2.线性模型

在求解参数模型参数时，涉及到模型选择问题，有两种策略，一种时经验风险最小化（如2.1），但容易出现过拟合，一种是结构风险最小化（如2.2），在经验风险最小化基础上加上一个正则化因子。

2.1 最小二乘法

解析解法

损失函数：

$$w = \arg\min \sum_{i=1}^{n} (y_i - f(x_i; w))^2 = \arg\min_w L$$

$$L = \sum_{i=1}^{n} (y_i - x_i^T w)^2 = ||y - Xw||^2 = (y - Xw)^T (y - Xw)$$

$$\nabla_w L = 2X^T Xw - 2X^T y = 0 \Rightarrow w_{LS} = (X^T X)^{-1} X^T y$$

假设空间：

$$y_i = f(x_i; w) = w_0 + \sum_{i=1}^{d} x_{ij} w_j$$

如果原始变量由向量xi组成，特征可以用基函数表示（类似于基本初等变化，对X进行空间变换），可以依旧使用上述公式求解。 线性回归模型中的各个基函数如何选择，模型的函数就会有不同的结果。而且基函数的选择有时可以将线性模型扩展到一般的非线性形式，只要你将基函数定义为一个非线性函数就好，例如如果每一个基函数是三角函数，那么整个模型就是傅里叶变换的形式。

但是有个问题，一旦取逆不是满秩，那么它将不存在，不能使用解析解法，此时要用到梯度下降法。

梯度下降法

- 梯度下降法

考虑无约束优化问题 $\min_{\boldsymbol{x}} f(\boldsymbol{x})$

若能构造一个序列 $\boldsymbol{x}^0, \boldsymbol{x}^1, \boldsymbol{x}^2, \ldots$ 满足 $f(\boldsymbol{x}^{t+1}) < f(\boldsymbol{x}^t),\ t = 0, 1, 2, \ldots$

则不断执行该过程即可收敛到局部极小点

根据泰勒展式有 $f(\boldsymbol{x} + \Delta\boldsymbol{x}) \simeq f(\boldsymbol{x}) + \Delta\boldsymbol{x}^{\mathrm{T}} \nabla f(\boldsymbol{x})$

于是, 欲满足 $f(\boldsymbol{x} + \Delta\boldsymbol{x}) < f(\boldsymbol{x})$, 可选择

$$\Delta\boldsymbol{x} = -\gamma \nabla f(\boldsymbol{x})$$

其中步长 γ 是一个小常数. 这就是梯度下降法

$$w^{(\tau+1)} = w^{(\tau)} - \eta \nabla \mathcal{L}_n$$

$$w^{(\tau+1)} = w^{(\tau)} + \eta (y_n - w^{(\tau)T} x_n) x_n$$

其中τ表示迭代次数，η是学习率参数

2.2 正则化

为了解决过拟合的问题，引入正则化项，加入罚函数，第一项是经验误差，第二项是惩罚项。

$$w_{opt} = \arg\min_{W} E_D(w) + \lambda E_W(w)$$

惩罚项一般为范数惩罚，

$$||w||_p = (\sum_{j=1}^{d} |w_j|^p)^{\frac{1}{p}} \quad 0 < p < \infty$$

p=2时，E_D为线性函数时，是岭回归。

p=1时，E_D为线性函数时，是LASSO回归。

岭回归

有时光采用训练样本进行无偏估计不好用，要采取结构风险最小化，岭回归的名称是因为求取参数的解析解的时候，最后的表达式是在原来的基础上在求逆矩阵内部加上一个对角矩阵，就好像一条“岭”一样。加上这条岭以后，原来不可求逆的数据矩阵就可以求逆了。同时，对角矩阵其实是由一个参数λ和单位对角矩阵相乘组成。λ越大，说明偏差就越大，原始数据对回归求取参数的作用就越小，当λ取到一个合适的值，就能在一定意义上解决过拟合的问题：原先过拟合的特别大或者特别小的参数会被约束到正常甚至很小的值，但不会为零。

- 岭回归 *Ridge regression*

$$w_{\mathrm{RR}} = \arg\min_{w} \|y - Xw\|^2 + \lambda\|w\|^2$$

惩罚函数 $g(w) = \|w\|^2$ 将惩罚w中数值比较大的元素

$$\lambda \to 0 \ : \ w_{\mathrm{RR}} \to w_{\mathrm{LS}}$$

$$\lambda \to \infty \ : \ w_{\mathrm{RR}} \to \vec{0}$$

$$\mathcal{L} = \|y - Xw\|^2 + \lambda\|w\|^2 = (y - Xw)^T(y - Xw) + \lambda w^T w$$

$$\nabla_w \mathcal{L} = -2X^T y + 2X^T Xw + 2\lambda w = 0$$

$$w_{\mathrm{RR}} = (\lambda I + X^T X)^{-1} X^T y$$

λ是超参数，不是越小越好/越大越好，需要提前设定。这里二范数的正则项偏向于把散点平滑地连在一起，减小（但不是变成0）不重要变量的λ值。**岭回归没有真正解决变量选择的问题。**

LASSO回归

不仅可以解决过拟合问题，而且可以在参数缩减过程中，将一些重复的没必要的参数直接缩减为零，也就是完全减掉了。这可以达到提取有用特征的作用。但是lasso回归的计算过程复杂，毕竟一范数不是连续可导的。

$$w_{LASSO} = \arg\min_{w} ||y - Ww||_2^2 + \lambda||w||_1$$

关于稀疏解：

- 稀疏解

w 中的每个维度与数据 x 中的维度是对应的.

如果 $w_k = 0$, 则预测函数变为：

$$f(x, w) = x^T w = w_1 x_1 + \cdots + 0 \cdot x_k + \cdots + w_d x_d,$$

相当于预测将屏蔽了k个特征.

如果一个w 满足大多数的维度 $= 0$, 则它被称为一个*稀疏 sparse* 解

缩减参数的目的：消除噪声特征，消除关联特征。

2.3 概率机器学习

线性回归的前提都是：

$$Y = Xw + \epsilon \quad \epsilon \sim N(0, \sigma^2 I) \quad Y \sim N(Xw, \sigma^2)$$

$$P(Y|Xw, \sigma^2 I) = \frac{1}{(\sqrt{2\pi}\sigma^2)^n} \exp\{-\frac{1}{2\sigma^2}(Y - Xw)^{-1}(Y - Xw)\}$$

最小二乘回归：使用极大似然估计

$$w_{ML} = arg\max_w \ln P(Y|\mu = Xw, \sigma^2 I)$$

$$= arg\max_w -\frac{1}{2\sigma^2}||Y - Xw||^2 - \frac{n}{2}\ln(2\pi\sigma^2)$$

$$= arg\min_w ||Y - Xw||^2$$

会发现与之前的解析解相同。

岭回归：使用最大后验估计，w服从正态分布

在最小二乘回归的基础上加入了参数的先验分布 $p(w)$：

$w \sim N(0, \lambda^{-1} I)$

$p(w) = \left(\frac{\lambda}{2\pi}\right)^{\frac{d}{2}} e^{-\frac{\lambda}{2} w^T w}$

- 岭回归的模型学习

$$w_{\text{MAP}} = \arg\max_w \ln p(y|w, X) + \ln p(w)$$

$$= \arg\max_w -\frac{1}{2\sigma^2}(y - Xw)^T(y - Xw) - \frac{\lambda}{2} w^T w + \text{const.}$$

$$\nabla_w \mathcal{L} = \frac{1}{\sigma^2} X^T y - \frac{1}{\sigma^2} X^T X w - \lambda w = 0$$

$$w_{\text{MAP}} = (\lambda\sigma^2 I + X^T X)^{-1} X^T y$$

参数λ反比于了参数先验分布的方差，即λ越大，我们假设模型应方差越小，所以岭回归可以鼓励学得一个光滑的模型。

LASSO回归：使用最大后验估计，w服从拉普拉斯分布

$$w \sim Laplace(0, \lambda^{-1})$$

$$p(w) = \frac{\lambda}{2} e^{-\lambda|w|}$$

$$w_{MAP} = \arg\max P(Y \mid X, w) P(w)$$

$$= \arg\max \ln P(Y \mid X, w) \ln P(w)$$

$$= \arg\max(-\frac{1}{2\sigma^2}(Y - Xw)^T (Y - Xw) - \lambda \mid w \mid) + const$$

$$= \arg\min(\frac{1}{2\sigma^2}(Y - Xw)^T (Y - Xw) + \lambda \mid w \mid)$$

可以看出，与最小二乘法得出的结果一样