线性回归

- 线性模型试图学得一个通过**特征的线性组合**来进行预测的函数，即$f(\mathbf{x})=\mathbf{w}^{\mathsf{T}}\mathbf{x}+b$
  由于$\mathbf{w}$直观表达了**各特征在预测中的重要性**，因此线性模型有很好的可解释性

- **模型假设**：$Y=X\beta+\varepsilon$
  - 自变量X与因变量Y是线性关系
  - 误差项ε独立同分布，服从$N(0,\sigma^2)$且独立于X
  - 自变量间不存在完全共线性

- **最大似然估计**：根据模型假设$\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$，似然函数可表示为

$$L(\beta)=\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{\varepsilon_i^2}{2\sigma^2}}=\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{\left(Y^{(i)}-X^{(i)}\beta\right)^2}{2\sigma^2}}=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{\sum_{i=1}^{n}\left(Y^{(i)}-X^{(i)}\beta\right)^2}{2\sigma^2}}$$

  那么$\hat{\beta}=\arg\max_{\beta} L(\beta)=\hat{\beta}=\arg\min_{\beta}\sum_{i=1}^{n}\left(Y^{(i)}-X^{(i)}\beta\right)^2$

- **最小二乘法**：试图找到一条直线或一个超平面，使所有样本到直线或平面上的欧式距离最小，即$\hat{\beta}=\arg\min_{\beta}(Y-X\beta)^{\mathsf{T}}(Y-X\beta)=(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}Y$
  矩阵的逆不存在时的解决方法：（1）减少特征个数（2）正则化

- **梯度下降法**：沿着损失函数梯度下降的方向对参数进行调整直到收敛，即重复以下步骤直到收敛：$\beta_{j+1}:=\beta_j-\alpha\dfrac{\partial J(\beta)}{\partial\beta_j}$，其中
  - α是学习率（Learning Rate），用来控制下降每步的距离（太小会收敛很慢，太大则会跳过最优点）
  - $J(\cdot)$是损失函数（Loss Function），用来衡量模型的预测精度

- 更一般地，考虑单调可微函数$g(\cdot)$，令$y=g^{-1}(\mathbf{w}^{\mathsf{T}}\mathbf{x}+b)$得到“广义线性模型”（Generalized Linear Model），其中函数$g(\cdot)$称为“联系函数”（Link Function）

- **优点**：线性模型简单直观，可解释性强；可通过正则化来防止过拟合
  缺点：处理非线性关系数据时不够灵活，拟合效果较差；模型拟合容易受到异常值或离群点的影响，需要进行预处理；模型假设较为严格