线性回归

机器学习的四个关键部分

- 数据（经验）：你有什么样的数据；
- 模型（假设）：对给定的问题有什么样的假设；
- 损失函数（目标）：如何评估一个模型；
- 优化函数（改善）：如何获取最优的模型。

线性回归

模型

$$y=f_\theta(x)=\theta_0+\sum_{j=1}^{d}\theta_jx_j=\theta^\top x, x=(1,x_1,x_2,\ldots,x_d)$$

这个x_i可以是非线性的，对于$y=a\cos(x)+b\sin(x)+c$，可以令$x_1=\cos(x), x_2=\sin(x)$得到线性方程。

目标/损失函数

目标：

$$\min_\theta \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}\left(y_i, f_\theta\left(x_i\right)\right)$$

最常使用均方误差（MSE）：

$$J_\theta=\frac{1}{2N}\sum_{i=1}^{N}\left(y_i-f_\theta\left(x_i\right)\right)^2 \quad \min_\theta J_\theta$$

优化

使用最常用的梯度下降：$\theta_{\text{new }}\leftarrow\theta_{\text{old }}-\eta\frac{\partial\mathcal{L}(\theta)}{\partial\theta}$

梯度下降可以细分为以下三种：

方法	损失函数	参数更新方程	优缺点
BGD	$J(\theta)=\frac{1}{2N}\sum_{i=1}^{N}\left(y_i-f_\theta\left(x_i\right)\right)^2 \quad \min_\theta J(\theta)$	$\theta_{\text{new }}=\theta_{\text{old }}+\eta\frac{1}{N}\sum_{i=1}^{N}\left(y_i-f_\theta\left(x_i\right)\right)x_i$	更稳定，但是收敛速度慢，容易获得局部最优解。
SGD	$J^{(i)}(\theta)=\frac{1}{2}\left(y_i-f_\theta\left(x_i\right)\right)^2 \quad \min_\theta \frac{1}{N}\sum_i J^{(i)}(\theta)$	$\theta_{\text{new }}=\theta_{\text{old }}+\eta\left(y_i-f_\theta\left(x_i\right)\right)x_i$	收敛速度快，但是可能会导致波动和不确定性。
MBGD	$J^{(k)}(\theta)=\frac{1}{2N_k}\sum_{i=1}^{N_k}\left(y_i-f_\theta\left(x_i\right)\right)^2$	$\theta_{\text{new }}=\theta_{\text{old }}-\eta\frac{\partial J^{(k)}(\theta)}{\partial\theta}$	结合了SGD和BGD的优点，并且适合并行处理。

学习率

学习率是一个需要用户选择的超参数。如果学习率太小，收敛速度太慢；如果太大，可能会波动甚至发散。

矩阵形式

目标：$J(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) \quad \min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$

梯度下降：$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})$

最优解：

$$
\begin{aligned}
\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0} &\Rightarrow \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) = \mathbf{0} \\
&\Rightarrow \boldsymbol{X}^\top\boldsymbol{y} = \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\theta} \\
&\Rightarrow \hat{\boldsymbol{\theta}} = \left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}
\end{aligned}
$$

正则化

有可能$\boldsymbol{X}^\top\boldsymbol{X}$是不可逆的，此时可以引入正则化：$J(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$

梯度下降：$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}$

最优解：

$$
\begin{aligned}
\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0} &\rightarrow -\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) + \lambda\boldsymbol{\theta} = \mathbf{0} \\
&\rightarrow \boldsymbol{X}^\top\boldsymbol{y} = \left(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}\right)\boldsymbol{\theta} \\
&\rightarrow \hat{\boldsymbol{\theta}} = \left(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}
\end{aligned}
$$