的横坐标位置 u^* 总是在 δ 右侧，也就是 $\mathrm{f}\left(x^*\right)\mathrm{H}\left(x^*\right)\geq 0>0$, 这说明 H(x) 与 f(x) 同正负号。因此 $\operatorname{sign}\left(\mathrm{H}\left(x^*\right)\right)=\mathrm{f}\left(x^*\right)$, 其结果与最小化 0/1 损失函数结果一致, 是一致替代函数。

HW10 & HW11 Reference

注意：方法不唯一，言之成理即可！

1 HW10

1.1 记 $\mathrm{err}^*(\boldsymbol{x}) = 1 - \max_{c\in\mathcal{Y}} P(c|\boldsymbol{x})$，$\mathrm{err}(\boldsymbol{x}) = 1 - \sum_c P(c|\boldsymbol{x})P(c|\boldsymbol{z})$，其中 $\boldsymbol{z}$ 为 $\boldsymbol{x}$ 的最近邻，试证明在样本无穷多时

$$\mathrm{err}^*(\boldsymbol{x}) \le \mathrm{err}(\boldsymbol{x}) \le \mathrm{err}^*(\boldsymbol{x}) \left(2 - \frac{|\mathcal{Y}|}{|\mathcal{Y}|-1} \times \mathrm{err}^*(\boldsymbol{x})\right)$$

提示：柯西-施瓦兹不等式 $(\sum_i a_i)^2 \le n(\sum_i a_i^2)$。

证明. 先证明左边不等式：

$$\begin{aligned}
\mathrm{err}^*(\boldsymbol{x}) &= 1 - \max_{c\in\mathcal{Y}} P(c|\boldsymbol{x}) = 1 - \max_{c\in\mathcal{Y}} P(c|\boldsymbol{x}) \cdot \sum_c P(c|\boldsymbol{z}) \\
&= 1 - \sum_c \max_{c\in\mathcal{Y}} P(c|\boldsymbol{x}) \cdot P(c|\boldsymbol{z}) \\
&\le 1 - \sum_c P(c|\boldsymbol{x}) \cdot P(c|\boldsymbol{z}) = \mathrm{err}^*(\boldsymbol{x})
\end{aligned}$$

令 $c^* = \arg\max_c P(c|\boldsymbol{x})$，再证明右边不等式：

$$\begin{aligned}
\mathrm{err}^*(\boldsymbol{x}) &= 1 - \sum_c P(c|\boldsymbol{x}) \cdot P(c|\boldsymbol{z}) \backsimeq 1 - \sum_c P(c|\boldsymbol{x})^2 \\
&\le 1 - P(c^*|\boldsymbol{x})^2 - \sum_{c\ne c^*} P(c|\boldsymbol{x})^2 \\
&\le 1 - P(c^*|\boldsymbol{x})^2 - \frac{1}{|\mathcal{Y}|-1}\left(\sum_{c\ne c^*} P(c|\boldsymbol{x})\right)^2 \\
&= 1 - P(c^*|\boldsymbol{x})^2 - \frac{1}{|\mathcal{Y}|-1}\left(1 - P(c^*|\boldsymbol{x})\right)^2 \\
&= \left(1 - P(c^*|\boldsymbol{x})\right) \cdot \left(1 + P(c^*|\boldsymbol{x}) - \frac{1}{|\mathcal{Y}|-1}\left(1 - P(c^*|\boldsymbol{x})\right)\right) \\
&= \left(1 - P(c^*|\boldsymbol{x})\right) \cdot \left(2 - \frac{|\mathcal{Y}|}{|\mathcal{Y}|-1}\left(1 - P(c^*|\boldsymbol{x})\right)\right) \\
&= \mathrm{err}^*(\boldsymbol{x}) \cdot \left(2 - \frac{|\mathcal{Y}|}{|\mathcal{Y}|-1} \cdot \mathrm{err}^*(\boldsymbol{x})\right)
\end{aligned}$$

综上所述，证毕！ □

1.2 在实践中，协方差矩阵 $\mathbf{X}\mathbf{X}^\top$ 的特征值分解常由中心化后的样本矩阵 $\mathbf{X}$ 的奇异值分解替代，试述其原因。

解. • 仅供参考，言之成理即可。

令 $\mathbf{X}\mathbf{X}^\top$ 的特征值分解为

$$\mathbf{X}\mathbf{X}^\top = \mathbf{Y}\mathbf{\Lambda}\mathbf{Y}^\top \tag{1}$$

令 $\mathbf{X}$ 的奇异值分解为 $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$，可得

$$\mathbf{X}\mathbf{X}^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{V}\mathbf{\Sigma}^\top\mathbf{U}^\top$$

因为 $\mathbf{X}$ 是经过中心化的样本矩阵，因此 $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$，$\mathbf{U}^\top\mathbf{U} = \mathbf{I}$，所以

$$\mathbf{X}\mathbf{X}^\top = \mathbf{U}(\mathbf{\Sigma}\mathbf{\Sigma}^\top)\mathbf{U}^\top \tag{2}$$

如果令 $\mathbf{Y} = \mathbf{U}$、$\mathbf{\Lambda} = \mathbf{\Sigma}\mathbf{\Sigma}^\top$，不难发现式(1)和(2)是等价的，也是就是协方差矩阵的特征值分解与中心化后的样本矩阵的奇异值分解其实是等价的。

除此外，相较于特征值分解，奇异值分解的运算要更加高效，节省存储空间。 □

1.3 求解优化问题

$$\begin{aligned} \max_{\mathbf{W}} \quad & \operatorname{tr}\left(\mathbf{W}^\top\mathbf{X}\mathbf{X}^\top\mathbf{W}\right) \\ \text{s.t.} \quad & \mathbf{W}^\top\mathbf{W} = \mathbf{I}_{d'} \end{aligned}$$

解. 先将问题转化为等价问题，

$$\begin{aligned} \min_{\mathbf{W}} \quad & -\operatorname{tr}\left(\mathbf{W}^\top\mathbf{X}\mathbf{X}^\top\mathbf{W}\right) \\ \text{s.t.} \quad & \mathbf{W}^\top\mathbf{W} = \mathbf{I}_{d'} \end{aligned}$$

然后使用拉格朗日乘子法，构造拉格朗日函数

$$L(\mathbf{W}, \mathbf{\Lambda}) = -\operatorname{tr}\left(\mathbf{W}^\top\mathbf{X}\mathbf{X}^\top\mathbf{W}\right) + \operatorname{tr}\left(\mathbf{\Lambda}(\mathbf{W}^\top\mathbf{W} - \mathbf{I}_{\mathbf{d}'})\right)$$

其中，$\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \cdots, \lambda_{d'})$，于是令

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{W}} &= -\frac{\partial}{\partial \mathbf{W}}\operatorname{tr}\left(\mathbf{W}^\top\mathbf{X}\mathbf{X}^\top\mathbf{W}\right) + \frac{\partial}{\partial \mathbf{W}}\operatorname{tr}\left(\mathbf{\Lambda}(\mathbf{W}^\top\mathbf{W} - \mathbf{I}_{\mathbf{d}'})\right) \\ &= -2\mathbf{X}\mathbf{X}^\top\mathbf{W} + 2\mathbf{W}\mathbf{\Lambda} \\ &= 0 \end{aligned}$$

解得

$$\mathbf{X}\mathbf{X}^\top\mathbf{W} = \mathbf{W}\mathbf{\Lambda}$$

这意味着

$$\mathbf{X}\mathbf{X}^\top\boldsymbol{w}_i = \lambda_i\boldsymbol{w}_i$$

也就是说，取 $\mathbf{X}\mathbf{X}^\top$ 的最大的前 d' 个特征值所对应的特征向量即可得 $\mathbf{W}$。将之代入到目标函数即可得

$$\operatorname{tr}\left(\mathbf{W}^\top\mathbf{X}\mathbf{X}^\top\mathbf{W}\right) = \sum_{i=1}^{d'}\lambda_{d'}.$$

□

1.4 令 $\mathbf{M}=\mathbf{P}\mathbf{P}^\top$，那么下列问题还是凸优化问题吗？试证明之。

$$\begin{aligned}\min_{\mathbf{P}} \quad & \sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{M}} \| \boldsymbol{x}_i-\boldsymbol{x}_j \|_{\mathbf{M}}^2 \\ \text{s.t.} \quad & \sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{C}} \| \boldsymbol{x}_i-\boldsymbol{x}_j \|_{\mathbf{M}}^2 \geq 1\end{aligned}$$

凸优化问题一般具有如下形式

$$\begin{aligned}\min \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x)\leq b_i,\ i=1,\dots,m \\ & h_i(x)=0,\ i=1,\dots,p\end{aligned}$$

其中函数 $f_0,f_1,\dots,f_m$ 是凸函数，$h_0,h_1,\dots,h_p$ 是仿射函数。

证明. 令

$$\begin{aligned}\boldsymbol{\Delta}_{i,j}&=\boldsymbol{x}_i-\boldsymbol{x}_j \\ f(\mathbf{P})&=\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{M}} \| \boldsymbol{x}_i-\boldsymbol{x}_j \|_{\mathbf{M}}^2=\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{M}} \boldsymbol{\Delta}_{i,j}^\top\mathbf{P}\mathbf{P}^\top\boldsymbol{\Delta}_{i,j} \\ g(\mathbf{P})&=1-\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{C}} \boldsymbol{\Delta}_{i,j}^\top\mathbf{P}\mathbf{P}^\top\boldsymbol{\Delta}_{i,j}\end{aligned}$$

此时，可将上述优化问题转化为

$$\begin{aligned}\min \quad & f(\mathbf{P}) \\ \text{s.t.} \quad & g(\mathbf{P})\leq 0\end{aligned}$$

根据凸优化问题的一般形式可知，只要证明 $f(\mathbf{P})$ 和 $g(\mathbf{P})$ 同时都为凸函数，即可证该问题是一个凸优化问题。接下来证明 $f(\mathbf{P})$ 和 $g(\mathbf{P})$ 是否为凸函数：

$$\begin{aligned}\frac{\partial f}{\partial \mathbf{P}}&=2\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{M}} \boldsymbol{\Delta}_{i,j}\boldsymbol{\Delta}_{i,j}^\top\mathbf{P} \\ \frac{\partial^2 f}{\partial \mathbf{P}\partial \mathbf{P}^\top}&=2\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{M}} \boldsymbol{\Delta}_{i,j}^\top\boldsymbol{\Delta}_{i,j} \\ \frac{\partial g}{\partial \mathbf{P}}&=-2\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{C}} \boldsymbol{\Delta}_{i,j}\boldsymbol{\Delta}_{i,j}^\top\mathbf{P} \\ \frac{\partial^2 g}{\partial \mathbf{P}\partial \mathbf{P}^\top}&=-2\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{C}} \boldsymbol{\Delta}_{i,j}^\top\boldsymbol{\Delta}_{i,j}\end{aligned}$$

因为 $\boldsymbol{\Delta}_{i,j}^\top\boldsymbol{\Delta}_{i,j}=(\boldsymbol{x}_i-\boldsymbol{x}_j)^\top(\boldsymbol{x}_i-\boldsymbol{x}_j)\geq 0$，那么

$$\begin{aligned}\frac{\partial^2 f}{\partial \mathbf{P}\partial \mathbf{P}^\top}&\geq 0 \\ \frac{\partial^2 g}{\partial \mathbf{P}\partial \mathbf{P}^\top}&\leq 0\end{aligned}$$

因此 $f(\mathbf{P})$ 是凸函数，$g(\mathbf{P})$ 不是凸函数，所以该问题不是凸优化问题。 □

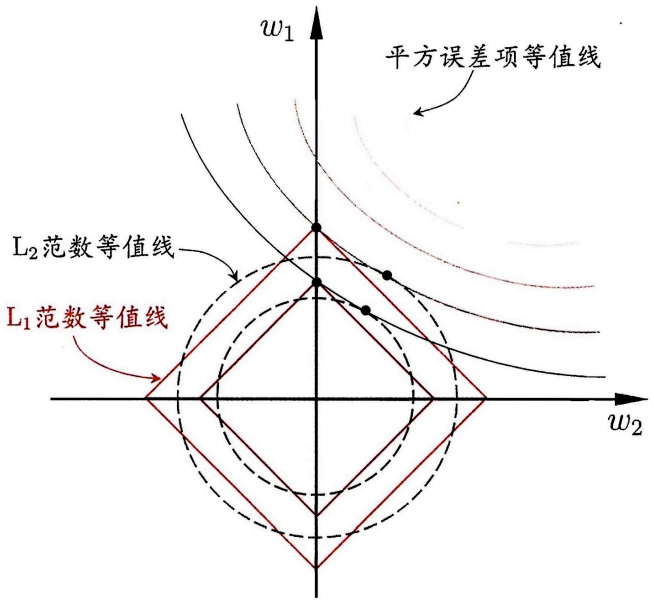


图 11.2 L_1 正则化比 L_2 正则化更易于得到稀疏解

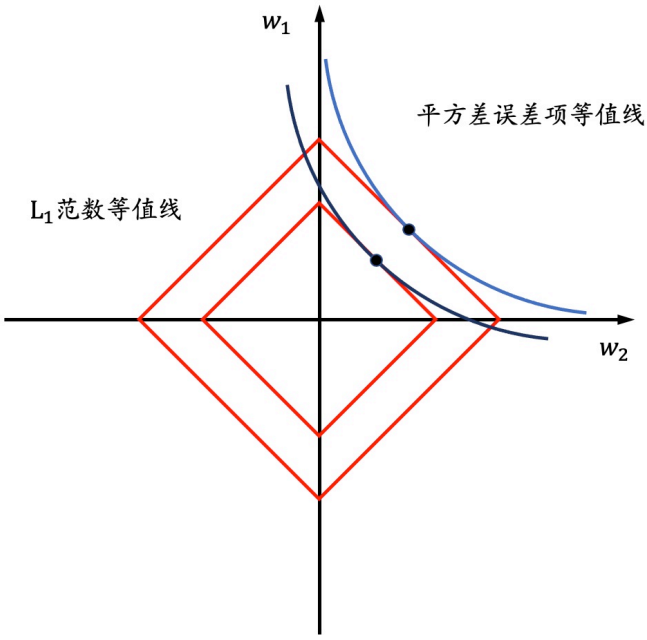


图 1: 情况演示图

2 HW11

2.1 [课本习题 11.5] 结合图 11.2，试举例说明 L_1 正则化在何种情形下不能产生稀疏解。

解. 如图1所示，当平方差误差项等值线的斜率较大的时候，其与 L_1 范数等值线的交点就不再位于坐标轴上，因此将无法产生稀疏解。

□

2.2 [课本习题 11.7] 试述直接求解 L_0 范数正则化会遇到的困难。

解. L_0 范数是统计向量非零元素的个数，不连续、不可微、非凸，无法通过凸优化的方式求解，需要采用遍历方式才能找到最优解，因此难度是 NP-难的。 □

2.3 [PPT 20 页] 证明回归和对率回归的损失函数的梯度是否满足 L-Lipschitz 条件，并求出 L。

证明. 先证明线性回归函数，其损失函数为

$$E(\boldsymbol{w}) = (\boldsymbol{y} - \mathbf{X}\boldsymbol{w})^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{w})$$

不难发现该函数为凸函数，其微分算子 ∇E 表示为

$$\nabla E(\boldsymbol{w}) = 2\mathbf{X}^\top (\mathbf{X}\boldsymbol{w} - \boldsymbol{y})$$

对于 $\forall \boldsymbol{w}, \boldsymbol{w}'$，都有

$$\begin{aligned} \| \nabla E(\boldsymbol{w}) - \nabla E(\boldsymbol{w}') \|_2 &= \| 2\mathbf{X}^\top \mathbf{X}(\boldsymbol{w} - \boldsymbol{w}') \|_2 \\ &\leq 2 \| \mathbf{X}^\top \mathbf{X} \|_2 \cdot \| \boldsymbol{w} - \boldsymbol{w}' \|_2 \end{aligned}$$

令 $L = 2 \| \mathbf{X}^\top \mathbf{X} \|_2 > 0$，可以发现线性回归函数的损失函数满足 L-Lipschitz 条件。

接着证明对率回归函数，其损失函数为

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m} \left(-y_i \boldsymbol{\beta}^\top \boldsymbol{x}_i + \ln \left(1 + e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i} \right) \right)$$

该函数是关于 $\boldsymbol{\beta}$ 的高阶可导连续凸函数，其微分算子 $\nabla \ell$ 表示为

$$\nabla \ell(\boldsymbol{\beta}) = \sum_{i=1}^{m} \left(-y_i \boldsymbol{x}_i + \frac{\boldsymbol{x}_i e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}}{1 + e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}} \right) = \sum_{i=1}^{m} \left(\frac{1}{1 + e^{-\boldsymbol{\beta}^\top \boldsymbol{x}_i}} - y_i \right) \boldsymbol{x}_i$$

对于 $\forall \boldsymbol{\beta}, \boldsymbol{\beta}'$，都有

$$\| \nabla \ell(\boldsymbol{\beta}) - \nabla \ell(\boldsymbol{\beta}') \|_2 = \| \sum_i \left(\frac{1}{1 + e^{-\boldsymbol{\beta}^\top \boldsymbol{x}_i}} - \frac{1}{1 + e^{-\boldsymbol{\beta}'^\top \boldsymbol{x}_i}} \right) \boldsymbol{x}_i \|_2$$

注意到 Sigmod 函数 $f(x) = \frac{1}{1+e^{-x}}$ 上任意两点连线的斜率小于等于 $f'(0)$，因此可知

$$\frac{1}{1 + e^{-\boldsymbol{\beta}^\top \boldsymbol{x}_i}} - \frac{1}{1 + e^{-\boldsymbol{\beta}'^\top \boldsymbol{x}_i}} \leq \left(\frac{1}{1 + e^{-\boldsymbol{\beta}^\top \boldsymbol{x}_i}} \right)' \Bigg|_{\boldsymbol{\beta}^\top \boldsymbol{x}_i = 0} \left(\boldsymbol{\beta}^\top - \boldsymbol{\beta}'^\top \right) \boldsymbol{x}_i = \frac{1}{4} \boldsymbol{x}_i^\top \left(\boldsymbol{\beta} - \boldsymbol{\beta}' \right)$$

因此可得

$$\begin{aligned} \| \nabla \ell(\boldsymbol{\beta}) - \nabla \ell(\boldsymbol{\beta}') \|_2 &\leq \| \sum_i \frac{1}{4} \boldsymbol{x}_i^\top \left(\boldsymbol{\beta} - \boldsymbol{\beta}' \right) \boldsymbol{x}_i \|_2 \\ &\leq \frac{1}{4} \| \sum_i \boldsymbol{x}_i^\top \boldsymbol{x}_i \|_2 \cdot \| \boldsymbol{\beta} - \boldsymbol{\beta}' \|_2 \end{aligned}$$

令 $L = \frac{1}{4} \| \sum_i \boldsymbol{x}_i^\top \boldsymbol{x}_i \|_2 > 0$，可以发现对率回归函数的损失函数满足 L-Lipschitz 条件。

□