统计学习

李 波

统计学习

贝叶斯准则

- 人类往往会根据自身的经验，形成一些对世界的假设(hypothesis)；
- 令 e 为经验（evidence），即为训练数据，H 为假设.；
- 根据经验 e，假设为 H 的概率为

$$p(H|e)=\frac{p(H\cap e)}{p(e)}=\frac{p(H)p(e|H)}{p(e)}$$

假设空间(Hypothesis space)

- 所有关于模型可能假设的集合;
- 令 e 为经验（evidence），即为训练数据，H 为假设;
- 根据经验 e，假设为 H 的概率为

$$p(H|e)=\frac{p(H\cap e)}{p(e)}=\frac{p(H)p(e|H)}{p(e)}$$

统计学习

贝叶斯准则

$$p(H|e) = \frac{p(H \cap e)}{p(e)} = \frac{\overbrace{p(H)}^{\text{先验概率}}\overbrace{p(e|H)}^{\text{似然概率}}}{\underbrace{p(e)}_{\text{经验概率}}}$$

- $p(H)$ 为先验概率(prior): 在没有经验 e 的时候，假设H的概率；
- $p(e|H)$ 为似然概率: 在假设H成立的前提下，经验发生的概率；
- $p(e)$ 为经验发生的概率。

统计学习

经验，即训练数据，为 $e=\{(\boldsymbol{x}_1,y_1),\dots(\boldsymbol{x}_n,y_n)\}$。对于一个预测样本 $\boldsymbol{x}$，如何根据经验（训练数据）预测样本的标签?

$$
\begin{aligned}
p(y|e,\boldsymbol{x}) &= \sum_{i=1}^{k} p(y,H_i|e,\boldsymbol{x}) \\
&= \sum_{i=1}^{k} p(y|H_i,e,\boldsymbol{x})p(H_i|e,\boldsymbol{x}) \\
&= \sum_{i=1}^{k} p(y|H_i,\boldsymbol{x})p(H_i|e)
\end{aligned}
$$

统计学习

第i个假设的后验概率

给定第 i 个假设和训练数据，预测标签为y的概率

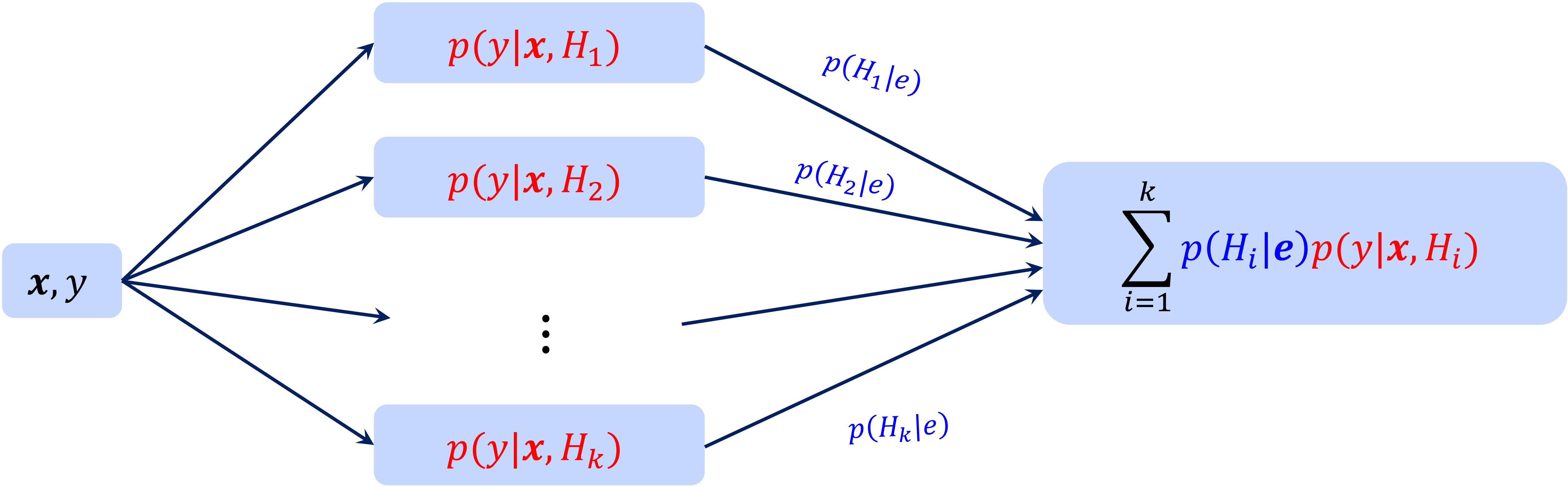
$$p(y|e,\boldsymbol{x}) = \sum_{i=1}^{k} p(H_i|e)p(y|H_i,\boldsymbol{x})$$

- 预测的标签概率为各假设条件下预测标签概率的加权平均值；
- 假设作为经验（或训练数据）与预测标签的一个中间量。

From CS480/680 Spring 2019 Pascal Poupart University of Waterloo

统计学习

$$p(y|e,\boldsymbol{x}) = \sum_{i=1}^{k} p(H_i|e)p(y|H_i,\boldsymbol{x})$$



统计学习

$$p(y|e,\boldsymbol{x}) = \sum_{i=1}^{k} p(H_i|e)p(y|H_i,\boldsymbol{x})$$

$$p(H_i|e) = \frac{p(H_i,e)}{p(e)} = \frac{p(H_i,e)}{\sum_{j=1}^{k} p(H_i,e)} = \frac{p(e|H_i)p(H_i)}{\sum_{j=1}^{k} p(e|H_i)p(H_i)}$$

From CS480/680 Spring 2019 Pascal Poupart University of Waterloo

统计学习

例子

- 有两种糖，分别是奶糖、巧克力；
- 有两种糖纸，分别是绿色、红色；
- 两种糖混合在一起，不知道那种糖是什么颜色糖纸；
- 希望通过糖纸颜色判断糖的种类。

统计学习

列出如下假设

- $\mathbf{H_0}$

$$\begin{cases} p(y=\text{奶糖}|x=\text{红色}\,)=0.7 \\ p(y=\text{巧克力}|x=\text{红色})=0.3 \end{cases} \qquad \begin{cases} p(y=\text{奶糖}|x=\text{绿色}\,)=0.3 \\ p(y=\text{巧克力}|x=\text{绿色})=0.7 \end{cases}$$

- $\mathbf{H_1}$

$$\begin{cases} p(y=\text{奶糖}|x=\text{红色}\,)=0.3 \\ p(y=\text{巧克力}|x=\text{红色})=0.7 \end{cases} \qquad \begin{cases} p(y=\text{奶糖}|x=\text{绿色})=0.7 \\ p(y=\text{巧克力}|x=\text{绿色})=0.3 \end{cases}$$

先验概率为

$$p(H_0)=0.1\,,p(H_1)=0.9$$

统计学习

- H0: 红色糖大多(70%)都是奶糖，绿色糖大多(70%)都是巧克力。
- H1: 红色糖大多(70%)都是巧克力，绿色糖大多(70%)都是奶糖，

先验概率为 $p(H_0) = 0.1$, $p(H_1) = 0.9$

随机拿 n 块糖，并打开，可以形成如下经验集合或者训练样本集合

$$e = \{(x_1, y_1), \ldots (x_n, y_n)\}$$

, 其中 x_i, y_i分别代表糖纸颜色和糖类别。下表给出了经验集合的统计数据。

(红色, 奶糖)	(红色, 巧克力)	(绿色, 奶糖)	(绿色, 巧克力)
6	4	5	5

统计学习

- H0: 红色糖大多(70%)都是奶糖，绿色糖大多(70%)都是巧克力。
- H1: 红色糖大多(70%)都是巧克力，绿色糖大多(70%)都是奶糖。

先验概率为 $p(H_0) = 0.1$, $p(H_1) = 0.9$

(红色, 奶糖)	(红色, 巧克力)	(绿色, 奶糖)	(绿色, 巧克力)
6	4	5	5

- 根据先验概率和经验数据，哪种假设成立可能性大?
- 如果选一块糖是**红色**，这块糖是奶糖还是巧克力?

- H0: 红色糖大多(70%)都是奶糖，绿色糖大多(70%)都是巧克力
- H1: 红色糖大多(70%)都是巧克力，绿色糖大多(70%)都是奶糖

先验概率为 $p(H_0) = 0.1$, $p(H_1) = 0.9$

(红色, 奶糖)	(红色, 巧克力)	(绿色, 奶糖)	(绿色, 巧克力)
6	4	5	5

$$p(e|H_0) = 0.7^6 \times 0.3^4 \times 0.3^5 \times 0.7^5 \approx 3.89 \times 10^{-7}$$

$$p(e|H_1) = 0.3^6 \times 0.7^4 \times 0.7^5 \times 0.3^5 \approx 7.15 \times 10^{-8}$$

$$\begin{aligned} p(e) &= p(H_0)p(H_0|e) + p(H_1)p(H_1|e) \\ &= 0.1 \times 3.89 \times 10^{-7} + 0.9 \times 7.15 \times 10^{-8} \approx 1.03 \times 10^{-7} \end{aligned}$$

$$p(H_0|e) = \frac{p(H_0)p(H_0|e)}{p(e)} = \frac{0.1 \times 3.89 \times 10^{-7}}{1.03 \times 10^{-7}} \approx 0.38$$

$$p(H_1|e) = \frac{p(H_1)p(H_1|e)}{p(e)} = \frac{0.9 \times 7.15 \times 10^{-8}}{1.03 \times 10^{-7}} \approx 0.62$$

- H0: 红色糖大多(70%)都是奶糖，绿色糖大多(70%)都是巧克力
- H1: 红色糖大多(70%)都是巧克力，绿色糖大多(70%)都是奶糖

先验概率为 $p(H_0) = 0.1$, $p(H_1) = 0.9$

(红色, 奶糖)	(红色, 巧克力)	(绿色, 奶糖)	(绿色, 巧克力)
6	4	5	5

$$p(y|e,\boldsymbol{x}) = \sum_{i=1}^{k} p(H_i|e)p(y|H_i,\boldsymbol{x})$$

如果 $y =$ 奶糖

$$p(y = \text{奶糖}|x = \text{红色}, H_0) = 0.7 \qquad p(y = \text{奶糖}|x = \text{红色}, H_1) = 0.3$$

$$p(y = \text{奶糖}|x = \text{红色}, e) = p(H_0|e)p(y = \text{奶糖}|x = \text{红色}, H_0) + p(H_1|e)p(y = \text{奶糖}|x = \text{红色}, H_1)$$
$$= 0.38 \times 0.7 + 0.62 \times 0.3 \approx 0.45$$

- H0: 红色糖大多(70%)都是奶糖，绿色糖大多(70%)都是巧克力
- H1: 红色糖大多(70%)都是巧克力，绿色糖大多(70%)都是奶糖

先验概率为 $p(H_0) = 0.1$, $p(H_1) = 0.9$

(红色, 奶糖)	(红色, 巧克力)	(绿色, 奶糖)	(绿色, 巧克力)
6	4	5	5

$$p(y|e,\boldsymbol{x}) = \sum_{i=1}^{k} p(H_i|e)p(y|H_i,\boldsymbol{x})$$

如果 $y =$ 巧克力

$p(y = \text{巧克力}|x = \text{红色}, H_0) = 0.3$

$p(y = \text{巧克力}|x = \text{红色}, H_1) = 0.7$

$$p(y = \text{巧克力}|x = \text{红色}, e) = p(H_0|e)p(y = \text{巧克力}|x = \text{红色}, H_0) + p(H_1|e)p(y = \text{巧克力}|x = \text{红色}, H_1)$$
$$= 0.38 \times 0.3 + 0.62 \times 0.7 \approx 0.55$$

- H0: 红色糖大多(70%)都是奶糖，绿色糖大多(70%)都是巧克力
- H1: 红色糖大多(70%)都是巧克力，绿色糖大多(70%)都是奶糖

先验概率为 $p(H_0) = 0.1$, $p(H_1) = 0.9$

(红色, 奶糖)	(红色, 巧克力)	(绿色, 奶糖)	(绿色, 巧克力)
6	4	5	5

$$p(y|e,\boldsymbol{x}) = \sum_{i=1}^{k} p(H_i|e)p(y|H_i,\boldsymbol{x})$$

$p(y = \text{奶糖}|e, x = \text{红色}) = 0.45$ $p(y = \text{巧克力}|e, x = \text{红色}) = 0.55$

$p(y = \text{奶糖}|e, x = \text{红色}) < p(y = \text{巧克力}|e, x = \text{红色})$ → **预测标签为巧克力**

统计学习

$$p(y|e,\boldsymbol{x}) = \sum_{i=1}^{k} p(H_i|e)p(y|H_i,\boldsymbol{x})$$

$x = $ 红色, $y = $ 奶糖

$p(y = \text{奶糖}|x = \text{红色}, H_0) = 0.7$

$p(y = \text{奶糖}|x = \text{红色}, H_1) = 0.3$

$p(H_0|e) = 0.38$

$p(H_1|e) = 0.62$

$0.38 \times 0.7 + 0.62 \times 0.3 \approx 0.45$

统计学习

$$p(y|e,\boldsymbol{x}) = \sum_{i=1}^{k} p(H_i|e)p(y|H_i,\boldsymbol{x})$$

$x =$ 红色,
$y =$ 巧克力

$p(y = \text{巧克力}|x = \text{红色}, H_0) = 0.3$

$p(H_0|e) = 0.38$

$p(y = \text{巧克力}|x = \text{红色}, H_1) = 0.7$

$p(H_1|e) = 0.62$

$0.38 \times 0.3 + 0.62 \times 0.7 \approx 0.55$

贝叶斯预测

- **最优**：如果先验给定，没有更准确的预测方法；
- **不存在过拟合**：所有可能假设（hypothesis）都考虑到并加权；
- **代价**：
  - 当假设个数太多时，贝叶斯预测无法进行。比如，对假设求和（或求积分）无法计算；
  - 解决方案：近似贝叶斯预测。

最大后验(Maximum a posteriori, MAP)预测

- 选择一个最可能的假设（hypothesis）做预测

贝叶斯预测: 计算所有假设预测的加权平均，每种假设的概率作为权重

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, p(y|e,x) = \underset{y}{\operatorname{argmax}} \sum_{i=1}^{k} p(H_i|e)p(y|H_i,\boldsymbol{x})$$

最大后验预测: 利用最可能的假设，对输入做预测

$$H = \underset{\{H_i\}}{\operatorname{argmax}}\, p(H_i|e) \quad \rightarrow \quad \hat{y} = \underset{\{y\}}{\operatorname{argmax}}\, p(y|x,H)$$

统计学习

$$p(y|e,\boldsymbol{x}) = \sum_{i=1}^{k} p(H_i|e)p(y|H_i,\boldsymbol{x})$$

$x=$ 红色,
$y=$ 巧克力

$p(y=\text{巧克力}|x=\text{红色},H_0)=0.3$

$p(H_0|e)=0.38$

$p(y=\text{巧克力}|x=\text{红色},H_1)=0.7$

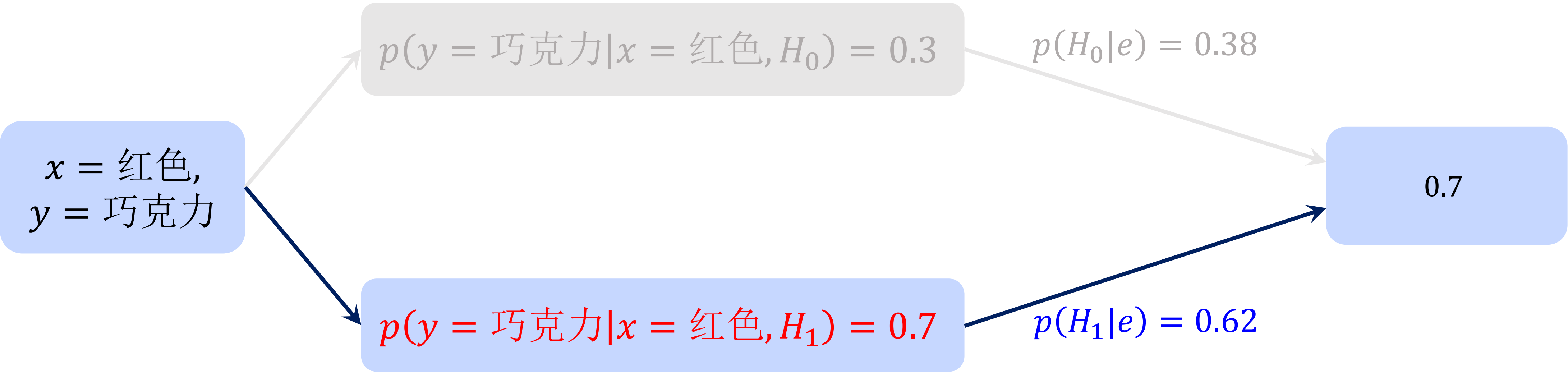
$p(H_1|e)=0.62$

$0.38\times0.3+0.62\times0.7\approx0.55$

- 因为 $p(H_1|e) > p(H_0|e)$， 所以利用H_1假设做预测。

统计学习

$$p(y|e,\boldsymbol{x}) = \sum_{i=1}^{k} p(H_i|e)p(y|H_i,\boldsymbol{x})$$



- 因为 $p(H_1)p(e|H_1) > p(H_0)p(e|H_0)$， 所以利用$H_1$假设做预测；
- 根据H_1假设，预测结果为巧克力的概率为0.7。

最大后验(MAP)预测

- 由于只利用一个假设（hypothesis），MAP 预测没有贝叶斯预测准确;
- 但是当数据量很大时，最大后验 预测接近贝叶斯预测;
- 控制过拟合: 先验概率可以用于降低假设（模型）的复杂程度;
- 不容易确定假设的先验概率;
- 优化问题可能难以解决。

最大似然(Maximum Likelihood, ML)预测

- 在最大后验预测基础上，假设先验概率均匀分布

贝叶斯预测: 计算所有假设预测的加权平均，每种假设的概率作为权重

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, p(y|e, x) = \underset{y}{\operatorname{argmax}} \sum_{i=1}^{k} p(H_i|e) p(y|H_i, \boldsymbol{x})$$

最大后验预测: 利用最可能的假设，对输入做预测

$$H = \underset{H_i}{\operatorname{argmax}}\, p(H_i) p(e|H_i) \rightarrow \hat{y} = \underset{y}{\operatorname{argmax}}\, p(y|x, H)$$

最大似然预测: 先验概率均匀分布，利用最可能的假设，对输入做预测

$$H = \underset{H_i}{\operatorname{argmax}}\, p(e|H_i) \rightarrow \hat{y} = \underset{y}{\operatorname{argmax}}\, p(y|x, H)$$

统计学习

最大似然(maximum likelihood, ML)

- 对MAP测进行简化，假设**先验概率为均匀分布。**；
- 相比于MAP，ML预测选择假设的标准变了：MAP利用先验概率和似然选择假设，而**ML只使用似然选择假设**；
- ML预测没有贝叶斯预测和MAP预测准确，因为ML预测忽略了假设（hypothesis）的先验信息，而且只根据一个假设做预测；
- 但当数据量足够大时，ML、MAP、贝叶斯预测都趋于相同；
- ML预测可能会过拟合，因为缺少先验概率。先验概率提供经验数据之外关于假设的信息；
- 相比MAP预测，ML预测更容易找到一个最优假设。

统计学习

- 在线性回归模型中，特征$\boldsymbol{x}$和标签y存在如下关系

$$y = \boldsymbol{\omega}^T \boldsymbol{x} + e$$

- e服从高斯分布，即 $e \sim \mathcal{N}(0, \sigma^2)$. 似然函数为

$$f_{X|Y}(\boldsymbol{x}|y) = \mathcal{N}(y - \boldsymbol{\omega}^T \boldsymbol{x}, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y - \boldsymbol{\omega}^T \boldsymbol{x})^2}{2\sigma^2}\right\}$$

- 对于n个训练数据 $(\boldsymbol{x}_1, \mathrm{y}_1), (\boldsymbol{x}_2, \mathrm{y}_2), \cdots, (\boldsymbol{x}_n, \mathrm{y}_n)$, 似然函数为

$$f_{Y|X}(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n | \mathrm{y}_1, \mathrm{y}_2, \cdots, \mathrm{y}_n) = \prod_{i=1}^{n} f_N(\boldsymbol{x}_i | y_i) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\left\{-\frac{\sum_{i=1}^{n}(y_i - \boldsymbol{\omega}^T \boldsymbol{x}_i)^2}{2\sigma^2}\right\}$$

统计学习

$$f_{Y|X}\left(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\,|\mathrm{y}_1,\mathrm{y}_2,\cdots,\mathrm{y}_n\right)=\prod_{i=1}^{n}f_{Y|X}(\boldsymbol{x}_i|y_i)=\frac{1}{\left(\sqrt{2\pi}\sigma\right)^n}\exp\left\{-\frac{\sum_{i=1}^{n}(y_i-\boldsymbol{\omega}^T\boldsymbol{x}_i)^2}{2\sigma^2}\right\}$$

- 利用最大似然准则估计模型参数

$$\max_{\boldsymbol{\omega}} f_{X|Y}\left(\mathrm{y}_1,\mathrm{y}_2,\cdots,\mathrm{y}_n\,|\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\right)$$

$$或\quad \min_{\boldsymbol{\omega}} \sum_{i=1}^{n}(y_i-\boldsymbol{\omega}^T\boldsymbol{x}_i)^2$$

$$或\quad \min_{\boldsymbol{\omega}} \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\omega}\|^2$$

- 上述优化问题为线性回归的优化问题，模型为**线性回归**；
- 可以看出，线性回归模型的损失函数来源于噪声服从高斯分布。

统计学习

$$f_{Y|X}(y_1, y_2, \cdots, y_n | \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n) = \prod_{i=1}^{n} f_N(y_i | \boldsymbol{x}_i) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\left\{-\frac{\sum_{i=1}^{n}(y_i - \boldsymbol{\omega}^T \boldsymbol{x}_i)^2}{2\sigma^2}\right\}$$

- 假设$\boldsymbol{\omega}$的先验信息为均值为0方差为 σ_0^2 的高斯分布，即

$$f_W(\boldsymbol{\omega}) = \mathcal{N}(0, \sigma_0^2) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{\|\boldsymbol{\omega}\|^2}{2\sigma_0^2}\right\}$$

- $\boldsymbol{\omega}$的最大后验估计为

$$\max_{\boldsymbol{\omega}} f_W(\boldsymbol{\omega}) f_{Y|X}(y_1, y_2, \cdots, y_n | \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)$$

或 $$\min_{\boldsymbol{\omega}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\omega}\|^2 + \frac{\sigma^2}{\sigma_0^2} \|\boldsymbol{\omega}\|^2$$

- 上述优化问题为带正则项的线性回归问题或岭回归问题，其中 $\lambda = \frac{\sigma^2}{\sigma_0^2}$；
- 可以看出，岭回归问题中损失函数中的正则项对应于参数先验分布为高斯分布。

统计学习

- 假设$\boldsymbol{\omega}$的先验信息为均值为$\boldsymbol{\mu}$方差为 σ_0^2 的高斯分布，即

$$f_W(\boldsymbol{\omega}) = \mathcal{N}(0, \sigma_0^2) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{\|\boldsymbol{\omega}-\boldsymbol{\mu}\|^2}{2\sigma_0^2}\right\}$$

- $\boldsymbol{\omega}$的最大后验估计为解如下优化问题

$$\max_{\boldsymbol{\omega}} f_W(\boldsymbol{\omega}) f_N(\mathrm{y}_1, \mathrm{y}_2, \cdots, \mathrm{y}_n \,|\, \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)$$

$$\text{或} \quad \min_{\boldsymbol{\omega}} \frac{1}{\sigma^2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\omega}\|^2 + \frac{1}{\sigma_0^2}\|\boldsymbol{\omega}-\boldsymbol{\mu}\|^2$$

- 当$\boldsymbol{\mu} = \mathbf{0}$上述优化问题为带正则项的线性回归问题或岭回归问题，其中 $\lambda = \frac{\sigma^2}{\sigma_0^2}$；
- 解上述优化问题，可得$\boldsymbol{\omega}$的最大后验估计为

$$\begin{aligned}\boldsymbol{\omega} &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}(\boldsymbol{X}^T\boldsymbol{y} + \lambda\boldsymbol{\mu}) \\ &= (\boldsymbol{I} + \lambda(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} + (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\mu}\end{aligned}$$

统计学习

- $\boldsymbol{\omega}$的最大后验估计

$$\boldsymbol{\omega} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}(\boldsymbol{X}^T\boldsymbol{y} + \lambda\boldsymbol{\mu})$$
$$= (\boldsymbol{I} + \lambda(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}\underbrace{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}}_{\text{最大似然估计}} + (\boldsymbol{X}^T\boldsymbol{X}/\lambda + \boldsymbol{I})^{-1}\underbrace{\boldsymbol{\mu}}_{\text{先验估计}}$$

- 从上式可以看出，最大后验估计值是最大似然估计值和先验估计的“加权”和，权重为一矩阵；
- 如果先验估计为0，即$\boldsymbol{\mu} = \boldsymbol{0}$，最大后验估计为最大似然估计的线性变换；
- $\lambda = \frac{\sigma^2}{\sigma_0^2}$，如果$\sigma^2 \to +\infty$，那么$\lambda \to +\infty$，$\boldsymbol{\omega} = \boldsymbol{\mu}$；
- 如果$\sigma_0^2 \to +\infty$，那么$\lambda \to 0$，$\boldsymbol{\omega} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$。