Chapter6有监督学习方法

定义

有监督学习Supervised Learning：从**有标记**的训练数据中学习推断函数。（既有特征x又有标签y）

目标函数(target function)：$y = f(x)$或$P(y|x)$

主要方法：Generative Model（GM），Discriminative Model（DM），Discriminative Function(DF)

1. **有监督学习方法一: 产生式模型 Generative Model**
   - 使用联合分布进行推断

$$p(x, y) = p(y)p(x|y)$$

> 💡 联合分布：
>
> 联合概率（Joint Probability）是指两个或多个随机事件同时发生的概率。在概率论中，联合概率用来描述事件之间的关系。

   - 使用贝叶斯定理来计算条件分布 $p(y|x)$

$$p(y|x) = \frac{p(x, y)}{p(x)} = \frac{p(x|y)p(y)}{\int p(x|y)p(y)dy}$$

   - 利用条件概率密度来预测
2. **有监督学习方法二: 判别式模型Discriminative Model**
   - 直接估计条件概率 $P(y|x)$或条件概率密度函数 $p(y|x)$。
   - 根据估计的函数确定输出 。
3. **有监督学习方法三: 判别函数Discriminative Function**

- 寻找一个函数 $f(x)$， 将每个输入直接映射到目标输出。
- 在其中，概率不起直接作用。
  - 不能直接获取后验概率。

- f通常旨在近似条件分布 $p(y|x)$

判别式模型和生成式模型都是使后验概率最大化，判别式是直接对后验概率建模，而生成式模型通过贝叶斯定理这一“桥梁”使问题转化为求联合概率。

回归任务

1. 线性回归→参考**统计机器学习note**

■输入: N个 i.i.d 训练样本 $(\boldsymbol{x}^i, y^i) \in X \times R,\ i = 1,2,...,N$

■目标函数: $f \in \mathcal{F}$

■损失函数: $L(f;x,y) = (f(x) - y)^2$

■期望风险: $\int (f(x) - y)^2 dP(x,y)$

■如果f是线性函数，最优化问题为: $\min\limits_{\boldsymbol{w}} J(\boldsymbol{w}) = \sum\limits_{i=1}^{N} (\boldsymbol{w}^T \boldsymbol{x}^i - y^i)^2$

损失函数最优解推导

$$w^* = (X^T X)^{-1} X^T Y$$

2. 最小二乘/均方误差(Least Mean Squares Algorithm)

■最优化问题: $\min\limits_{\boldsymbol{w}} J(\boldsymbol{w}) = \sum\limits_{i=1}^{N} (\boldsymbol{w}^T \boldsymbol{x}^i - y^i)^2$

■梯度下降: $\dfrac{\partial J(\boldsymbol{w})}{\partial w_j} = 2\sum\limits_{i=1}^{N} x_j^i (\boldsymbol{w}^T \boldsymbol{x}^i - y^i)$

- 批梯度下降BGD：Batch Gradient Descent

$$w_j = w_j - 2\alpha \sum_{i=1}^{N} (\boldsymbol{w}^T \boldsymbol{x}^i - y^i) x_j^i,\ \alpha > 0$$

批梯度下降，**BGD**
Batch Gradient Descent

学习率（可动态调整）

优点：

一次迭代是对所有样本进行计算，此时利用矩阵进行操作，实现了并行。

由全数据集确定的方向能够更好地代表样本总体，从而更准确地朝向极值所 在的方向。当目标函数为凸函数时，BGD一定能够得到全局最优。

缺点：

当样本数目 *N*很大时，每迭代一步都需要对所有样本计算，训练过程会很慢

- 随机梯度下降SGD：Stochastic Gradient Descent

$$w_j = w_j - 2\alpha(\boldsymbol{w}^T\boldsymbol{x}^i - y^i)x_j^i,\ \alpha > 0$$

随机梯度下降，SGD
Stochastic Gradient Descent

3. 利用非线性基进行线性回归(广义线性回归)

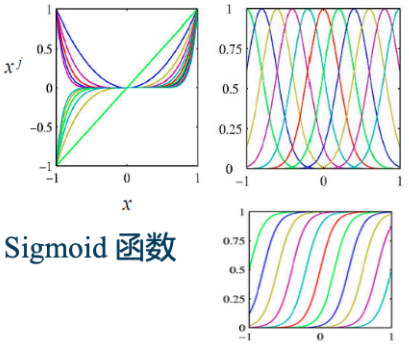
■对非线性基进行线性组合：

$$f(\boldsymbol{w},\boldsymbol{x}) = w_0 + \sum_{j=1}^{K} w_j\phi_j(\boldsymbol{x})$$

$$\Phi = (1,\phi_1,...,\phi_K)$$

■非线性基函数

- $\phi(\boldsymbol{x}) = (1,x,x^2,...,x^K)$ 多项式基函数
- $\phi_j(\boldsymbol{x}) = \exp\left(-\frac{(x-\mu_j)^2}{2s^2}\right)$ 高斯函数
- $\phi_j(\boldsymbol{x}) = \sigma\left(\frac{x-\mu_j}{s}\right), \sigma(a) = \frac{1}{1+\exp(-a)}$ Sigmoid 函数



可以求得闭式解，如线性回归方法

■最优化问题: $\min_{w} J(w)=\sum_{i=1}^{N}(w^T\phi(x^i)-y^i)^2$

■梯度: $\frac{\partial J(w)}{\partial w_j}=2\sum_{i=1}^{N}\phi_j(x^i)(w^T\phi(x^i)-y^i)$

■闭式解: $w^*=(\Phi^T\Phi)^{-1}\Phi^T y$

其中 $\Phi=\begin{pmatrix}\phi_0(x^1) & \dots & \phi_K(x^1)\\ \vdots & \vdots & \vdots \\ \phi_0(x^N) & \dots & \phi_K(x^N)\end{pmatrix}, y=(y^1,\dots,y^N)^T$

4. 最大似然估计 Maximum Likelihood Estimation MLE

$$p(y\mid x,w,\beta)=\mathcal{N}(y\mid f(x,w),\beta^{-1})$$

$$\prod_{i=1}^{N}\mathcal{N}(y^i\mid w^T x^i,\beta^{-1})$$

$$\hat{x}_{MLE}(y)=\underset{x}{ARGmax}\ f_Y(y|x)$$

- in which the f is the pdf of diatribution of y
- 似然和概率：

  概率是在特定环境下某件事情发生的可能性，也就是结果没有产生之前依据环境所对应的参数来预测某件事情发生的可能性

  似然是在确定的结果下去推测产生这个结果的可能环境（参数）
- 结论：在高斯噪声模型下，**最小化平方误差与最大似然的解相同**
- 推导

对数似然 $\sum_{i=1}^{N}\ln\mathcal{N}(y^i\mid w^T x^i,\beta^{-1})=\frac{N}{2}\ln\beta-\frac{N}{2}\ln 2\pi-\frac{1}{2}\beta J(w)$

其中 $J(w)=\sum_{i=1}^{N}(w^T x^i-y^i)^2$

5. 最大化后验概率Maximum A Posteriori MAP

最大后验估计方法是在MLE的基础上引入了先验信息，即在进行参数估计时，不仅依赖数据，还结合了模型的先验分布。

$$\hat{x}_{maxPrior} = \underset{X}{ARGmax}\ f_x(x)$$

$$\hat{x}_{MAP} = \underset{x}{ARGmax}\ f_x(x|y) = \underset{x}{ARGmax}\ \frac{f_Y(y|x)f_X(x)}{f_Y(y)}$$

for the $f_Y(y)$, $f_Y(y) = \int_X f_Y(y|x)f_X(x)dx$. It does not depend on x. Thus we can ignore it. The difference between the MLE and MAP is $f_X(x)$

■贝叶斯定理: $p(\boldsymbol{w} \mid \boldsymbol{y}) = p(\boldsymbol{y} \mid \boldsymbol{w})p(\boldsymbol{w}) / p(\boldsymbol{y})$

●似然函数: $p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{i=1}^{N} \mathcal{N}(y^i \mid \boldsymbol{w}^T \boldsymbol{x}^i, \beta^{-1})$

●先验: $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{0}, \alpha^{-1}\boldsymbol{I})$

■后验概率依然是高斯分布，对后验取对数:

$$\ln(p(\boldsymbol{w} \mid \boldsymbol{y})) = -\beta \sum_{i=1}^{N} (y^i - \boldsymbol{w}^T \boldsymbol{x}^i)^2 - \lambda \boldsymbol{w}^T \boldsymbol{w} + constant$$

■最大化后验等同于最小化带有正则项的平方和误差

$$\min_{\boldsymbol{w}} \sum_{i=1}^{N} (\boldsymbol{w}^T \boldsymbol{x}^i - y^i)^2 + \lambda \boldsymbol{w}^T \boldsymbol{w},\ \lambda = \frac{\alpha}{\beta}$$

结论：最大化后验等同于最小化带有正则项的平方和误差

推导：带有正则项的平方和误差

推导：最大化后验

6. 总结

最大似然估计（MLE）可以看作是使用均匀分布作为先验的最大后验估计（MAP）。

MLE	MAP
判别模型	产生模型

MLE	MAP
频率学派	贝叶斯学派

MLE：

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

MAP:

$$\hat{\theta}_{MAP} = \text{argmax}_{\theta} P(\theta|D) = \text{argmax}_{\theta} P(D|\theta)P(\theta)$$

分类任务

■输入: N i.i.d 训练样本 $(\boldsymbol{x}^i, y^i) \in X \times C, \quad i = 1,2,...,N$

■目标函数：$f \in \mathcal{F}$

■损失函数：$L(f; x, y) = I_{\{f(\boldsymbol{x}) \neq y\}}$

■期望风险（损失）：$\int I_{\{f(\boldsymbol{x}) \neq y\}} dP(\boldsymbol{x}, y) = P(f(\boldsymbol{x}) \neq y)$

在回归任务中，直接使用传统的回归损失函数（如**L2损失**）来求解时，可能会遇到**异常值（outliers）**问题

判别函数法

判别式模型

Logistic 回归

是最常见的判别式模型之一。它通过建模输入特征 x 与标签 y 之间的条件概率关系来进行分类任务。

> Logistic 回归用于二分类问题，目标是通过一个输入向量 $\mathbf{x}$（或称模式向量）来预测一个二分类输出 $y \in \{0, 1\}$。其基本思想是将线性回归模型的输出映射到一个概率值区间 [0, 1]，使其能够表示为一个分类概率。

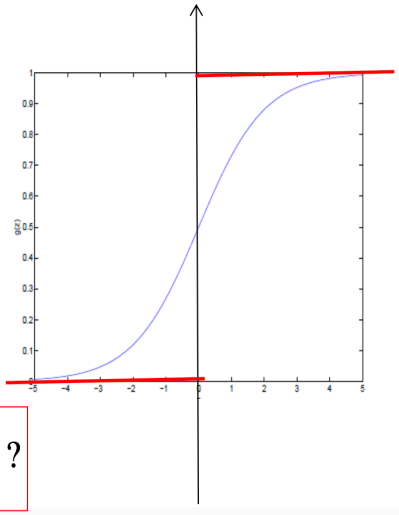
■估计后验概率 $p(y|\boldsymbol{x})$

$$P(y=1|\boldsymbol{x})=f(\boldsymbol{x},\boldsymbol{w})=g(\boldsymbol{w}^T\boldsymbol{x})=\frac{1}{1+\exp(-\boldsymbol{w}^T\boldsymbol{x})}$$

Logistic function

$$g(z)=\frac{1}{1+e^{-z}}$$

Sigmoid function



■ $g(z)$的性质

- $z\to\infty$ 时, $g(z)\to 1$
- $z\to-\infty$ 时, $g(z)\to 0$
- $g(z)$的0-1之间
- $g'(z)=g(z)(1-g(z))$

$$\ln\frac{P(y=1|\boldsymbol{x})}{1-P(y=1|\boldsymbol{x})}=?$$

求解：最大似然估计(Maximum Likelihood Estimator）MLE

- 概率分布：$P(y|\boldsymbol{x},\boldsymbol{w})=(f(\boldsymbol{x},\boldsymbol{w}))^y(1-f(\boldsymbol{x},\boldsymbol{w}))^{1-y}$ *Bernoulli*
- 似然:

$$L(\boldsymbol{w})=\prod_{i=1}^{N}P(y^i|\boldsymbol{x}^i,\boldsymbol{w})=\prod_{i=1}^{N}\left(f(\boldsymbol{x}^i,\boldsymbol{w})\right)^{y^i}\left(1-f(\boldsymbol{x}^i,\boldsymbol{w})\right)^{1-y^i}$$

- 最大化log 似然:

$$l(\boldsymbol{w})=\log L(\boldsymbol{w})=\sum_{i=1}^{N}\left(y^i\log f(\boldsymbol{x}^i,\boldsymbol{w})+(1-y^i)\log\left(1-f(\boldsymbol{x}^i,\boldsymbol{w})\right)\right)$$

- 梯度:

$$\frac{\partial l(\boldsymbol{w})}{\partial w_j}=(y^i-f(\boldsymbol{x}^i,\boldsymbol{w}))x_j^i,\forall(\boldsymbol{x}^i,y^i)$$

- SGD:

$$w_j=w_j+\alpha(y^i-f(\boldsymbol{x}^i,\boldsymbol{w}))x_j^i$$

$$\mu(\boldsymbol{x}) = \sigma\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}\right)$$

$$p(y|\boldsymbol{x};\mu) = \mu(\boldsymbol{x})^{y}(1-\mu(\boldsymbol{x}))^{(1-y)}$$

多类logistic回归

■Softmax 函数取代logistic sigmoid: $P(C_k \mid \boldsymbol{w}, \boldsymbol{x}) = \dfrac{\exp(\boldsymbol{w}_k^T \boldsymbol{x})}{\sum_{j=1}^{K} \exp(\boldsymbol{w}_j^T \boldsymbol{x})}$

这里 $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_K$ 是待学习的参数。

■y可以看作是取K值之一的离散变量

- 将y表示为K维的向量
- 如果 y =3, 那么$\boldsymbol{y}$=(0,0,1,0,…,0)（第三个元素为1，其它元素为0）
- K维向量满足: $\sum_{i=1}^{K} P(y_i \mid \boldsymbol{w}, \boldsymbol{x}) = 1$

One hot 表示
独热表示

■概率分布: $\mu_i = P(y_i = 1 \mid \boldsymbol{w}, \boldsymbol{x})$

$$P(\boldsymbol{y} \mid \boldsymbol{\mu}) = \prod_{i=1}^{K} \mu_i^{y_i}$$

Generalized Bernoulli
广义伯努里分布