1.6 线性回归模型

CSDN学院

2017年10月

▶ 线性回归

不止于代码

- 模型
  – 目标函数（损失函数、正则）
    - 概率解释
- 优化求解
- 模型选择

▶ 线性回归

- 给定训练数据$\mathcal{D}=\{\mathbf{x}_i, y_i\}_{i=1}^N$，其中$y\in\mathbb{R}$，回归学习一个从输入$\mathbf{x}$到输出$y$的映射$f$
- 对新的测试数据$\mathbf{x}$，用学习到的映射对其进行预测：$\hat{y}=f(\mathbf{x})$
- 若假设映射f是一个线性函数，即

$$y=f(\mathbf{x}\,|\,\mathbf{w})=\mathbf{w}^T\mathbf{x}$$

- 我们称之为线性回归模型。

目标函数

- 目标函数通常包含两项：损失函数和正则项

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{N} L(f(\mathbf{x}_i;\boldsymbol{\theta}), y_i) + \lambda\Omega(\boldsymbol{\theta})$$

- 对回归问题，损失函数可以采用L2损失，得到

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{N} L(y_i, \hat{y}_i) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{N} (y_i - \mathbf{w}^T \mathbf{x}_i)^2$$

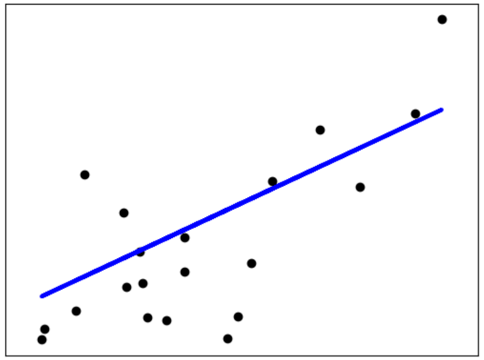
残差平方和(residual sum of squares，RSS)

▶ 线性回归的正则项

- 由于线性模型比较简单，实际应用中有时正则项为空，得到最小二乘线性回归（Ordinary Least Square，OLS）

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{N} L(y_i, \hat{y}_i) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{N} (y_i - \mathbf{w}^T \mathbf{x}_i)^2$$



▶ 线性回归的正则项

- 正则项可以为L2正则，得到岭回归（Ridge Regression）模型：

$$J(\mathbf{w}) = \underbrace{\sum_{i=1}^{N}\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2}_{\text{训练集上残差平方和}} + \lambda\underbrace{\|\mathbf{w}\|_2^2}_{\text{L2正则}}$$

- 正则项也可以选L1正则，得到Lasso模型：

$$J(\mathbf{w}) = \sum_{i=1}^{N}\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2 + \lambda|\mathbf{w}|$$

  – 当λ取合适值时，Lasso（ least absolute shrinkage and selection operator）的结果是稀疏的（w的某些元素系数为0），起到特征选择作用。

▶ 为什么L_1正则的解是稀疏的?

- 考虑两个优化问题：

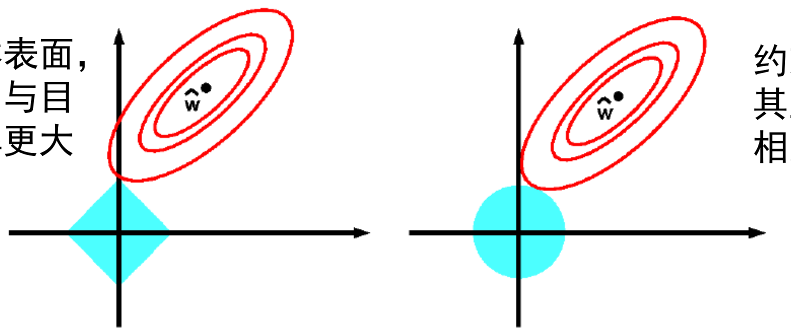
$$\min_{\mathbf{w}} RSS(\mathbf{w}) + \lambda \|\mathbf{w}\|_1 \qquad \min_{\mathbf{w}} RSS(\mathbf{w}) + \lambda \|\mathbf{w}\|_2^2$$

- 分别等价于（连续的带约束的优化问题）

$$\min_{\mathbf{w}} RSS(\mathbf{w}) \quad s.t. \quad \|\mathbf{w}\|_1 \le B \qquad \min_{\mathbf{w}} RSS(\mathbf{w}) \quad s.t. \quad \|\mathbf{w}\|_2^2 \le B$$

约束表面为立方体表面，由于角更突出，角与目标函数相交的概率更大

角：有些系数为0
→稀疏



约束表面为球形表面，其上各点与目标函数相交的概率相同

- 例：如 $\mathbf{w} = (1,0)^T$, $\left(1/\sqrt{2}, 1/\sqrt{2}\right)^T$ 的L_2 模相同(1)，但L_1模分别为1和$\sqrt{2}$

▶ 线性回归模型的概率解释

- 最小二乘（线性）回归等价于极大似然估计
- 正则（线性）回归等价于高斯先验（L2正则）或Laplace先验下（L1正则）的贝叶斯估计

▶ 最小二乘线性回归 等价于 极大似然估计

- 假设：$y = f(\mathbf{x}) + \varepsilon = \mathbf{w}^T \mathbf{x} + \varepsilon$
- 其中ε为线性预测和真值之间的残差

- 我们通常假设残差的分布为 $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ ，因此线性回归可写成：$p(y \mid \mathbf{x}, \boldsymbol{\theta}) \sim \mathcal{N}(y \mid \mathbf{w}^T \mathbf{x}, \sigma^2)$
- 其中 $\boldsymbol{\theta} = (\mathbf{w}, \sigma^2)$

▶ Recall：极大似然估计

- 极大似然估计（Maximize Likelihood Estimator, MLE）定义为

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log p(\mathcal{D} \mid \boldsymbol{\theta})$$

- 其中（log）似然函数为

$$l(\boldsymbol{\theta}) = \log p(\mathcal{D} \mid \boldsymbol{\theta}) = \sum_{i=1}^{N} \log p(y_i \mid x_i, \boldsymbol{\theta})$$

- 表示在参数为$\boldsymbol{\theta}$的情况下，数据 $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$ 出现的概率.
- 极大似然：选择数据出现概率最大的参数。

▶ 线性回归的MLE

$$p\left(y_i \mid \mathbf{x}_i, \mathbf{w}, \sigma^2\right) = \mathcal{N}\left(y_i \mid \mathbf{w}^T \mathbf{x}_i, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}\left(\left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2\right)\right)$$

- OLS的似然函数为

$$l(\boldsymbol{\theta}) = \log p(\mathcal{D} \mid \boldsymbol{\theta}) = \sum_{i=1}^{N} \log p\left(y_i \mid x_i, \boldsymbol{\theta}\right)$$

- 极大似然可等价地写成极小负log似然损失(negative log likelihood， NLL)

$$\begin{aligned}
NLL(\boldsymbol{\theta}) &= -\sum_{i=1}^{N} \log p\left(y_i \mid x_i, \boldsymbol{\theta}\right) \\
&= -\sum_{i=1}^{N} \log\left[\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}\left(\left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2\right)\right)\right] \\
&= \frac{N}{2}\log\left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2
\end{aligned}$$

▶ 正则回归等价于贝叶斯估计

- 假设残差的分布为 $\varepsilon \sim \mathcal{N}(0,\sigma^2)$，线性回归可写成：

$$p(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}) \sim \mathcal{N}(y_i \mid \mathbf{w}^T\mathbf{x}_i, \sigma^2)$$

$$p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{y} \mid \mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}_N) \propto \exp\left(-\frac{1}{2\sigma^2}\left[(\mathbf{y}-\mathbf{X}\mathbf{w})^T(\mathbf{y}-\mathbf{X}\mathbf{w})\right]\right)$$

- 若假设参数w的先验分布为 $w_j \sim \mathcal{N}(0,\tau^2)$
  - 偏向较小的系数值，从而得到的曲线也比较平滑

$$p(\mathbf{w}) = \prod_{j=1}^{D} \mathcal{N}(w_j \mid 0, \tau^2) \propto \exp\left(-\frac{1}{2\tau^2}\sum_{j=1}^{D}\mathbf{w}_j^2\right) = \exp\left(-\frac{1}{2\tau^2}\left[\mathbf{w}^T\mathbf{w}\right]\right)$$

  - 其中$1/\tau^2$ 控制先验的强度

▶ 正则回归等价于贝叶斯估计

- 根据贝叶斯公式，得到参数的后验分布为

$$p\left(\mathbf{w} \mid \mathbf{X}, \mathbf{y}, \sigma^2\right) \propto \exp\left(-\frac{1}{2\sigma^2}\left[\left(\mathbf{y}-\mathbf{X}\mathbf{w}\right)^T\left(\mathbf{y}-\mathbf{X}\mathbf{w}\right)\right]-\frac{1}{2\tau^2}\left[\mathbf{w}^T\mathbf{w}\right]\right)$$

- 则最大后验估计（MAP）等价于最小目标函数

$$J(\mathbf{w})=(\mathbf{y}-\mathbf{X}\mathbf{w})^T(\mathbf{y}-\mathbf{X}\mathbf{w})+\frac{\sigma^2}{\tau^2}\mathbf{w}^T\mathbf{w}$$

- 对比岭回归的目标函数

$$J(\mathbf{w})=\sum_{i=1}^{N}\left(y_i-\mathbf{w}^T\mathbf{x}_i\right)^2+\lambda\|\mathbf{w}\|_2^2$$

▶ 小结

- 线性回归模型也可以放到机器学习一般框架
  - 损失函数：L2损失…
  - 正则：无正则、L2正则、L1正则…
- 正则回归模型可视为先验为正则、似然为高斯分布的贝叶斯估计