一文读不懂系列之”线性回归“

1. 模型

- 训练集$\mathcal{D}=\{(\mathbf{x}^{(n)},y^{(n)})\}_{i=1}^{N}$，自变量为样本的特征向量$\mathbf{x}\in\mathbb{R}^{D}$，因变量为$y\in\mathbb{R}$;
- 权重向量$\mathbb{w}\in\mathbb{R}^{D}$和偏置$b\in\mathbb{R}$为可学习参数;
- **线性模型**：函数$f(\mathbf{x};\mathbf{w},b)$

$$f(\mathbf{x};\mathbf{w},b)=\mathbf{w}^{T}\mathbf{x}+b$$

下面就把增广权重向量和增广特征向量统一为$\mathbf{w}$和$\mathbf{x}$，那么线性模型就简写成

$$f(\mathbf{x};\mathbf{w})=\mathbf{w}^{T}\mathbf{x}$$

- 训练集$\mathcal{D}$上的**经验风险**$\mathcal{R}(\mathbf{w})$
  - 取平方损失函数$\mathcal{L}(y^{(n)},f(\mathbf{x}^{(n)};\mathbf{w}))=\frac{1}{2}(y^{(n)}-\mathbf{w}^{T}\mathbf{x}^{(n)})$

$$\mathcal{R}(\mathbf{w})=\frac{1}{2}||\mathbf{y}-\mathbf{X}^{T}\mathbf{w}||^{2}$$

2. 参数估计

$$\begin{cases} y=h(\mathbf{x}) \\ \\ p(y|\mathbf{x}) \end{cases}$$

2.1 LSM | Least Square Method：平方损失的经验风险最小化

$$\begin{aligned}\frac{\partial\mathcal{R}(\mathbf{w})}{\partial\mathbf{w}}&=\frac{1}{2}\frac{\partial||\mathbf{y}-\mathbf{X}^{T}\mathbf{w}||^{2}}{\partial\mathbf{w}}\\&=-\mathbf{X}(\mathbf{y}-\mathbf{X}^{T}\mathbf{w})\end{aligned}$$

令$\frac{\partial\mathcal{R}(\mathbf{w})}{\partial\mathbf{w}}=0$

$$\mathbf{w}^{*}=(\mathbf{X}\mathbf{X}^{T})^{-1}\mathbf{X}\mathbf{y}$$

可以看到在LSM中我们需要$\mathbf{X}\mathbf{X}^{T}\in\mathbb{R}^{(D+1)\times(D+1)}$是可逆的（$rank(\mathbf{X}\mathbf{X}^{T})=D+1$），即可以知道$\mathbf{X}$的行向量是线性不相关，换句话说特征之间是互相独立的(不存在完美的多重共线性[说人话就是不存在精确的线性关系]);

存在的问题

- 当$\mathbf{X}\mathbf{X}^{T}$不可逆时，比较常见的情况是样本数量N小于特征数量$D+1$，则此时$rank(\mathbf{X}\mathbf{X}^{T})=N$，就会有无穷多解$\mathbf{w}^{*}$使得$\mathcal{R}(\mathbf{w}^{*})=0$;
  - 解决方案
    1. 预处理时采用PCA等方法消除不同特征之间的相关性，再使用LSM进行参数估计;
    2. 使用LMS(梯度下降迭代)求解参数;
- 当$\mathbf{X}\mathbf{X}^{T}$可逆时，有可能存在多重共线性(数据集$\mathbf{X}$上小的扰动会导致$(\mathbf{X}\mathbf{X}^{T})^{-1}$发生大的改变)，使得LSM的计算变得不稳定;
  - 解决方案：
    1. 岭回归：$(\mathbf{X}\mathbf{X}^{T}+\lambda I)$，最优参数为

$$\mathbf{w}^* = (\mathbf{X}\mathbf{X}^T + \lambda I)^{-1}\mathbf{X}\mathbf{y}$$

岭回归也可以看作结构风险最小化准则下的LSM，其中

$$\mathcal{R}(\mathbf{w}) = \frac{1}{2}||\mathbf{y} - \mathbf{X}^T\mathbf{w}||^2 + \frac{1}{2}\lambda||\mathbf{w}||^2$$

> 其实多重共线性还有很多其他解决办法【挖🕳待补】

2.2 LMS | Least Mean Square

$$\mathbf{w} \leftarrow \mathbf{w} - \alpha\frac{\partial}{\partial \mathbf{w}}\mathcal{R}(\mathbf{w})$$
$$\mathbf{w} \leftarrow \mathbf{w} + \alpha\mathbf{X}(\mathbf{y} - \mathbf{X}^T\mathbf{w})$$

- 多个样本时梯度下降两种方式
    1. 批量梯度下降法 | Batch Gradient Descent
        - 每一步检查整个训练集中的所有样本；
        - 容易被局部最小值影响；[此处不会，$\mathcal{R}(\mathbf{w})$为凸函数，极小值就是最小值]
    2. 随机梯度下降法 | Stochastic Gradient Descent
        - 每次遇到一个样本就对参数进行更新，对整个训练集进行循环遍历；
        - 训练集(N)很大的时候，一般偏向于选择SGD(BGD需要对整个训练集进行扫描，引起性能开销)

2.3 MLE | Maximum Likelihood Estimation

[Maximum Likelihood Estimation] Choose value that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \underset{\theta}{argmax}\ P(\mathcal{D}|\theta)$$

> 条件概率$p(y|\mathbf{x})$ 角度

假设随机变量y由函数$f(\mathbf{x};\mathbf{w}) = \mathbf{w}^T\mathbf{x}$和随机噪声$\epsilon(\epsilon \sim \mathcal{N}(0, \sigma^2))$，即

$$y = f(\mathbf{x};\mathbf{w}) + \epsilon$$
$$= \mathbf{w}^T\mathbf{x} + \epsilon$$

由此我们可以得到随机变量$y|\mathbf{x};\mathbf{w},\sigma \sim \mathcal{N}(\mathbf{w}^T\mathbf{x}, \sigma^2)$

$$p(y|\mathbf{x};\mathbf{w},\sigma) = \mathcal{N}(y;\mathbf{w}^T\mathbf{x},\sigma^2)$$

则参数$\mathbf{w}$在训练集$\mathcal{D}$上的似然函数为

$$p(\mathbf{y}|\mathbf{X};\mathbf{w},\sigma) = \prod_{n=1}^{N} p(y^{(n)}|\mathbf{x}^{(n)};\mathbf{w},\sigma)$$
$$= \prod_{n=1}^{N} \mathcal{N}(y^{(n)};\mathbf{w}^T\mathbf{x}^{(n)},\sigma^2)$$

其对数似然函数为

$$logp(\mathbf{y}|\mathbf{X};\mathbf{w},\sigma) = \sum_{n=1}^{N} \mathcal{N}(y^{(n)};\mathbf{w}^T\mathbf{x}^{(n)},\sigma^2)$$

则MLE转化为

$$\begin{aligned}\mathbf{w}_{MLE} &= \underset{\mathbf{w}}{argmax}\ p(\mathbf{y}|\mathbf{X};\mathbf{w},\sigma) \\ &= \underset{\mathbf{w}}{argmax}\ logp(\mathbf{y}|\mathbf{X};\mathbf{w},\sigma)\end{aligned}$$

令$\frac{\partial log\, p(\mathbf{y}|\mathbf{X};\mathbf{w},\sigma)}{\partial \mathbf{w}} = 0$，计算得

$$\hat{\mathbf{w}}_{MLE} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{y}$$

2.4 MAP | Maximum A Posterior Estimation

[Maximum A Posterior Estimation] Choose value that is most probable given observed data and prior belief

$$\begin{aligned}\hat{\theta}_{MAP} &= \underset{\theta}{argmax}\ P(\theta|\mathcal{D}) \\ &= \underset{\theta}{argmax}\ P(\mathcal{D}|\theta)P(\theta)\end{aligned}$$

在MLE的假设基础上，进一步假设我们掌握了一些关于参数$\mathbf{w}$的信息，即参数$\mathbf{w}$先验分布为$p(\mathbf{w};v)$，由贝叶斯公式，我们能得到

$$\begin{aligned}\mathbf{w}_{MAP} &= \underset{\mathbf{w}}{argmax}\ p(\mathbf{y}|\mathbf{X};\mathbf{w},\sigma)p(\mathbf{w};v) \\ &= \underset{\mathbf{w}}{argmax}\ log\, p(\mathbf{y}|\mathbf{X};\mathbf{w},\sigma) + log\, p(\mathbf{w};v)\end{aligned}$$

若我们假设这个先验分布为各向同性的高斯分布($p(\mathbf{w};v) = \mathcal{N}(\mathbf{w};\mathbf{0},v^2I)$)，则

$$\begin{aligned}\mathbf{w}_{MAP} &= \underset{\mathbf{w}}{argmax}\ log\, p(\mathbf{y}|\mathbf{X};\mathbf{w},\sigma) + log\, p(\mathbf{w};v) \\ &= -\frac{1}{2\sigma^2}||\mathbf{y}-\mathbf{X}^T\mathbf{w}||^2 - \frac{1}{2v^2}\mathbf{w}^T\mathbf{w}\end{aligned}$$

- 我们看到MAP实际上等价于平方损失的结构风险最小化(正则化系数为$\lambda = \frac{\sigma^2}{v^2}$)
- 当先验分布$p(\mathbf{w};v)$退化为均匀分布时(大白话就是你的先验信息获取了和没获取一样)，此时MAP退化为MLE；

* 整理自

1. nndl
2. cs290 notes1
3. cmu 10-715 lecture1b