Examen. A.D.

Classification Bayésienne 贝叶斯分类

贝叶斯定理：

后验概率：给定观察向量 x，某个特定类别的概率 $P(y|x)$.

Bayésienne： $P(y,x) = P(y|x)P(x) = P(x|y)P(y)$

似然 Likelihood　Class Prior Probability，类先验概率

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} = \frac{P(x|y)P(y)}{\sum_i P(x|y_i)P(y_i)}$$

贝叶斯定理：

因此，我们选择最大后验概率的结果作为预测结果

$$y^* = \underset{i}{\operatorname{argmax}}\ P(y_i|x)$$

那么，我们犯错的可能性有多大？

$$P(err|x) = \begin{cases} P(y_2|x), & \text{但是决策为 } y_1; \\ P(y_2|x), & \text{但是决策为 } y_2. \end{cases}$$

$$P(err|x) = \operatorname{argmin}\left[P(y_1|x), P(y_2|x)\right]$$

预测时值作为 P_e，当决策数为时，为剩下的概率求和.

错不同的错误会带来不同的损失，λ_{ij}：预测 ← 真实

条件风险（期望损失）：$R(y_i|x) = \sum_{j=1}^{n} \lambda_{ij} P(y_j|x)$

决策规则：预测 y_i 的期望损失 = 预测为 y_i 但实际是 y_j 的损失条数 λ_{ij} × y_j 发生的概率.

0-1 条件风险：$R(y_i|x) = 1 - P(y_j|x)$

贝叶斯最优分类：对每一个样本都选取风险最小的决策，我们就可以得到最优分类.

$$h^*(x) = \underset{y \in Y}{\operatorname{argmax}}\ P(y|x)$$ ……即最大后验概率（在 0-1 条件风险中）.

朴素贝叶斯：条件独立性假设：$P(X=x \mid Y=c_k) = P(X^{(1)}=x^{(1)} \dots X^{(n)}=x^{(n)} \mid Y=c_k)$

eg. $P(x_1, x_2, x_3) = P(x_1) \cdot P(x_2) \cdot P(x_3)$

$$= \prod_{i=1}^{n} P(X^{(i)}=x^{(i)} \mid Y=c_k)$$

贝叶斯定理：$P(Y=c_k \mid X=x) = \dfrac{P(X=x \mid Y=c_k) \cdot P(Y=c_k)}{\sum_k P(X=x \mid Y=c_k) P(Y=c_k)}$

代入上式，得 $P(Y=c_k \mid X=x) = \dfrac{P(Y=c_k) \prod_j P(X^{(j)}=x^{(j)} \mid Y=c_k)}{\sum_k P(Y=c_k) \prod_j P(X^{(j)}=x^{(j)} \mid Y=c_k)}$

TD. Exercice 1:

① 假设一个最大后验概率 $d^*(x)$

② 计算出一个 x 的临界值 α.

③ 计算条件风险概率 P_e.

1: 假设 $x \in \omega_1$. 且 $d^*(x) = f(x \mid \omega_1) P(\omega_1)$

$\therefore f(x \mid \omega_1) P(\omega_1) \geq f(x \mid \omega_2) P(\omega_2)$

又：这两个分类 ω_1 和 ω_2 同概率.

$\therefore f(x \mid \omega_1) \geq f(x \mid \omega_2) \Rightarrow \dfrac{x}{\sigma_1^2} e^{\frac{-x^2}{2\sigma_1^2}} \amalg_R \geq \dfrac{x}{\sigma_2^2} e^{\frac{-x^2}{2\sigma_2^2}} \amalg_R$

$\Rightarrow -\ln(\sigma_1^2) - \dfrac{x^2}{2\sigma_1^2} \geq -\ln(\sigma_2^2) - \dfrac{x^2}{2\sigma_2^2} \Rightarrow x^2 \left(\dfrac{1}{2\sigma_2^2} - \dfrac{1}{2\sigma_1^2}\right) \geq \ln(\sigma_1^2) - \ln(\sigma_2^2)$

$\Rightarrow x^2 \geq \dfrac{2\sigma_2^2 \sigma_1^2}{\sigma_1^2 - \sigma_2^2} \cdot \ln\left(\dfrac{\sigma_1^2}{\sigma_2^2}\right) = \alpha^2 \Rightarrow \alpha = \sqrt{\dfrac{2\sigma_1^2 \sigma_2^2}{\sigma_1^2 - \sigma_2^2} \cdot \ln\left(\dfrac{\sigma_1^2}{\sigma_2^2}\right)}$

2: 条件风险 $P_e = \sum_{i=1}^{2} \lambda_{ij} P(y_j \mid x)$

$= \frac{1}{2} \times P(x \geq \alpha \mid \omega_2) + \frac{1}{2} \times P(x < \alpha \mid \omega_1)$

Examen . AD　　Analyse en Composantes Principales

ACP et Classification (解题过程)

① 求X, 以及 $\bar{X}$

$$X = \begin{bmatrix} -4 & -2 \\ -3 & -1 \\ -1 & 0 \\ 2 & 0 \\ 2 & 1 \\ 4 & 2 \end{bmatrix}, \quad \bar{X} = \left[\tfrac{1}{6}(-4-3-1+2+2+4),\ \tfrac{1}{6}(-2-1+0+0+1+2)\right] = [0,\ 0]$$

② 求出 X_c

$$X_c = X - \bar{X} = X$$

③ 求 Σ Calcul la matrice de variance-covariance Σ

$$\Sigma = \frac{1}{n} X_c^T X_c = \frac{1}{6}\begin{bmatrix} -4 & -3 & -1 & 2 & 2 & 4 \\ -2 & -1 & 0 & 0 & 1 & 2 \end{bmatrix}\begin{bmatrix} -4 & -2 \\ -3 & -1 \\ -1 & 0 \\ 2 & 0 \\ 2 & 1 \\ 4 & 2 \end{bmatrix} = \frac{1}{6}\begin{bmatrix} 50 & 21 \\ 21 & 10 \end{bmatrix}$$

④ 计算主向量 Le premier vecteur principale.

设系数 λ, $\det(6\Sigma - \lambda I_d) = \begin{vmatrix} 50-\lambda & 21 \\ 21 & 10-\lambda \end{vmatrix} = (\lambda - 59)(\lambda - 1) = 0$

取最大的λ作为主向量.

得 $\lambda_1 = 59$, $\lambda_2 = 1$

Le premier vecteur principal est associé la plus grande valeur λ.

故 $\begin{bmatrix} 50-59 & 21 \\ 21 & 10-59 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = 0 \Rightarrow y = \frac{3}{7}x$

故 $v_1 = \begin{bmatrix} 1 \\ 3/7 \end{bmatrix}$

⑤ 计算相应的主分量 La composante principale correspondante

Calcul de la composante principale :

$$X_c \cdot v_1 = \begin{bmatrix} -4 & -2 \\ -3 & -1 \\ -1 & 0 \\ 2 & 0 \\ 2 & 1 \\ 4 & 2 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 3/7 \end{bmatrix} = \begin{bmatrix} -34/7 \\ -24/7 \\ -1 \\ 2 \\ 17/7 \\ 34/7 \end{bmatrix}$$

a −34
b −24
c −7
d 14
e 17
f 34

⑥ 分类 Classification
确定阈值 Seuil.
阶数 k

$k = 1$, seuil $= 20/7$. 我们可将所有数据乘7. 我们可作出以下图表:

① 列 ② 行

dist (距离)	a	b	c	d	e	f
a	0	10	27	48	51	68
b		0	17	38	41	58
c			0	21	24	41
d				0	3	20
e					0	17
f						0

- dist a→b = 10 < 20, 其它不满足
  → $C_1 = \{a, b\}$ 可理解为集合(a,b)
- dist c→b = 17 < 20, 其它不满足
  ⇒ → 将c加入 C_1, $C_1 = \{a, b, c\}$
- dist e→d = 3 < 20, 其它不满足
  → 新建 $C_2 = \{e, d\}$
- dist f→e = 17 < 20,
  → 将f加入 C_2, $C_2 = \{e, d, f\}$

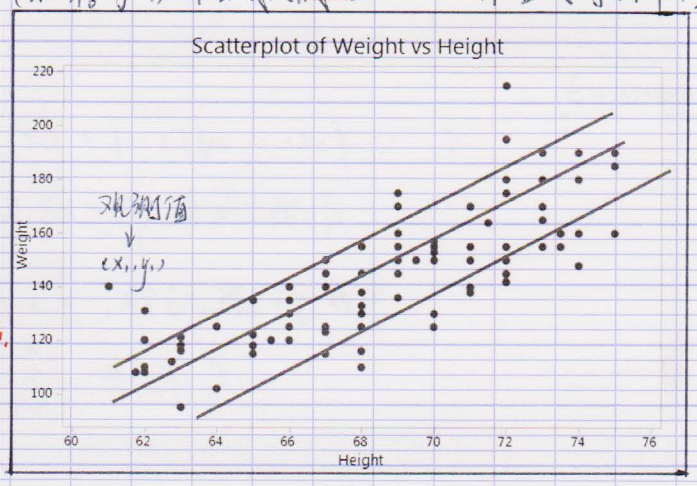
Examen. A.D.

Moindres carrés (最小二乘法)

Modèle:

构建一元回归方程: $\hat{y} = \beta_0 + \beta_1 x$ (1)

($\hat{y}$: 因变量的拟合值; x: 自变量的观察值; $x=0, \beta_0 = \hat{y} = ?$ 回归直线的截距; β_1: 回归直线的斜率 $(\frac{\Delta y}{\Delta x})$)



当拟合值与观测值的误差为最小时，拟合的线是最佳的。

综合所有点的误差

课件的方法:

上述的回归方程(1)可以写成自变量 x 和参数 β 的方程 $f(x, \beta)$

$$f(x, \beta) = \sum_{k=1}^{m} \beta_k \phi_k(x) \quad <2>$$

由于拟合值和观察值的误差必须体现出所有点的观察值与拟合值的差，因此要求和。用 S 来表示误差的平方和，则有：

$S(\beta_0, \beta_1)$ 计算最小值.

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \quad \text{pour } <1>$$

$$S(\beta_k, k=1,2,3\ldots n) = \sum_{i=1}^{n} \left[y_i - \beta_1 \phi_1(x) - \beta_2 \phi_2(x) - \ldots \beta_3 \phi_3(x) \right] \quad <2>$$

为了让 $S(\beta_0, \beta_1)$ 取得最小值，将 $S(\beta_0, \beta_1)$ 分别对 β_0 和 β_1 求偏导数，并令它们等于0. 联立可得：

导数为0时取最大值*（一般来讲）

$$\begin{cases} \dfrac{\partial S}{\partial \beta_0} = \dfrac{\partial \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_0} = 0 \\ \dfrac{\partial S}{\partial \beta_1} = \dfrac{\partial \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_1} = 0 \end{cases} \rightarrow \text{正规方程组 (Normal Equations)}$$

它的解为：

$$\begin{cases} \beta_1 = \dfrac{n \sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \\ \beta_0 = \bar{y} - \beta_1 \bar{x} \end{cases} \quad \text{或} \quad \begin{cases} \beta_1 = \dfrac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\ \beta_0 = \bar{y} - \beta_1 \bar{x} \end{cases}$$

其中，$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$，$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$，分别是观测值 x 和 y 的均值.

此时，我们需要求出矩阵 $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ 以解出回归直线以及 $S(\beta_0, \beta_1)$

详见课件 « Moindres carrés (cf TD n°2) » - Partie 2.

TD. Exercice 3.

1. ∵ $f(t) = a\sqrt{|t-1|} + bt^2$ 中有2个未知数，$f(0)=1$，$f(1)=3$，$f(2)=7$

∴ 不能确定唯一的 a，b 值.

2. $\hat{y}$ 为 y 的预测值，$\hat{y} = \beta_0 + \beta_1 x = a\sqrt{|x-1|} + bx^2$

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{3} (y_i - a\sqrt{|x_i - 1|} - bx_i^2)$$

$$= (1-a)^2 + (3-b)^2 + (7-a-4b)^2$$

3. 对 S 分别对 β_0，β_1 求偏导.

误差计算

$$\hat{y} = \beta_0 + \beta_1 x$$

如何合理地计算误差呢?

Answer:
拟合值和观察值的误差必须体现出所有点的观察值和拟合值的差，因此需要进行求和。
因为数据散落在回归直线的上下两侧，为了防止正负误差相互抵消，因此将误差进行平方后再求和。
也就是说当$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$最小时，曲线是最佳的估计，用$S$来表示误差的平方和，则有:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

正规方程组

$\hat{y} = \beta_0 + \beta_1 x$ $S(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$

为了让$S(\beta_0, \beta_1)$取得最小值，将$S(\beta_0, \beta_1)$分别对β_0和β_1求偏导数，并令它们为0，联立可得:

$$\begin{cases} \frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0 \end{cases}$$

上面的这个公式称为正规方程组(Normal Equations)，它的解为

$$\begin{cases} \beta_1 = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \\ \beta_0 = \bar{y} - \beta_1 \bar{x} \end{cases} \quad \text{或者} \quad \begin{cases} \beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\ \beta_0 = \bar{y} - \beta_1 \bar{x} \end{cases}$$

其中$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$， $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$，分别是观测值$x$和$y$的均值

Examen. A.D.

Arbre de décision. (avec l'indice de Gini)

决策树（基尼指数相关）

相关概念：

信息熵（Entropy）：

$$H(X) = -\sum_{i=1}^{n} P(X=i) \log_2 P(X=i)$$

基尼指数（l'indice Gini）：

基尼指数表示在样本集合中一个随机选中的样本被分错的概率。

Gini 指数越高 ↑ ⇒ 样本被分错的概率越小 ↓

$$Gini(p) = \sum_{i=1}^{n} P_i \cdot (1-P_i) = 1 - \sum_{i=1}^{n} P_i^2$$

其中，P_i 表示选中的样本属于第 i 个的概率。

（解题过程）

① l'indice de Gini associé à base d'apprentissage

计算根结点的 Gini 指数：

$$Gini(p) = \sum_{i=1}^{n} P_i(1-P_i) = \frac{1}{2}\times(1-\frac{1}{2}) + \frac{1}{2}\times(1-\frac{1}{2}) = \frac{1}{2}$$

② Les indices de Gini associé à chaque variable.

计算加取 Gini 指数：

在 "Ciel" 中：

Ciel	Oui	Non	Fréquence	P
Soleil	2	0	2	2/6
couvert	0	1	1	1/6
pluie	1	2	3	3/6

②.1 列出在当前情况下，各个取值的情况

②.2 计算 Gini 增益

分类的个数 $\sum_{i=1}^{} P_i \cdot Gini(\text{次级1}, \text{次级2})$

$= \sum P_i \cdot Gini(1 - P_{\text{次级1}}^2 - P_{\text{次级2}}^2)$

与 "Ciel" 相关的 Gini 指数为：（Gini 增益）

$$\frac{2}{6}Gini(2,0) + \frac{1}{6}Gini(0,1) + \frac{3}{6}Gini(1,2)$$

$$= \frac{2}{6}\times(1-1^2-0^2) + \frac{1}{6}\times(1-0^2-1^2) + \frac{3}{6}\times[1-(\frac{1}{3})^2-(\frac{2}{3})^2] = 0 + 0 + \frac{3}{6}\times\frac{4}{9} = \frac{2}{9}$$

在 "Température" 中：

Température	Oui	Non	F	P
chaud	2	1	3	3/6
froid	1	2	3	3/6

与 "Température" 相关的 Gini 指数为：（Gini 增益）

$$\frac{3}{6}Gini(2,1) + \frac{3}{6}Gini(1,2) = \frac{3}{6}\times[1-(\frac{2}{3})^2-(\frac{1}{3})^2] + \frac{3}{6}\times[1-(\frac{1}{3})^2-(\frac{2}{3})^2]$$

$$= \frac{3}{6}\times\frac{4}{9} + \frac{3}{6}\times\frac{4}{9}$$

$$= \frac{4}{9}$$

在"Vent"中：

Vent	Oui	Non	Σ	P
faible	1	3	4	$4/6$
fort	2	0	2	$2/6$

与"Vent"相关的 Gini 指数为：(Gini 增益)

$$\frac{4}{6}\,Gini(1,3)+\frac{2}{6}\,Gini(2,0)=\frac{4}{6}\times\left[1-\left(\frac{1}{4}\right)^2-\left(\frac{3}{4}\right)^2\right]+\frac{2}{6}\left[1-\left(\frac{2}{2}\right)^2-\left(\frac{0}{2}\right)^2\right]$$

$$=\frac{4}{6}\times\frac{6}{16}+\frac{2}{6}\times 0$$

$$=\frac{1}{4}$$

CART 树：从根节点，t=1 时开始，从所有可能属性 S 集合中搜索，

使 Gini 增益最大的划分 S，然后，使用划分后的 S 被划分

成两个结点，t=2，t=3，在 t=2，t=3 节点上继续划分迭代。

在上题中，$Gini\_ciel=\frac{2}{7}$

$Gini-temperature=\frac{4}{7}$ (max，最佳分解特征)

$Gini-vent=\frac{1}{4}$

]343

5.1 公式(3)

$$A\beta = b \Rightarrow (A+E)\beta = b+z$$

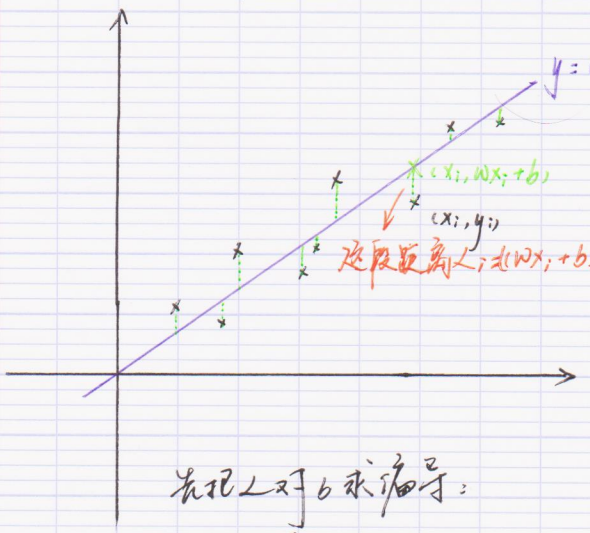
$$[A \quad b]$$

为SVD:

$$[A+\underbrace{\Delta A}_{E} \quad b+\underbrace{\Delta b}_{z}]$$

$$[A \quad b] = [U_A \quad U_b]\begin{bmatrix}\Sigma_A & \\ & \Sigma_b\end{bmatrix}\begin{bmatrix} \\ \end{bmatrix} = 0$$

最小二乘法(解决数据拟合)



直线 $y = wx+b$ 为要拟合出的结果

L 为所有 i 的和, 即:

$$L = \sum_{i=1}^{n}[(wx_i+b)-y_i]^2$$

② 距离为正, 故要取平方.

预测值 y_i 与实际值 y 距离的总和.

目的: 找出 L 的最小值.

先把 L 对 b 求偏导:

$$\frac{\partial L}{\partial b} = \frac{\partial \sum_{i=1}^{n}[(wx_i+b)-y_i]^2}{\partial b}$$

$$= 2\left[\sum_{i=1}^{n} wx_i + \sum_{i=1}^{n} b \cdot 0 - \sum_{i=1}^{n} y_i\right] \quad ①$$

令 $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \Rightarrow n\bar{x} = \sum_{i=1}^{n} x_i$

①式可化简为: $2[0 \cdot n\bar{x} + nb - n\bar{y}]$

令导数 $= 0$, 即 $2n(w\bar{x}+b-\bar{y}) = 0$

得 $b = \bar{y} - w\bar{x}$ ③

接着对 w 求偏导:

$$\frac{\partial L}{\partial w} = \frac{\partial \sum_{i=1}^{n}[(wx_i+b)-y_i]^2}{\partial w}$$

$$= \sum_{i=1}^{n} 2(wx_i+b-y_i)x_i$$

结合③式, 化简得

原式 $= 2\left[w(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) + n\bar{x}\bar{y} - \sum x_i y_i\right]$

令 $\frac{\partial L}{\partial w} = 0 \Rightarrow w = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$