

Machine

Learning - P12

最大后验概率 (maximum a posteriori, MAP), 一种参数估计方法

对于 n 个服从独立同分布的样本的训练集 $D = \{x_1, x_2, \dots, x_n\}$

对于 $x_i, i=1, 2, \dots, n$, 有概率密度函数 $p(x_i | \theta)$

则对概率密度函数 $p(x_i | \theta)$ 中的参数 θ 进行估计

$$\begin{aligned}\text{即 } \theta_{\text{MAP}} &= \operatorname{argmax}_{\theta} \Pr\{\theta | D\} \\ &= \operatorname{argmax}_{\theta} \frac{\Pr(D|\theta) \Pr\{\theta\}}{\Pr(D)} \\ &= \operatorname{argmax}_{\theta} \Pr\{D|\theta\} \Pr\{\theta\} \quad \text{由于对于 } \theta \text{ 而言 } \Pr\{D\} \text{ 是常数}\end{aligned}$$

$$= \operatorname{argmax}_{\theta} [\Pr\{x_1, x_2, \dots, x_n | \theta\} \Pr\{\theta\}]$$

$$= \operatorname{argmax}_{\theta} (\prod_{i=1}^n \Pr\{x_i | \theta\}) \Pr\{\theta\}, \text{ 由于 } x_i \text{ 服从独立同分布}$$

$$= \operatorname{argmax}_{\theta} \log (\prod_{i=1}^n \Pr\{x_i | \theta\}) \Pr\{\theta\}, \text{ 取指数不改变最大值对应的 } \theta \text{ 值}$$

$$= \operatorname{argmax}_{\theta} [\log \Pr\{\theta\} + \sum_{i=1}^n \log \Pr\{x_i | \theta\}]$$

注意其中 $\Pr\{\theta\}$ 表示参数 θ 的先验概率率

并假定参数 θ 的先验概率率有概率密度函数 $q(\theta)$

$$\text{则有 } \theta_{\text{MAP}} = \operatorname{argmax}_{\theta} (\log q(\theta) + \sum_{i=1}^n \log p(x_i | \theta))$$

与最大似然估计 (MLE) 相比, 只多了一项参数 θ 的先验概率率

“共轭”的先验概率率分布 (“conjugate” prior)

指参数 θ 的后验概率率分布与先验概率率分布有相同形式 (same function form)

parameter distribution

Bernoulli p Beta

Binomial p Beta

Poisson λ Gamma

Exponential λ Gamma

Multinomial p_i Dirichlet

Normal μ Normal

Normal σ^2 Inverse Gamma

狄利克雷分布 (Dirichlet distribution), 对于维度 $d \geq 2$, 有参数 $\alpha_1, \alpha_2, \dots, \alpha_d > 0$

对于支撑集 (support) 为 d 维向量 $\bar{x} = (x_1, x_2, \dots, x_d)$

满足 $0 < x_i < 1, i=1, 2, \dots, d$ 且 $\sum_{i=1}^d x_i = 1$

有概率密度函数 $p(x_1, x_2, \dots, x_d | \alpha_1, \alpha_2, \dots, \alpha_d) = \frac{1}{B(\bar{\alpha})} \prod_{i=1}^d x_i^{\alpha_i - 1}$

其中多项 Beta 函数 $B(\bar{\alpha}) = \prod_{i=1}^d \Gamma(\alpha_i) / \Gamma(\sum_{i=1}^d \alpha_i)$, $\bar{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_d)$

对称狄利克雷分布 (symmetric Dirichlet distribution), 对于参数 $\alpha > 0$

$$p(x_1, \dots, x_d | \alpha) = \frac{1}{[\Gamma(\alpha)]^d / \Gamma(d\alpha)} \prod_{i=1}^d x_i^{\alpha - 1}$$

Machine

Learning - P13

泊松分布的

对于包含 n 个样本的训练集 $D = \{x_1, x_2, \dots, x_n\}$

最大后验概率

假设 $x_i \sim \text{Poisson}(\lambda)$, $i = 1, 2, \dots, n$, 且 x_1, x_2, \dots, x_n 服从独立同分布
且泊松分布的参数入先验地服从 Γ 分布, 即 $\lambda \sim \text{P}(a, b)$

则估计参数入对于训练集 D 的后验概率

$$\text{首先有 } P(\lambda | D) = \frac{P(D|\lambda) P(\lambda)}{P(D)}$$

$$\text{又有 } P(D|\lambda) = \prod_{i=1}^n e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}$$

$$P(\lambda) = \frac{b^a}{\text{P}(a)} \lambda^{a-1} e^{-b\lambda}$$

$$\text{于是 } P(\lambda | D) = \prod_{i=1}^n e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \cdot \frac{b^a}{\text{P}(a)} \lambda^{a-1} e^{-b\lambda} / P(D)$$

$$= \frac{b^a}{\text{P}(a) \cdot \prod_{i=1}^n x_i! \cdot P(D)} \cdot \lambda^{\sum_{i=1}^n x_i + a - 1} e^{-\lambda n - b\lambda}$$

由于 $\frac{b^a}{\text{P}(a) \cdot \prod_{i=1}^n x_i! \cdot P(D)}$ 为一个常数

$$\text{于是 } P(\lambda | D) \propto \lambda^{\sum_{i=1}^n x_i + a - 1} e^{-(b+n)\lambda}$$

于是可知 $P(\lambda | D) \propto P(a + \sum_{i=1}^n x_i, b + n)$

即 $P(\lambda | D)$ 可以看作服从参数为 $a + \sum_{i=1}^n x_i, b + n$ 的 Γ 分布

指数分布的

对于包含 n 个独立同分布的样本的训练集 $D = \{x_1, x_2, \dots, x_n\}$

最大后验概率

假设 $x_i \sim \text{Exponential}(\theta)$, $i = 1, 2, \dots, n$

且指数分布的参数 θ 先验地服从参数为 a, b 的 Γ 分布, 即 $\theta \sim \text{P}(a, b)$

则估计参数 θ 对于训练集 D 的后验概率

$$\text{首先有 } P(\theta | D) = \frac{P(D|\theta) P(\theta)}{P(D)}$$

$$\text{又有 } P(D|\theta) = \prod_{i=1}^n \theta e^{-\theta x_i}$$

$$P(\theta) = \frac{b^a}{\text{P}(a)} \theta^{a-1} e^{-b\theta}$$

$$\text{于是 } P(\theta | D) = \prod_{i=1}^n \theta e^{-\theta x_i} \cdot \frac{b^a}{\text{P}(a)} \theta^{a-1} e^{-b\theta} / P(D)$$

$$= \frac{b^a}{\text{P}(a) P(D)} \cdot \theta^{n+a-1} e^{-\sum_{i=1}^n x_i \theta - b\theta}$$

由于 $b^a / \text{P}(a) P(D)$ 为一个常数

$$\text{于是 } P(\theta | D) \propto \theta^{n+a-1} e^{-(\sum_{i=1}^n x_i + b)\theta}$$

于是可知 $P(\theta | D) \propto P(a+n, b + \sum_{i=1}^n x_i)$

即 $P(\theta | D)$ 可以看作服从参数为 $a+n, b + \sum_{i=1}^n x_i$ 的 Γ 分布

伽马分布 (Gamma distribution), 对于服从伽马分布的随机变量 $X \sim \text{P}(a, b)$, 或记为 $X \sim \text{G}(a, b)$

其概率密度函数 $P(x | a, b) = \frac{b^a}{\text{P}(a)} x^{a-1} e^{-bx}$, 其中 $x > 0$

其中参数 a 为形状参数 (shape parameter), b 为逆尺度参数 (inverse scale parameter)

特别地有, 当 $a=1$ 时, 伽马分布转化为参数为 b 的指数分布

即 $P(x | 1, b) = \frac{b^1}{\text{P}(1)} x^{1-1} e^{-bx} = b e^{-bx}$

Machine Learning - P14

多项式分布的 对于有 k 个不同结果的试验，出现第 i 个结果的概率为 $0 < \theta_i < 1$ ，其中 $i = 1, 2, \dots, k$

~~极大似然估计~~ 训练集 $D = \{x_1, x_2, \dots, x_N\}$ 表示 N 次独立试验的结果

则估计多项式分布的参数 $\theta_1, \theta_2, \dots, \theta_k$

首先对于训练集 $D = \{x_1, x_2, \dots, x_N\}$

其中第 i 个结果出现了 N_i 次，即有 $\sum_{i=1}^k N_i = N$

则取对数似然函数 $L(\theta_1, \theta_2, \dots, \theta_k) = \ln P(\theta_1, \dots, \theta_k | D)$

$$L(\theta_1, \dots, \theta_k) = \ln P(\theta_1, \dots, \theta_k | D)$$

$$= \ln (\prod_{i=1}^k \theta_i^{N_i})$$

$$= \sum_{i=1}^k N_i \ln \theta_i$$

又参数 $\theta_1, \dots, \theta_k$ 满足 $\sum_{i=1}^k \theta_i = 1$

则取拉格朗日函数 $L_G(\theta_1, \dots, \theta_k, \lambda) = \sum_{i=1}^k N_i \ln \theta_i + \lambda (\sum_{i=1}^k \theta_i - 1)$

分别对 $\theta_1, \dots, \theta_k, \lambda$ 求 $L_G(\theta_1, \dots, \theta_k, \lambda)$ 的偏导数并等于 0

则有 $\begin{cases} \frac{N_i}{\theta_i} + \lambda = 0 \\ \sum_{i=1}^k \theta_i - 1 = 0 \end{cases}$

于是有 $\theta_i = \frac{N_i}{\sum_{i=1}^k N_i} = N_i / N$ 即 $\frac{N_i}{N}$

于是对多项式分布参数 θ_i 的估计为 N 次试验中第 i 个结果出现次数的比例。

多项式分布的 对于有 k 个不同结果的试验，出现第 i 个结果的概率为 $0 < \theta_i < 1$ ，其中 $i = 1, 2, \dots, k$

最大后验分布 多项式分布的参数 $\theta_1, \dots, \theta_k$ 先验地服从参数为 $\alpha_1, \dots, \alpha_n$ 的狄利克雷分布

如有训练集 $D = \{x_1, \dots, x_N\}$ ，其中第 i 个结果共出现了 N_i 次，即 $\sum_{i=1}^k N_i = N$

则估计参数 $\theta_1, \dots, \theta_k$ 对于训练集 D 的后验分布

令参数向量 $\vec{\theta} = (\theta_1, \dots, \theta_k)$, $\vec{\alpha} = (\alpha_1, \dots, \alpha_n)$

于是 $\text{Dir}(\vec{\theta} | \vec{\alpha}) = \frac{1}{B(\vec{\alpha})} \prod_{i=1}^k \theta_i^{\alpha_i - 1}$

则后验概率 $P(\vec{\theta} | D) = \frac{P(D | \vec{\theta}) P(\vec{\theta})}{P(D)}$

又 $P(D | \vec{\theta}) = \prod_{i=1}^k \theta_i^{N_i}$

于是 $P(\vec{\theta} | D) = P(D | \vec{\theta}) P(\vec{\theta}) / P(D)$

$$= \prod_{i=1}^k \theta_i^{N_i} \cdot \frac{1}{B(\vec{\alpha})} \prod_{i=1}^k \theta_i^{\alpha_i - 1} / P(D)$$

$$= \frac{1}{B(\vec{\alpha}) P(D)} \prod_{i=1}^k \theta_i^{N_i + \alpha_i - 1}$$

于是 $P(\vec{\theta} | D)$ 可以看作服从参数为 $N_1 + \alpha_1, \dots, N_k + \alpha_k$ 的狄利克雷分布

$$\vec{\theta}_{MAP} = \arg \max_{\vec{\theta}} P(\vec{\theta} | D) = \arg \max_{\vec{\theta}} \frac{1}{B(\vec{\alpha}) P(D)} \prod_{i=1}^k \theta_i^{N_i + \alpha_i - 1}$$

$$= \arg \max_{\vec{\theta}} \ln \prod_{i=1}^k \theta_i^{N_i + \alpha_i - 1}$$

$$= \arg \max_{\vec{\theta}} \sum_{i=1}^k (N_i + \alpha_i - 1) \ln \theta_i, \quad \text{又 } \sum_{i=1}^k \theta_i = 1$$

$$\text{于是可知 } \theta_{i, MAP} = \frac{\alpha_i - 1}{\sum_{i=1}^k \alpha_i - k} = \frac{N_i + \alpha_i - 1}{N + \sum_{i=1}^k \alpha_i - k}$$

特别地当 $\alpha_1 = \alpha_2 = \dots = \alpha_k = 1$ 时， $\theta_{i, MAP} = \frac{N_i}{N}$

Machine Learning - P15

朴素贝叶斯 (naive Bayes classifier, NBC), 假设特征间强(朴素)独立的简单概率分类器

朴素 (naive): 指各个特征之间相互独立

贝叶斯 (Bayes): 基于贝叶斯定理

对于输入空间 $X \subseteq \mathbb{R}^D$ 为 D 维向量的集合

输出空间 $Y = \{\text{类}\}$ 为分类标记集合 $\{1, 2, \dots, C\}$

训练集 $D = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \dots, (\vec{x}_N, y_N)\}$, 其中 $\vec{x}_i \in X$, $y_i \in Y$

由 $P(X, Y)$ 独立同分布产生, 于是朴素贝叶斯模型是生成模型 (generative model)

朴素贝叶斯模型通过训练集学习联合概率分布 $P(X, Y)$

于是 $P(X, Y) = P(X = \vec{x} | Y) \cdot P(Y)$

令随机向量 $X = (X_1, X_2, \dots, X_D)$, $\vec{x} = (x_1, x_2, \dots, x_D)$

则 $P(X, Y) = P(X_1 = x_1, X_2 = x_2, \dots, X_D = x_D | Y) \cdot P(Y)$

注意如果对于 X_i , $i=1, \dots, D$, 有 S_i 种不同的取值, Y 有 C 个类别

则需要估计的向量 X 的数量有 $\prod_{i=1}^D S_i$ 种不同向量

于是需要估计的参数空间有 $C \cdot \prod_{i=1}^D S_i$ 个

即使对于 D 维向量的每个特征都是伯努利变量, 即 $S_i = 2$

参数空间的大小为 $O(C \cdot 2^D)$, 则是指数级别的, 实际中不可行

于是假设特征间相互独立, 即对于类别 Y 是条件独立的

$P(X, Y) = P(X_1 = x_1 | Y) P(X_2 = x_2 | Y) \cdots P(X_D = x_D | Y) P(Y)$

于是此时对于 X_i , 有 $S_i - 1$ 个不同参数需要估计

则参数空间大小为 $C \cdot \sum_{i=1}^D (S_i - 1) + (C - 1)$

当特征为伯努利变量时, 参数空间是 $O(CD)$ 的

对于朴素贝叶斯分类, 以及给定的输入 $\vec{x} = (x_1, x_2, \dots, x_D)$

以使得后验概率 $P(Y=c | X=\vec{x})$ 最大的 $y=c$ 作为输出

$$\text{则 } P(Y=c | X=\vec{x}) = \frac{P(X=\vec{x} | Y=c) P(Y=c)}{P(X=\vec{x})} = \frac{P(X=\vec{x} | Y=c) P(Y=c)}{\sum_{j=1}^C P(X=\vec{x} | Y=j) P(Y=j)}$$

于是 $y_{NBC} = \operatorname{argmax}_c P(Y=c | X=\vec{x})$

$$= \operatorname{argmax}_c \frac{P(X=\vec{x} | Y=c) P(Y=c)}{\sum_{j=1}^C P(X=\vec{x} | Y=j) P(Y=j)}, \text{ 又分母可视为常数}$$

$$= \operatorname{argmax}_c P(X=\vec{x} | Y=c) P(Y=c)$$

$$= \operatorname{argmax}_c P(Y=c) \prod_{i=1}^D P(X_i=x_i | Y=c)$$

$$= \operatorname{argmax}_c (\ln P(Y=c) + \sum_{i=1}^D \ln P(X_i=x_i | Y=c))$$

则在估计 $C-1$ 个参数 $P(Y=c)$ 和 $C \cdot \sum_{i=1}^D (S_i - 1)$ 个参数 $P(X_i=x_i | Y=c)$ 后

可以使用朴素贝叶斯分类对给定输入进行分类

Machine

Learning - P16

朴素贝叶斯模型的参数估计，对于输入空间 $X \subseteq \mathbb{R}^D$ 为 D 维向量的集合

输出空间 Y 为分类标记集合 $\{1, 2, \dots, C\}$

有训练集 $D = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \dots, (\vec{x}_N, y_N)\}$

首先估计先验概率 $P(Y=c)$

使用只包含输出的训练集 $D' = \{y_1, y_2, \dots, y_N\}$

令其中结果 $c = 1, 2, \dots, C$ 的次数分别为 N_1, N_2, \dots, N_C

又有 $\sum_{c=1}^C N_c = N$ ，则可以使用 MLE 估计 $P(Y=c)$ 的先验概率

$$\text{即 } P(Y=c) = \frac{\sum_{i=1}^N I(y_i=c)}{N} = \frac{N_c}{N}, \text{ 其中 } c = 1, 2, \dots, C$$

再估计条件概率 $P(X_i = x_i | Y=c)$ ，其中 $i = 1, 2, \dots, D$

$$\text{由于 } P(X_i = x_i | Y=c) = \frac{P(X_i = x_i, Y=c)}{P(Y=c)}, \quad c = 1, 2, \dots, C$$

则对于输入向量的第 i 个维度的条件概率估计

使用只包含第 i 个维度为输入的训练集

$$D_{i,c} = \{(\vec{x}_{1i}, y_1), (\vec{x}_{2i}, y_2), \dots, (\vec{x}_{Ni}, y_N)\}$$

令其中第 i 个维度有 S_i 种可能的选择

$$\text{即 } x_{i,j} \in \{1, 2, \dots, S_i\}$$

$N_{1c}, N_{2c}, \dots, N_{Sc}$

又取训练集中输入 $x_i = 1, 2, \dots, S_i$ ，输出为 $c = 1, 2, \dots, C$ 的次数分别为

又有 $\sum_{k=1}^{S_i} N_{kc} = N_c$ ，则使用多项式分布的 MLE 估计 $P(X_i = x_i | Y=c)$

$$\text{即 } P(X_i = x_i | Y=c) = \frac{\sum_{j=1}^{S_i} I(x_{ij} = x_i | Y=c)}{\sum_{j=1}^{S_i} I(Y=c)} = \frac{N_{ic}}{N_c}$$

当 S_i 平凡地等于 2 时， D 特征均为伯努利变量

而当 $S_i \geq 2$ 时， D 特征则可视为多项式变量， $i = 1, 2, \dots, D$

NAIVE-BAYES-CLASSIFIER-FITTING ($D = \{(\vec{x}_1, y_1), \dots, (\vec{x}_N, y_N)\}$)

令 $\vec{y}_j = [0, 0, \dots, 0]$ ，表示 $P(Y=c)$ 的概率

θ_{cik} 为三维数组 [\vec{y}_j] [\vec{x}_i] [y_j] 表示 $P(x_i = k | Y=c)$

for $j := 1$ to N

$y_j := y_j$] 提取输出状态 y_j 的值 整体时间复杂度

$JL_c := JL_c + 1$

为 $O(ND)$

for $i := 1$ to D

$k := x_{ji}$] 提取输入的第 i 个特征 x_{ji}

$\theta_{cik} := \theta_{cik} + 1$

时间复杂度

对于每个 $c = 1, 2, \dots, C$ ，子数组 θ_c 除以 JL_c] 计算 $P(x_i = k | Y=c)$ 的条件概率

对于 JL_c ，除以总样本数 N

计算 $P(Y=c)$

Machine

Learning - P17

拉普拉斯平滑 (Laplace smoothing), 对于训练集 $T = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \dots, (\vec{x}_n, y_n)\}$

其中输入空间 $X \subseteq \mathbb{R}^D$ 为 D 维向量的集合

其中第 i 维度有 S_i 种不同输入状态标记, 即 $X_i = \{1, 2, \dots, S_i\}$

输出空间 Y 为分类标记集合 $\{1, 2, \dots, C\}$

令 N_1, N_2, \dots, N_C 表示输出 $Y=1, 2, \dots, C$ 在训练集中出现的次数

N_{ikc} 表示对于输入中的第 i 个特征, $1 \leq i \leq N$

对于这个特征的第 k 个可能状态标记, $1 \leq k \leq S_i$

且对应输出为 $Y=c$ 的次数 n_{ikc} , $1 \leq c \leq C$

于是可知参数估计

参数 π_c 表示先验概率 $P(Y=c)$, 向量元 $= [\pi_1, \dots, \pi_C]$

于是有 $\hat{\pi}_c = N_c / N$

参数 θ_{cik} 表示给定输出为 $Y=c$, 第 i 个特征标记为 k 的条件概率 $P(X_i=k | Y=c)$

于是有 $\hat{\theta}_{cik} = N_{ikc} / N_c$

注意当训练集 T 不够大而无法覆盖所有 $\prod_{i=1}^D S_i$ 个输入特征标记时

会出现对于某些特征的某些状态, 在训练集中并未出现

这是在训练集中对应的参数 θ_{cik} 对于所有 $c=1, 2, \dots, C$ 均为 0

如果此时新的样本中这个特征的这个状态出现

则在计算后验概率 $P(Y=c | \vec{x})$ 时

由于 $P(Y=c | \vec{x}) = P(Y=c) \cdot \prod_{i=1}^D P(X_i=k | Y=c) / P(\vec{x})$, 依据条件独立

则对于所有 $c=1, 2, \dots, C$, 都有 $P(Y=c | \vec{x}) = 0$

这是导致无法计算 $\operatorname{argmax}_c P(Y=c | \vec{x})$

于是需对 $\hat{\theta}_{cik}$ 的参数估计进行调整

取一个实数参数 α , 通常取 1

则对 $\hat{\theta}_{cik}$ 的估计中, 对每一个可能状态 k 的计数都加上 α .

于是总计数需要加上 $S_i \cdot \alpha$

$$\text{即 } \hat{\theta}_{cik} = \frac{\alpha + \sum_{j=1}^N I(X_{ji}=k, Y=j)}{S_i \cdot \alpha + \sum_{j=1}^N I(Y=j)} = \frac{\alpha + N_{ikc}}{S_i \cdot \alpha + N_c}$$

当 $\alpha=1$ 时, 即称为拉普拉斯平滑 (Laplace smoothing)

而 $\alpha \in [0, 1]$ 时, 称为 Lidstone 平滑

也可以认为给定输出为 $Y=c$ 的第 i 个特征的 k 个状态分别出现的频率

先验地服从参数 $\alpha=2$ 的对称狄利克雷分布 $\text{Dir}(\theta_{c11}, \theta_{c12}, \dots, \theta_{c1S_i} | \alpha=2)$

$$\text{于是有最大后验分布估计为 } \theta_{cik} \text{ MAP} = \frac{N_{ikc} + \alpha - 1}{N_c + \sum_{k=1}^{S_i} \alpha - S_i} = \frac{N_{ikc} + 1}{N_c + S_i}$$

Machine

Learning - P18

拉普拉斯平滑 对于训练集 $T = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \dots, (\vec{x}_N, y_N)\}$

其中输出空间 Y 为分类标记集合 $\{1, 2, \dots, C\}$

参数 $\hat{\pi}_c$ 表示根据训练集估计的 Y 的概率分布 $P(Y=c)$

当训练集 T 的样本不足时，可能出现 $\{1, 2, \dots, C\}$ 中的类别并未全部出现的情形

假设对于 $Y=c$ ，在训练集中出现次数为 0

$$\text{则对其的参数估计 } \hat{\pi}_c = \frac{N_c}{N} = 0$$

尽管从现实的角度考虑，对于训练集中预期存在但未出现的分类并不妥

即 对于任意样本 \vec{x} ，在其分类时都有

$$P(Y=c|\vec{x}) = P(\vec{x}|Y=c) P(Y=c) / P(\vec{x}) = 0$$

从而使得不论对任何 \vec{x} ， $\operatorname{argmax}_c P(Y=c|\vec{x}) \neq c$

但这实际上忽略了 一种可能出现的情形

即 \vec{x} 距离 $\{1, \dots, C\} \setminus \{c\}$ 是够远，从而使得距离 c 是够近，而分类为 c

或者说在排除了已知的（通过训练集学习的）可能后，选择了“看起来最不可能”的选项

于是参考对特征条件概率的拉普拉斯平滑

$$\text{即 取参数估计 } \hat{\pi}_c = \frac{N_c + 1}{N + C}$$

也可以说认为 参数向量 $\pi = (\pi_1, \pi_2, \dots, \pi_C)$ 服从参数为 $\alpha=2$ 的对称狄利克雷分布

即有 $\text{Dir}(\pi | \alpha=2)$ 。于是 通过训练集的 最大后验概率估计

$$\pi_{c \text{MAP}} = \frac{N_c + \alpha - 1}{N + C \cdot \alpha - C} = \frac{N_c + 1}{N + C}$$

多元正态分布 (multivariate normal distribution, MVN)，对于 D 维随机变量列向量 $\vec{x} = (x_1, x_2, \dots, x_D)^T$

对于 D 维期望列向量 $\mu = (\mu_1, \mu_2, \dots, \mu_D)^T$

$D \times D$ 的协方差矩阵 (covariance matrix) Σ ，行列式为 $|\Sigma|$

则称 \vec{x} 服从期望为 μ ，方差为 Σ 的多元正态分布，记为 $\vec{x} \sim N_p(\mu, \Sigma)$

$$\text{有概率密度函数 } N(\vec{x} | \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^D |\Sigma|^{\frac{1}{2}}}} \cdot e^{-\frac{1}{2}(\vec{x} - \mu)^T \Sigma^{-1} (\vec{x} - \mu)}$$

注意其中的每个维度 x_i 都服从正态分布， $i=1, 2, \dots, D$

$$\text{又协方差矩阵 } \Sigma = E[(\vec{x} - \mu)(\vec{x} - \mu)^T]$$

$$\text{于是 } \Sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)] = \text{cov}(x_i, x_j), \text{ 其中 } 1 \leq i, j \leq D$$

$$\text{则可知 } x_i \sim N(\mu_i, \Sigma_{ii}), \text{ 其中 } i=1, 2, \dots, D$$

对 协方差矩阵 Σ 进行特征值分解 (eigendecomposition)

$$\text{即 } \Sigma = U \Delta U^T, \text{ 其中 } U \text{ 是正交矩阵 (orthonormal matrix of eigenvector)}$$

Δ 是对角矩阵 (diagonal matrix of eigenvalue)

$$\text{则 } \Sigma^{-1} = (U \Delta U^T)^{-1} = U^T \Delta^{-1} U^{-1} = U \Delta^{-1} U^T = \sum_{i=1}^D \frac{1}{\lambda_i} \vec{u}_i \vec{u}_i^T$$

其中 \vec{u}_i 是 U 中的第 i 列，即第 i 个特征向量 (eigenvector)

Machine

Learning - P19

多元正态分布 对于D维随机变量列向量 $\vec{X} = (X_1, X_2, \dots, X_D)$ 服从多元正态分布 $N_D(\vec{\mu}, \Sigma)$

其中 $\vec{\mu}$ 为 D 维期望列向量 $(\mu_1, \mu_2, \dots, \mu_D)$

Σ 为 $D \times D$ 的协方差矩阵，其中 $\Sigma_{ij} = E[(X_i - \mu_i)(X_j - \mu_j)^T]$, $1 \leq i, j \leq D$

有概率密度函数 $N(\vec{X} | \vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp[-\frac{1}{2} (\vec{X} - \vec{\mu})^T \Sigma^{-1} (\vec{X} - \vec{\mu})]$

Σ 的特征值分解为 $U \Lambda U^{-1}$, 其中 U 是正交矩阵, Λ 是对角矩阵

$\Sigma^{-1} = \sum_{i=1}^D \frac{1}{\lambda_i} \vec{u}_i \cdot \vec{u}_i^T$, 其中向量 \vec{u}_i 是矩阵 U 的第 i 列, 即第 i 个正交向量

注意到 $(\vec{X} - \vec{\mu})^T \Sigma^{-1} (\vec{X} - \vec{\mu})$ 为 D 维向量 \vec{X} 与 D 维期望向量 $\vec{\mu}$ 的 Mahalanobis 矩离

Mahalanobis 矩离是一种有效计算两个未知样本集的相似度的方法

表示数据相对于协方差矩阵 Σ 的 D 维向量 \vec{x} 与分布 D

与欧式距离相比, 考虑了各个特性之间的关系

并且是独立于测量尺度的 (scale-invariant)

对于服从同一分布且协方差矩阵 Σ 的随机变量 \vec{x}, \vec{y} ,

其 Mahalanobis 矩离为 $(\vec{x} - \vec{y})^T \Sigma^{-1} (\vec{x} - \vec{y})$

则 $(\vec{x} - \vec{y})^T \Sigma^{-1} (\vec{x} - \vec{y}) = (\vec{x} - \vec{\mu})^T (\sum_{i=1}^D \frac{1}{\lambda_i} \vec{u}_i \cdot \vec{u}_i^T) (\vec{x} - \vec{\mu})$

$= \sum_{i=1}^D \frac{1}{\lambda_i} (\vec{x} - \vec{\mu})^T \vec{u}_i \cdot \vec{u}_i^T (\vec{x} - \vec{\mu})$

令 $y_i = (\vec{x} - \vec{\mu})^T \vec{u}_i$, 由于 $(\vec{x} - \vec{\mu})^T \vec{u}_i$ 是标量, 则 $y_i = \vec{u}_i^T (\vec{x} - \vec{\mu})$

于是 $(\vec{x} - \vec{y})^T \Sigma^{-1} (\vec{x} - \vec{y}) = \sum_{i=1}^D \frac{y_i^2}{\lambda_i}$

参数估计：对于训练集 $D = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_N\}$

取对数似然函数 $L(\vec{\mu}, \Sigma) = \log P(D | \vec{\mu}, \Sigma) = \frac{N}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^N (\vec{x}_i - \vec{\mu})^T \Sigma^{-1} (\vec{x}_i - \vec{\mu})$

令向量 $\vec{y}_i = \vec{x}_i - \vec{\mu}$

$$\frac{\partial}{\partial \vec{\mu}} (\vec{x}_i - \vec{\mu})^T \Sigma^{-1} (\vec{x}_i - \vec{\mu}) = \frac{\partial}{\partial \vec{y}_i} \vec{y}_i^T \Sigma^{-1} \vec{y}_i \cdot \frac{\partial \vec{y}_i}{\partial \vec{\mu}} = -I(\Sigma^{-1} + \Sigma^{-T}) \vec{y}_i$$

$$\text{于是 } \frac{\partial}{\partial \vec{\mu}} L(\vec{\mu}, \Sigma) = -\frac{1}{2} \sum_{i=1}^N -2 \Sigma^{-1} (\vec{x}_i - \vec{\mu}) = \Sigma^{-1} \cdot \sum_{i=1}^N (\vec{x}_i - \vec{\mu}) = 0$$

$$\text{则 } \hat{\vec{\mu}}_{MLE} = \frac{1}{N} \sum_{i=1}^N \vec{x}_i = \bar{\vec{x}}$$

矩阵的迹 (trace) 为, 对于 $n \times n$ 矩阵 A , $tr(A) = \sum_{i=1}^n A_{ii}$, 即对角线元素之和

trace trick: scalar inner product $\vec{x}^T A \vec{x} = tr(\vec{x}^T A \vec{x}) = tr(\vec{x} \vec{x}^T A) = tr(A \vec{x} \vec{x}^T)$

于是 $L(\vec{\mu}, \Sigma) = \frac{N}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^N tr[(\vec{x}_i - \vec{\mu})(\vec{x}_i - \vec{\mu})^T \Sigma]$, 其中 $\Delta = \Sigma^{-1}$

$$= \frac{N}{2} \log |\Delta| - \frac{1}{2} tr[S_\mu \Delta], \text{ 其中 } S_\mu = \sum_{i=1}^N (\vec{x}_i - \vec{\mu})(\vec{x}_i - \vec{\mu})^T$$

$$\frac{\partial L(\Delta)}{\partial \Delta} = \frac{N}{2} \Delta^{-T} - \frac{1}{2} S_\mu^T = 0, \text{ scatter matrix centered on } \vec{\mu}$$

$$\Delta^{-T} = \frac{1}{N} S_\mu, \text{ 于是 } \Sigma = \Delta^{-1} = \Delta^{-T} = \frac{1}{N} S_\mu$$

$$\text{则 } \hat{\Sigma}_{MLE} = \frac{1}{N} \sum_{i=1}^N (\vec{x}_i - \vec{\mu})(\vec{x}_i - \vec{\mu})^T$$

Machine

Learning - P20

高斯判别分析 (Gaussian discriminant analysis, GDA), 对于训练集 $T = \{(\vec{x}_1, y_1), \dots, (\vec{x}_N, y_N)\}$

其中输出空间 Y 为分类标记集合 $\{1, 2, \dots, C\}$

对于 D 维随机变量列向量 $\vec{x} = (x_1, x_2, \dots, x_D)$ 服从多元正态分布 $N_D(\vec{\mu}_c, \Sigma_c)$

其中 $\vec{\mu}_c$ 是给定分类为 $Y=c$ 的 D 维期望列向量

Σ_c 是给定分类为 $Y=c$ 的 $D \times D$ 协方差矩阵, 其中 $c=1, 2, \dots, C$

则有条件概率分布 $P(\vec{x}|Y=c, \theta) = N_D(\vec{x}|\vec{\mu}_c, \Sigma_c) = \frac{1}{(2\pi)^{D/2} |\Sigma_c|^{1/2}} \exp[-\frac{1}{2}(\vec{x}-\vec{\mu}_c)^T \Sigma_c^{-1} (\vec{x}-\vec{\mu}_c)]$

则有生成分类器 (generative classifier)

$$\hat{y}(\vec{x}) = \arg \max_c P(Y=c|\vec{x}, \theta)$$

$= \arg \max_c P(\vec{x}|Y=c, \theta) P(Y=c) / P(\vec{x})$, 由于其中 $P(\vec{x})$ 为常数

$$= \arg \max_c N_D(\vec{x}|\vec{\mu}_c, \Sigma_c) P(Y=c)$$

$$= \arg \max_c \ln [N_D(\vec{x}|\vec{\mu}_c, \Sigma_c) P(Y=c)]$$

$$= \arg \max_c \left\{ \ln P(Y=c) + \ln \frac{1}{(2\pi)^{D/2} |\Sigma_c|^{1/2}} + [-\frac{1}{2}(\vec{x}-\vec{\mu}_c)^T \Sigma_c^{-1} (\vec{x}-\vec{\mu}_c)] \right\}$$

$$= \arg \max_c [\ln P(Y=c) - \frac{1}{2}(\vec{x}-\vec{\mu}_c)^T \Sigma_c^{-1} (\vec{x}-\vec{\mu}_c)] - \frac{1}{2} \ln |\Sigma_c|$$

如果 $P(Y=c)$ 先验地服从均匀分布

$$\text{则有 } \hat{y}(\vec{x}) = \arg \max_c [-\frac{1}{2}(\vec{x}-\vec{\mu}_c)^T \Sigma_c^{-1} (\vec{x}-\vec{\mu}_c)]$$

$$= \arg \min_c [(\vec{x}-\vec{\mu}_c)^T \Sigma_c^{-1} (\vec{x}-\vec{\mu}_c)]$$

如果将 $\vec{\mu}_c$ 视为类 $Y=c$ 的中心, 则为计量 \vec{x} 与 $\vec{\mu}_c$ 的 Mahalanobis 距离

即可视为一个最近中心分类器 (nearest centroids classifier)

特别地当协方差矩阵 Σ_c 为对角矩阵时, 即 $\Sigma_{cii} = \sigma_{ci}^2$, $\Sigma_{cij} = 0$, $1 \leq i, j \leq D$ 且 $i \neq j$

于是此时 Σ_c^{-1} 也是对角矩阵

$$\Sigma_c^{-1} = \begin{vmatrix} 1/\sigma_{c1}^2 & & \\ & \ddots & \\ & & 1/\sigma_{cD}^2 \end{vmatrix}$$

$$\text{且 } |\Sigma_c| = \sigma_{c1}^2 \times \sigma_{c2}^2 \times \dots \times \sigma_{cD}^2 = \prod_{i=1}^D \sigma_{ci}^2$$

$$\text{则有 } (\vec{x}-\vec{\mu}_c)^T \Sigma_c^{-1} (\vec{x}-\vec{\mu}_c) = \left[\frac{x_1-\mu_{c1}}{\sigma_{c1}^2}, \frac{x_2-\mu_{c2}}{\sigma_{c2}^2}, \dots, \frac{x_D-\mu_{cD}}{\sigma_{cD}^2} \right]^T \cdot (\vec{x}-\vec{\mu}_c)$$
$$= \sum_{i=1}^D \frac{(x_i-\mu_{ci})^2}{\sigma_{ci}^2}$$

$$\text{于是有 } P(Y=c|\vec{x}) = P(\vec{x}|Y=c) P(Y=c) / P(\vec{x})$$

$$= \arg \max_c P(Y=c) \cdot \arg \max_c P(\vec{x}|Y=c) P(Y=c)$$

$$= \arg \max_c P(Y=c) \cdot \frac{1}{(2\pi)^{D/2} |\Sigma_c|^{1/2}} \exp[-\frac{1}{2}(\vec{x}-\vec{\mu}_c)^T \Sigma_c^{-1} (\vec{x}-\vec{\mu}_c)]$$

$$= \arg \max_c P(Y=c) \cdot \frac{1}{(2\pi)^{D/2} \prod_{i=1}^D \sigma_{ci}^{1/2}} \exp[-\frac{1}{2} \sum_{i=1}^D \frac{(x_i-\mu_{ci})^2}{\sigma_{ci}^2}]$$

$$= \arg \max_c P(Y=c) \cdot \prod_{i=1}^D \frac{1}{(2\pi)^{1/2} \sigma_{ci}} \exp[-\frac{(x_i-\mu_{ci})^2}{2\sigma_{ci}^2}]$$

$$= \arg \max_c P(Y=c) \cdot \prod_{i=1}^D P(x_i|Y=c)$$

即此时 GDA 等价于朴素贝叶斯分析,

实际上协方差矩阵 Σ_c 是对角矩阵表示 x_1, x_2, \dots, x_D 是对于 $Y=c$ 条件为独立的

Machine

Learning - P21

对于训练集 $T = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \dots, (\vec{x}_N, y_N)\}$

其中输出空间 Y 为分类标记集合 $\{1, 2, \dots, C\}$

对于 D 维随机变量列向量 $\vec{x} = (x_1, x_2, \dots, x_D)$ 服从多元正态分布 $N_D(\vec{\mu}_c, \Sigma_c)$

其中 $\vec{\mu}_c$ 为给定 $Y=c$ 的 D 维期望列向量 $(\mu_{c1}, \mu_{c2}, \dots, \mu_{cD})$

Σ_c 为给定 $Y=c$ 的 $D \times D$ 协方差矩阵 $(c=1, 2, \dots, C)$

$P(Y=c)$ 先验地服从分布元，其中 $\pi_c = P(Y=c)$

则有条件概率 $P(Y=c | \vec{x}, \theta) = \pi_c \cdot \frac{1}{(2\pi)^{D/2} |\Sigma_c|^{1/2}} \exp[-\frac{1}{2}(\vec{x} - \vec{\mu}_c)^T \Sigma_c^{-1} (\vec{x} - \vec{\mu}_c)] / P(\vec{x})$

其中 $P(\vec{x}) = \sum_{c=1}^C \pi_c \cdot \frac{1}{(2\pi)^{D/2} |\Sigma_c|^{1/2}} \exp[-\frac{1}{2}(\vec{x} - \vec{\mu}_c)^T \Sigma_c^{-1} (\vec{x} - \vec{\mu}_c)]$

二次判别分析 (quadratic discriminant analysis, QDA)

首先进行参数估计，令 $[N_1, N_2, \dots, N_c]$ 分别表示结果 $Y=1, 2, \dots, C$ 在训练集中出现的次数

则有 $\vec{\mu}_c = \sum_{i=1}^N \vec{x}_i ; y_i=c \frac{\vec{x}_i}{N_c}$

$\Sigma_c = \sum_{i=1; y_i=c}^N (\vec{x}_i - \vec{\mu}_c)(\vec{x}_i - \vec{\mu}_c)^T / N_c$

$P(Y=c) = N_c / N$

又由于 $\operatorname{argmax}_c P(Y=c | \vec{x}, \theta) = \operatorname{argmax}_c \pi_c \cdot \frac{1}{(2\pi)^{D/2} |\Sigma_c|^{1/2}} \exp[-\frac{1}{2}(\vec{x} - \vec{\mu}_c)^T \Sigma_c^{-1} (\vec{x} - \vec{\mu}_c)] / P(\vec{x})$
 $= \operatorname{argmax}_c \ln \{ \pi_c \cdot \frac{1}{(2\pi)^{D/2} |\Sigma_c|^{1/2}} \exp[-\frac{1}{2}(\vec{x} - \vec{\mu}_c)^T \Sigma_c^{-1} (\vec{x} - \vec{\mu}_c)] \}$
 $= \operatorname{argmax}_c [\ln \pi_c + \ln |\Sigma_c|^{1/2} - \frac{1}{2}(\vec{x} - \vec{\mu}_c)^T \Sigma_c^{-1} (\vec{x} - \vec{\mu}_c)]$

又 $|\Sigma_c|^{-1} = |\Sigma_c|'$ ，即矩阵逆的行列式等于其行列式的倒数

则 $\operatorname{argmax}_c P(Y=c | \vec{x}, \theta) = \operatorname{argmax}_c [\ln \pi_c + \frac{1}{2} \ln |\Sigma_c|' - \frac{1}{2}(\vec{x} - \vec{\mu}_c)^T \Sigma_c^{-1} (\vec{x} - \vec{\mu}_c)]$

且二次判别函数 (quadratic discriminant function)

$$g_c(\vec{x}) = \ln \pi_c + \frac{1}{2} \ln |\Sigma_c|' - \frac{1}{2}(\vec{x} - \vec{\mu}_c)^T \Sigma_c^{-1} (\vec{x} - \vec{\mu}_c)$$

于是对 \vec{x} 的分类可以写作 $\operatorname{argmax}_c g_c(\vec{x})$

可以对二次判别函数进行进一步地转换

$$g_c(\vec{x}) = \ln \pi_c + \frac{1}{2} \ln |\Sigma_c|' - \frac{1}{2} [\vec{x}^T \Sigma_c^{-1} \vec{x} - \vec{x}^T \Sigma_c^{-1} \vec{\mu}_c - \vec{\mu}_c^T \Sigma_c^{-1} \vec{x} + \vec{\mu}_c^T \Sigma_c^{-1} \vec{\mu}_c]$$

又 $\vec{x}^T \Sigma_c^{-1} \vec{\mu}_c$ 与 $\vec{\mu}_c^T \Sigma_c^{-1} \vec{x}$ 都是标量值

$$\text{且 } (\vec{x}^T \Sigma_c^{-1} \vec{\mu}_c)^T = \vec{\mu}_c^T \Sigma_c^{-1} \vec{x} = \vec{\mu}_c^T \Sigma_c^{-1} \vec{x}$$

$$\text{于是 } g_c(\vec{x}) = \ln \pi_c + \frac{1}{2} \ln |\Sigma_c|' - \frac{1}{2} \vec{x}^T \Sigma_c^{-1} \vec{x} + \vec{x}^T \Sigma_c^{-1} \vec{\mu}_c - \frac{1}{2} \vec{\mu}_c^T \Sigma_c^{-1} \vec{\mu}_c$$

$$= -\frac{1}{2} \vec{x}^T \Sigma_c^{-1} \vec{x} + \vec{x}^T \Sigma_c^{-1} \vec{\mu}_c - \frac{1}{2} \vec{\mu}_c^T \Sigma_c^{-1} \vec{\mu}_c + \frac{1}{2} \ln |\Sigma_c|' + \ln \pi_c$$

$$\text{取 } \vec{\beta}_c = \Sigma_c^{-1} \vec{\mu}_c, \gamma_c = -\frac{1}{2} \vec{\mu}_c^T \Sigma_c^{-1} \vec{\mu}_c + \frac{1}{2} \ln |\Sigma_c|' + \ln \pi_c$$

$$\text{则 } g_c(\vec{x}) = -\frac{1}{2} \vec{x}^T \Sigma_c^{-1} \vec{x} + \vec{\beta}_c^T \vec{x} + \gamma_c$$

于是对于训练集 T 应用 QDA 时，对于输出 $y \in \{1, 2, \dots, C\}$

先进行参数估计，计算 $\vec{\mu}_c, \Sigma_c^{-1}, \pi_c$

对于新的样本 \vec{x} ，分别计算 $g_c(\vec{x}) = \ln \pi_c + \frac{1}{2} \ln |\Sigma_c|' - \frac{1}{2}(\vec{x} - \vec{\mu}_c)^T \Sigma_c^{-1} (\vec{x} - \vec{\mu}_c)$

于是可以求出 $\operatorname{argmax}_c g_c(\vec{x})$

Machine

Learning - P22

1/9 - 2023

线性判别分析 (linear discriminant analysis, LDA)

对于训练集 $T = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \dots, (\vec{x}_N, y_N)\}$

其中输出空间 Y 为分类标记集合 $\{1, 2, \dots, C\}$

对于 D 维随机变量列向量 $\vec{x} = (x_1, x_2, \dots, x_D)$ 服从多元正态分布 $N_D(\vec{\mu}_c, \Sigma_c)$

其中 $\vec{\mu}_c$ 为给定 $Y=c$ 的 D 维期望列向量 $(\mu_{c1}, \mu_{c2}, \dots, \mu_{cD})$

Σ_c 为给定 $Y=c$ 的 $D \times D$ 协方差矩阵

又 $P(Y=c)$ 服从分布 $\pi_c = (\pi_1, \pi_2, \dots, \pi_C)$

于是有二次判别函数 $g_c(\vec{x}) = \ln \pi_c + \frac{1}{2} \ln |\Sigma_c| - \frac{1}{2} (\vec{x} - \vec{\mu}_c)^T \Sigma_c^{-1} (\vec{x} - \vec{\mu}_c)$

即 $g_c(\vec{x}) = -\frac{1}{2} \vec{x}^T \Sigma_c^{-1} \vec{x} + \vec{\mu}_c^T \Sigma_c^{-1} \vec{x} - \frac{1}{2} \vec{\mu}_c^T \Sigma_c^{-1} \vec{\mu}_c + \frac{1}{2} \ln |\Sigma_c| + \ln \pi_c$

如果假定多元正态分布 $N_D(\vec{\mu}_c, \Sigma_c)$ 中的协方差矩阵是相等的 (tied/shared)

即对于 $c = 1, 2, \dots, C$, 有 $\Sigma_c = \Sigma$

则此时的判别函数 $g_c(\vec{x}) = \ln \pi_c + \frac{1}{2} \ln |\Sigma| - \frac{1}{2} (\vec{x} - \vec{\mu}_c)^T \Sigma^{-1} (\vec{x} - \vec{\mu}_c)$

于是此时其中 $-\frac{1}{2} \vec{x}^T \Sigma^{-1} \vec{x}$ 和 $\frac{1}{2} \ln |\Sigma|$ 与分类标记 $Y=c$ 无关

即 $\operatorname{argmax}_c g_c(\vec{x}) = \operatorname{argmax}_c [-\frac{1}{2} \vec{x}^T \Sigma^{-1} \vec{x} + \vec{\mu}_c^T \Sigma^{-1} \vec{x} - \frac{1}{2} \vec{\mu}_c^T \Sigma^{-1} \vec{\mu}_c + \frac{1}{2} \ln |\Sigma| + \ln \pi_c]$

= $\operatorname{argmax}_c [\vec{\mu}_c^T \Sigma^{-1} \vec{x} - \frac{1}{2} \vec{\mu}_c^T \Sigma^{-1} \vec{\mu}_c + \ln \pi_c]$

由于此时仅剩 \vec{x} 的一次项 $\vec{\mu}_c^T \Sigma^{-1} \vec{x}$

于是有线性判别函数 $g_c(\vec{x}) = \vec{\mu}_c^T \Sigma^{-1} \vec{x} - \frac{1}{2} \vec{\mu}_c^T \Sigma^{-1} \vec{\mu}_c + \ln \pi_c$

取 $\vec{\beta}_c = \Sigma^{-1} \vec{\mu}_c$, $\gamma_c = -\frac{1}{2} \vec{\mu}_c^T \Sigma^{-1} \vec{\mu}_c + \ln \pi_c$

则有 $g_c(\vec{x}) = \vec{\beta}_c^T \vec{x} + \gamma_c$

又条件概率 $P(Y=c|\vec{x}, \theta) = \pi_c \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp[-\frac{1}{2} (\vec{x} - \vec{\mu}_c)^T \Sigma^{-1} (\vec{x} - \vec{\mu}_c)] / P(\vec{x})$

又 $P(\vec{x}) = \sum_{c=1}^C \pi_c \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp[-\frac{1}{2} (\vec{x} - \vec{\mu}_c)^T \Sigma^{-1} (\vec{x} - \vec{\mu}_c)]$

又 $P(Y=c|\vec{x}, \theta) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \cdot \exp[-\frac{1}{2} \vec{x}^T \Sigma^{-1} \vec{x}] \cdot \exp[\vec{\mu}_c^T \Sigma^{-1} \vec{x} - \frac{1}{2} \vec{\mu}_c^T \Sigma^{-1} \vec{\mu}_c + \ln \pi_c]$

而 $[\frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \cdot \exp(-\frac{1}{2} \vec{x}^T \Sigma^{-1} \vec{x})]$ 是与分类标记 $Y=c$ 无关的常数项

于是 $P(Y=c|\vec{x}, \theta) = \exp[\vec{\mu}_c^T \Sigma^{-1} \vec{x} - \frac{1}{2} \vec{\mu}_c^T \Sigma^{-1} \vec{\mu}_c + \ln \pi_c] / \sum_{c=1}^C \exp[\vec{\mu}_c^T \Sigma^{-1} \vec{x} - \frac{1}{2} \vec{\mu}_c^T \Sigma^{-1} \vec{\mu}_c + \ln \pi_c]$

= $\exp[\vec{\beta}_c^T \vec{x} + \gamma_c] / \sum_{c=1}^C \exp[\vec{\beta}_c^T \vec{x} + \gamma_c]$

softmax 函数, 或称归一化指数函数, 是逻辑函数 (logistic function) 的推广

用于将任意实数的 D 维向量 \vec{x} 归一化 (normalize) 为 D 维概率分布

即 $S(\vec{x}) = [p_1, p_2, \dots, p_D]$, 且有 $0 < p_i < 1$, $\sum_{i=1}^D p_i = 1$, 其中 $i=1, 2, \dots, D$

于是令 $\vec{\eta} = [\vec{\beta}_1^T \vec{x} + \gamma_1, \vec{\beta}_2^T \vec{x} + \gamma_2, \dots, \vec{\beta}_D^T \vec{x} + \gamma_D]$

则 $P(Y=c|\vec{x}, \theta) = \exp[\vec{\beta}_c^T \vec{x} + \gamma_c] / \sum_{c=1}^C \exp[\vec{\beta}_c^T \vec{x} + \gamma_c]$

= $\exp[\vec{\eta}_c] / \sum_{c=1}^C \exp[\vec{\eta}_c] = S(\vec{\eta})_c$

Machine

Learning - P23

QDA - evaluation

线性判别分析，对于训练集 $T = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \dots, (\vec{x}_n, y_n)\}$

其中输出空间 Y 为分类标记集合 $\{1, 2, \dots, C\}$ ，服从概率分布 $\pi = (\pi_1, \pi_2, \dots, \pi_C)$

对于 D 维随机变量列向量 $\vec{x} = (x_1, x_2, \dots, x_D)$ 服从多元正态分布 $N_D(\vec{\mu}_c, \Sigma)$

其中 $\vec{\mu}_c$ 为给定 $Y=c$ 的 D 维期望列向量， Σ 为给定的 $D \times D$ 协方差矩阵

有线性判别函数 $g_c(\vec{x}) = \vec{\beta}_c^T \vec{x} + \gamma_c$

其中 $\vec{\beta}_c = \Sigma^{-1} \vec{\mu}_c$, $\gamma_c = -\frac{1}{2} \vec{\mu}_c^T \Sigma^{-1} \vec{\mu}_c + \ln \pi_c$

于是有 $P(Y=c|\vec{x}, \theta) = \exp[\vec{\beta}_c^T \vec{x} + \gamma_c] / \sum_{c=1}^C \exp[\vec{\beta}_c^T \vec{x} + \gamma_c]$

特别地当分类为二分类 (two-class) 时，即 $Y \in \{0, 1\}$

于是 $P(Y=1|\vec{x}, \theta) = \exp[\vec{\beta}_1^T \vec{x} + \gamma_1] / (e^{\vec{\beta}_1^T \vec{x} + \gamma_1} + e^{\vec{\beta}_0^T \vec{x} + \gamma_0})$
 $= 1 / (1 + e^{\vec{\beta}_1^T \vec{x} + \gamma_1 - \vec{\beta}_0^T \vec{x} - \gamma_0})$

单独看指数 $\vec{\beta}_1^T \vec{x} + \gamma_1 - \vec{\beta}_0^T \vec{x} - \gamma_0$ 并取相反数

$$\text{则 } \vec{\beta}_1^T \vec{x} + \gamma_1 - \vec{\beta}_0^T \vec{x} - \gamma_0 =$$

$$= (\vec{\beta}_1 - \vec{\beta}_0)^T \vec{x} + (\gamma_1 - \gamma_0)$$

$$= (\Sigma^{-1} \vec{\mu}_1 - \Sigma^{-1} \vec{\mu}_0)^T \vec{x} + [(-\frac{1}{2} \vec{\mu}_1^T \Sigma^{-1} \vec{\mu}_1 + \ln \pi_1) - (-\frac{1}{2} \vec{\mu}_0^T \Sigma^{-1} \vec{\mu}_0 + \ln \pi_0)]$$

$$= [\Sigma^{-1}(\vec{\mu}_1 - \vec{\mu}_0)]^T \vec{x} + [-\frac{1}{2}(\vec{\mu}_1^T \Sigma^{-1} \vec{\mu}_1 - \vec{\mu}_0^T \Sigma^{-1} \vec{\mu}_0) + \ln \pi_1 / \pi_0]$$

$$= [\Sigma^{-1}(\vec{\mu}_1 - \vec{\mu}_0)]^T \vec{x} + [-\frac{1}{2}(\vec{\mu}_1^T \Sigma^{-1} \vec{\mu}_1 + \vec{\mu}_1^T \Sigma^{-1} \vec{\mu}_0 - \vec{\mu}_1^T \Sigma^{-1} \vec{\mu}_0 - \vec{\mu}_0^T \Sigma^{-1} \vec{\mu}_0) + \ln \pi_1 / \pi_0]$$

$$= [\Sigma^{-1}(\vec{\mu}_1 - \vec{\mu}_0)]^T \vec{x} + [-\frac{1}{2}(\vec{\mu}_1 - \vec{\mu}_0)^T \Sigma^{-1}(\vec{\mu}_1 + \vec{\mu}_0) + \ln \pi_1 / \pi_0]$$

$$\text{取 } \vec{w} = \Sigma^{-1}(\vec{\mu}_1 - \vec{\mu}_0), \vec{x}_0 = \frac{1}{2}(\vec{\mu}_1 + \vec{\mu}_0) - (\vec{\mu}_1 - \vec{\mu}_0) \ln \pi_1 / \pi_0 / (\vec{\mu}_1 - \vec{\mu}_0)^T \Sigma^{-1}(\vec{\mu}_1 - \vec{\mu}_0)$$

$$\text{则 } \vec{\beta}_1^T \vec{x} + \gamma_1 - \vec{\beta}_0^T \vec{x} - \gamma_0 = \vec{w}^T \vec{x} - \vec{w}^T \vec{x}_0 = \vec{w}^T(\vec{x} - \vec{x}_0)$$

$$\text{于是 } P(Y=1|\vec{x}, \theta) = \frac{1}{1 + e^{-\vec{w}^T(\vec{x} - \vec{x}_0)}}$$

于是可以直观地描述为， \vec{x}_0 为分类 $Y=1$ 与 $Y=0$ 中的一点，且有 $P(Y=1|\vec{x}_0, \theta) = \frac{1}{2}$

而 $\vec{w}^T(\vec{x} - \vec{x}_0)$ 可以解释为向量 $\vec{x} - \vec{x}_0$ 在 \vec{w}^T 上的相对 \vec{x}_0 的偏移量

则 $\vec{w}^T(\vec{x} - \vec{x}_0) \rightarrow +\infty$ 时， $P(Y=1|\vec{x}, \theta) \rightarrow 1$ ，而 $\vec{w}^T(\vec{x} - \vec{x}_0) \rightarrow -\infty$ 时， $P(Y=0|\vec{x}, \theta) \rightarrow 0$

于是也可以记作 $P(Y=1|\vec{x}, \theta) = \text{sigmoid}[\vec{w}^T(\vec{x} - \vec{x}_0)]$

与 QDA 相比，LDA 产生线性决定边界 (linear decision boundary)

