Regularization
正则化

郑滕飞

2023.5.17

目录

多项式逼近问题

问题：给定 $(x_i, y_i), i = 0, \ldots, n-1$，且 x_i 互不相同，求逼近它们的一个至多 $n-1$ 次多项式 $p(x)$。

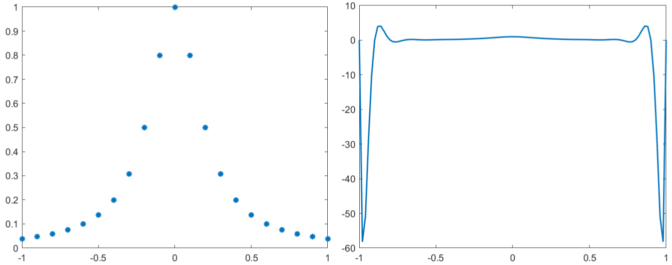
- 利用 Vandermonde 行列式容易看出存在唯一精确解
- 可证明 $x \in [a, b]$ 时多项式 $p(x)$ 插值 $f(x)$ 误差界

$$\frac{\prod_{i=0}^{n-1}(x - x_i)}{n!} \max_{t\in[a,b]} |f^{(n)}(t)|$$

- 若可以自由采样，选取合适结点 (Chebyshev point) 可以显著降低误差

龙格现象

一般结点 (如等距结点) 插值多项式，在边界处容易出现很高的误差。



泛化误差

对拟合问题，可以考虑简单的均方误差作为**泛化误差** (与模型泛化能力相关)：

$$MSE(\hat{f}(x)) = E[(\hat{f}(x) - f(x))^2] = E[\hat{f}(x) - f(x)]^2 + Var(\hat{f}(x))$$

- 包含**偏差项**与**方差项**两项
- 也对应近似误差与估计误差 (测试集结果未知)
- 近似误差降低，但估计误差反而升高的情况即为**过拟合**

偏差-方差权衡

- No Free Lunch Theorem
- 实际情况中基本不可能同时降低偏差与方差
- 无偏估计一般易于直接求解，于是考虑在偏差允许的范围内降低方差

正则化项的引入

回到多项式逼近问题，严格的逼近也就相当于

$$\min_{p\in\mathbb{R}_n[x]} \sum_i (p(x_i) - y_i)^2$$

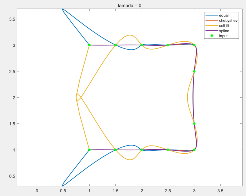
为控制 p 的 “陡峭程度”，改进为

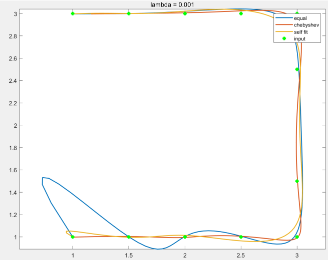
$$\min_{p\in\mathbb{R}_n[x]} \sum_i (p(x_i) - y_i)^2 + \lambda \int_a^b (p'')^2(x)\mathrm{d}x$$

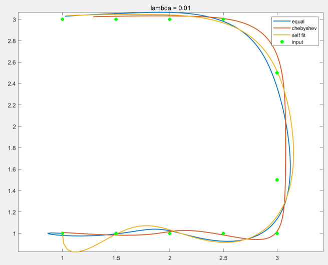
加入的部分即为**正则化项**。

光滑性效果

由两组多项式插值拟合平面参数曲线：



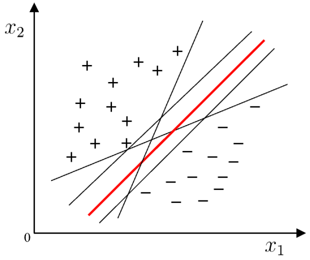




目录

SVM

支持向量机 (Support Vector Machine)：假设两类样本线性可分，寻找最优分割两类样本的超平面。



假设超平面 $w^T x + b = 0$，可得到此例子的优化模型 (标签 $y_i = \pm 1$)：

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

$$s.t. y_i(w^T x_i + b) \geq 1$$

软间隔 SVM

- SVM 是稀疏的
- 只有离分割超平面最近的点才影响值
- 对噪声非常敏感

改进方式：软间隔 SVM

$$\min_{w,b} \frac{1}{2}\|w\|^2 + \frac{\sigma}{2}\sum_i \|c_i(x)\|^2$$

$$c_i(w,b) = \max\{1 - y_i(w^T x_i + b), 0\}$$

罚函数

定义：对问题 $\min_{x\in S} f(x)$，满足 $\begin{cases} P(x) = f(x) & x \in S \\ P(x) > f(x) & x \notin S \end{cases}$ 的函数 P 称为**罚函数**。

硬约束化为软约束得到 $c(x)$，则 $f(x) + \frac{\sigma}{2}\|c(x)\|^2$ 是一种罚函数，且可证明：f 可微时，只要 σ 充分大，最优解即能任意接近真实最优解。

- 对抗噪声
- 便于求解

岭回归

$$\hat{y} = \beta^T x$$

$$\beta = \arg\min_{\beta} \|y - X\beta\|^2$$

$$\Downarrow$$

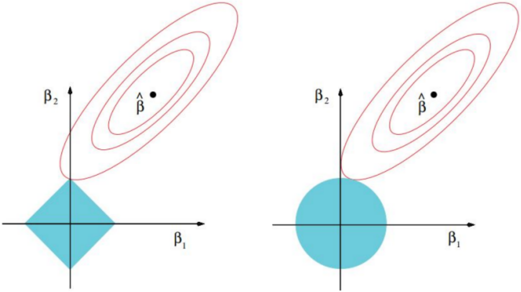
$$\beta = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda\|\beta\|^2$$

- 更广义的回归：$f^{-1}(y) = \beta^T x$
- 硬间隔版本：$\|\beta\|^2 \leq t$
- **权重衰减**作用

Lasso 回归

$$\beta = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda\|\beta\|_1$$

- 岭回归精确解 $\beta = (X^TX + \lambda I)^{-1}X^Ty$
- Lasso 回归难以直接写出精确解，且解关于 y 非线性
- 优点：λ 充分大时会将一些系数精确设定为 0，得到**稀疏解**



XGBoost

利用多个基模型组合成加法模型，假设共 M 个，则前 t 个输出为

$$y_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = y_i^{(t-1)} + f_t(x_i)$$

于是在前 $t-1$ 个固定时优化第 t 个应优化 (loss 表示误差函数，这里取均方误差 $\|\cdot\|^2$，p 表示惩罚项)

$$\sum_i \mathrm{loss}(y_i, y_i^{(t-1)} + f_t(x_i)) + p(f_t)$$

将 loss 泰勒展开，保留 $y_i^{(t-1)}$ 处一阶导数 g_i、二阶导数 h_i 得到优化目标近似为

$$\sum_i g_i f_t(x_i) + \sum_i \frac{1}{2} h_i^2 f_t(x_i) + p(f_t)$$

回归树

回归树模型：每步进行树状划分直到达到底层结点，本质是给出一个映射 q，将每个输入 x_i 映射到叶子 $q(x_i)$，并预测以这个叶子的权重 $w_{q(x_i)}$。

正则化项：叶子数 T 引起的正则化项 γT 与权重引起的正则化项 $\frac{1}{2}\lambda\|w\|^2$ 求和。

若树的映射已经确定，记 $G_j = \sum_{p(x_i)=j} g_i, H_j = \sum_{p(x_i=j)} h_i$，可得优化目标变为

$$\sum_{j=1}^{T}\left(G_j w_j + \frac{1}{2}(H_j+\lambda)w_j\right) + \gamma T \Rightarrow w_j = -\frac{G_j}{H_j+\lambda}$$

结点划分

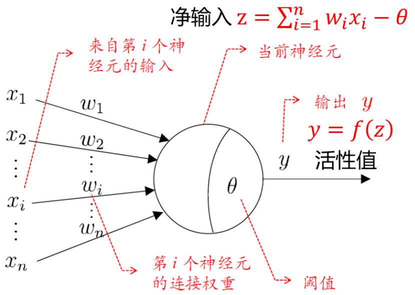
假设旧叶子结点对应 g, h 之和为 G, H，划分为左右后分别对应的和为 G_L, H_L, G_R, H_R，则优化目标的改进为

$$\frac{G_L^2}{2(H_L+\lambda)}+\frac{G_R^2}{2(H_R+\lambda)}-\frac{G^2}{2(H+\lambda)}-\gamma$$

由此，重复寻找最优划分直到无法改进目标函数，并预先给定基学习器的数目，就可以完成整个回归树为基的 XGBoost 构造。反之，若不进行正则化，目标为 $\frac{G_L^2}{2H_L}+\frac{G_R^2}{2H_R}-\frac{G^2}{2H}$，划分为单个才能终止，会产生严重的过拟合。

目录

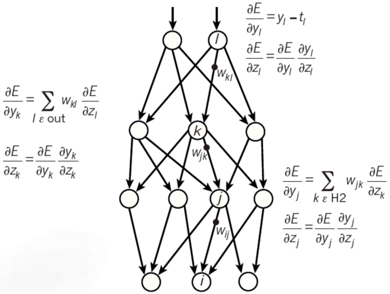
神经网络



- 多层线性回归中加以非线性处理
- 输入通过处理得到输出，可以多层叠加，进行不同连接
- 常用激活函数：ReLU($\max\{0, z\}$)、tanh 等

误差逆传播

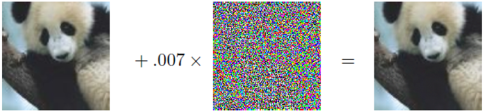
通过梯度下降的方式更新每个权重与阈值的值：



- 先前向计算出结果，再估计误差反向传播
- 当学习率设置得足够低时，一定能够逐渐逼近结果
- 足够大的网络可任意逼近任何连续函数 (万能逼近定理)

对抗学习

由于深度学习过程中的参数远多于实际函数空间的维数，过拟合非常容易发生，几乎必然会学习到无效的特征，由此导致容易构造看起来相同但深度学习确定的标签相反的例子：



x $\quad + .007 \times \quad$ $\text{sign}(\nabla_x J(\theta, x, y))$ $\quad = \quad$ $x + \epsilon\text{sign}(\nabla_x J(\theta, x, y))$

$y =$ 熊猫 w/ 57.7% confidence

线虫 w/ 8.2% confidence

长臂猿 w/ 99.3 % confidence

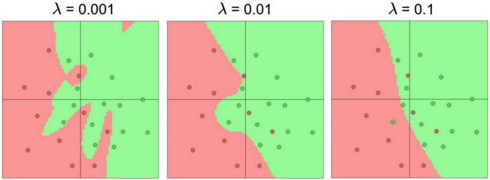
自然，我们可以给权重添加显式的一范数或二范数正则化，不过，现实中也常采用隐式的正则化技巧。

扰动数据集

给原本数据集添加一定噪声 $\epsilon \sim \mathcal{N}(0, \eta I)$，记扰动后为 y_ϵ，若不扰动，模型为 $\hat{y}$，扰动后训练的模型为 $\hat{y}_\epsilon$，并用泰勒展开近似 $\hat{y}_\epsilon = \hat{y} + \nabla_w y^T \epsilon$。
直接代入后利用正态分布的性质展开得到均方误差

$$E[(\hat{y}_\epsilon - y)^2] = E[(\hat{y} - y)^2] + \eta E[\|\nabla_w y\|^2]$$

第二项即为正则化项。



早停策略

每次训练完计算验证集损失，上升时停止训练。
假设损失函数在最优点附近可以二次近似

$$J(w) = J(w^*) + \frac{1}{2}(w - w^*)H(w - w^*)$$

H 代表最优点 Hesse 阵，则梯度为 $H(w - w^*)$。若每次以学习率 t 进行梯度下降，若初始为 0，计算可知下降 τ 次后

$$w^{(\tau)} = (I - (I - tH)^{\tau})w^*$$

对比直接进行 L2 正则化得到的 $w = (H - \alpha I)^{-1}Hw^*$，可近似得到 $\tau \approx \frac{1}{t\alpha}$，也即给定学习率下早停次数与正则化系数反比。

Dropout

每次训练时，概率将一些神经元的输出设为 0。考虑基于指数分布的广义线性模型

$$P_\beta(y|x) = h(y)\exp(yx^T\beta - A(x^T\beta)), A(t) = \ln\int h(y)\exp(yt)\mathrm{d}y$$

在此假设下可以分离出正则化项 $E_\xi[A((x.*\xi)^T\beta)] - A(x_i^T\beta)$ 其中 ξ 的每个分量独立以 σ 概率为 1，否则为 0，$.*$ 代表逐元素相乘。

- 防止特征的 co-adaptaion
- 可以看成一种模型集成
- 在大网络效果良好的正则化技术

The End

Thank you for listening!
部分内容参考自：

- 连德富老师机器学习、深度学习讲义
- 崔文泉老师机器学习方法讲义