2016-2017 学年秋季学期　　试题专用纸

中国科学院大学

试 题 专 用 纸

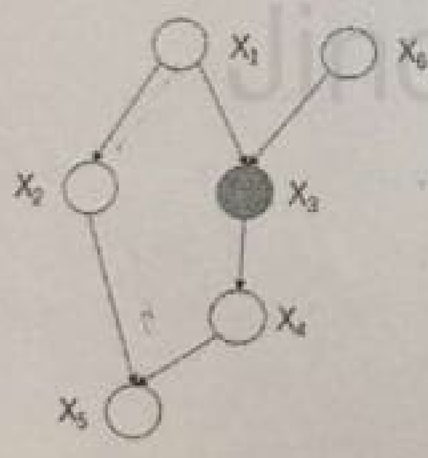
课程编号：091M4042H
课程名称：模式识别与机器学习
任课教师：黄庆明、山世光、兰艳艳、郭嘉丰

注意事项：
1.考试时间为 120 分钟，考试方式 闭 卷；
2.全部答案写在答题纸上；
3.考试结束后，请将本试卷和答题纸、草稿纸一并交回。

1. （6 分）简述模式的概念和它的直观特性，并简要说明模式分类有哪几种主要方法。

2. （8分）假设某研究者在ImageNet数据上使用线性支持向量机（Linear SVM）来做文本分类的任务，请说明在如下情况下分别如何操作才能得到更好的结果，并说明原因。
（1）训练误差5%，验证误差10%，测试误差10%。
（2）训练误差1%，验证误差10%，测试误差10%。
（3）训练误差1%，验证误差3%，测试误差10%。

3. （8分）给定如下概率图模型，其中变量X_3为已观测变量，请问变量X_4和X_6是否独立？并用概率推导证明之。



4. （10 分）（1）随机猜测作为一个分类算法是否一定比 SVM 差？借此阐述你对“No Free Lunch Theorem”的理解。（2）举例阐述你对“Occam's razor”的理解。

5. （10 分）详细描述 AdaBoost 的原理并给出算法，并解释为什么 AdaBoost 经常可以在训练误差为 0 后继续训练还可能带来测试误差的继续下降。

6. （10 分）用感知器算法求下列模式分类的解向量（取 w(1)为零向量）
ω_1: $\{(0\ 0\ 0)^T, (1\ 0\ 0)^T, (1\ 0\ 1)^T, (1\ 1\ 0)^T\}$
ω_2: $\{(0\ 0\ 1)^T, (0\ 1\ 1)^T, (0\ 1\ 0)^T, (1\ 1\ 1)^T\}$

7. （12 分）设以下模式类别具有正态概率密度函数：

ω_1：$\{(0\ 0\ 0)^T,\ (1\ 0\ 0)^T,\ (1\ 0\ 1)^T,\ (1\ 1\ 0)^T\}$

ω_2：$\{(0\ 1\ 0)^T,\ (0\ 1\ 1)^T,\ (0\ 0\ 1)^T,\ (1\ 1\ 1)^T\}$

若 $P(\omega_1)=P(\omega_2)=0.5$，求这两类模式之间的贝叶斯判别界面的方程式。

8. （12 分）假设有如下线性回归问题，

$$\min_{\beta}(y-X\beta)^2+\lambda||\beta||_2^2$$

其中y和β是n维向量，X是一个m × n的矩阵。

该线性回归问题的参数估计可看作一个后验分布的均值，其先验为高斯分布$\beta\sim N(0,\tau I)$，样本产生自高斯分布$y\sim N(X\beta,\sigma^2 I)$，其中$I$为单位矩阵，试推导调控系数$\lambda$与方差$\tau$和$\sigma^2$的关系。

9. （12 分）给定有标记样本集$D_l=\{(x_1,y_1),(x_2,y_2),\dots,(x_l,y_l)\}$和未标记样本$D_u=\{(x_{l+1},y_{l+1}),(x_{l+2},y_{l+2}),\dots,(x_{l+u},y_{l+u})\}, l\ll u, l+u=m$，假设所有样本独立同分布，且都是由同一个包含 N 个混合成分的高斯混合模型$\{(\alpha_i,\mu_i,\Sigma_i)|1\le i\le N\}$产生，每个高斯混合成分对应一个类别，请写出极大似然估计的目标函数（对数似然函数），以及用 EM 算法求解参数的迭代更新式。

10. （12 分）假定对一类特定人群进行某种疾病检查，正常人以 ω_1类代表，患病者以ω_2类代表。设被检查的人中正常者和患病者的先验概率分别为

正常人：$P(\omega_1)=0.9$

患病者：$P(\omega_2)=0.1$

现有一被检查者，其观察值为 x，从类条件概率密度分布曲线上查得

$P(x|\omega_1)=0.2$，$P(x|\omega_2)=0.4$

同时已知风险损失函数为

$$\begin{pmatrix}\lambda_{11} & \lambda_{12}\\ \lambda_{21} & \lambda_{22}\end{pmatrix}=\begin{pmatrix}0 & 6\\ 1 & 0\end{pmatrix}$$

其中λ_{ij}表示将本应属于第 j 类的模式判为属于第 i 类所带来的风险损失。试对该被检查者用以下两种方法进行分类：

（1）　基于最小错误率的贝叶斯决策，并写出其判别函数和决策面方程；

（2）　基于最小风险的贝叶斯决策，并写出其判别函数和决策面方程。

2016-2017

1. 模式是抽取自物体的信息集合，既包含空间部分，又包含时间部分。

直观特性：可观察性、可区分性、相似性

主要方法：监督学习：概念驱动，归纳假说

非监督学习：数据驱动，演绎假说

2. (1) 欠拟合，换用复杂度更高的模型

(2) 过拟合，换用复杂度更低的模型

(3) 测试数据与训练数据不是独立同分布的，更换测试数据集

4. (1) 不一定，在无先验知识的情况下，无法断言一个模型比另一个更好。

对特定的问题，为了获得更好的效果需要使用更复杂的模型。

(2) 训练数据来自添加高斯噪声的 $y=\sin x$ $(x\in[0,2\pi])$，

使用不同的多项式函数拟合，三次的效果最佳。在同等训练错误率的条件下，简单模型具有更小的方差，更好的泛化能力。

5. $\{x_i, y_i\}_{i=1}^{n}$

$D(i)=\frac{1}{n}$

for 1 to T:

训练弱分类器 h_i

$D_{(i)}^{t+1} = D_{(i)}^{t} \cdot e^{-\alpha_i y_i h_i(x_i)}$

通过弱分类器的组合，得到强分类器。每次训练弱分类器后，对分类错误的样本，增加权重使得后续的分类器更加"关注"该样本，以达到提升分类效果的目的。

其中 $\alpha_i = \frac{1}{2}\ln\frac{1-\varepsilon}{\varepsilon}$，$\varepsilon = P(h(x)\neq y) < 0.5$

$H_{final} = \sum \alpha_i h_i(x)$

当训练误差为零后，AdaBoost会继续增大分类间距，提升模型的泛化能力，减少测试误差。

6. $W(1) = (0,0,0,0)$，对 ω_2 的样本取相反数

增广数据：$\omega_1: \{(0,0,0,1)^T, (1,0,0,1)^T, (1,0,1,1)^T, (1,1,0,1)^T\}$

$\omega_2: \{(-0,-0,-1,-1)^T, (-0,-1,-1,-1)^T, (-0,-1,-0,-1)^T, (-1,-1,-1,-1)^T\}$

更新规则：$W(k+1) = \begin{cases} W(k) & W(k)^T X(k) > 0 \\ W(k) + X(k+1) & \text{其它} \end{cases}$ （以下省略不更新的步骤）

$X(1) = (0,0,0,1)^T$，$W(1)^T X(1) = 0$ 更新：$W(2) = (0,0,0,1)^T$

$X(2) = (0,0,-1,-1)^T$，$W(2)^T X(2) = -1$ 更新：$W(3) = (0,0,-1,0)^T$

$X(3) = (0,-1,0,-1)^T$，$W(3)^T X(3) = 0$ 更新：$W(4) = (0,-1,-1,-1)^T$

$X(4) = (0,0,0,1)^T$，$W(4)^T X(4) = -1$，更新：$W(5) = (0,-1,-1,0)^T$

$X(5) = (1,0,0,1)^T$，$W(5)^T X(5) = 0$，更新：$W(6) = (1,-1,-1,1)^T$

$X(6) = (0,0,-1,-1)^T$，$W(6)^T X(6) = 0$，更新：$W(7) = (1,-1,-2,0)^T$

$X(7) = (0,0,0,1)^T$，$W(7)^T X(7) = 0$，更新：$W(8) = (1,-1,-2,1)^T$

$X(8) = (1,0,1,1)^T$，$W(8)^T X(8) = 0$，更新：$W(9) = (2,-1,-1,2)^T$

$X(9) = (0,0,-1,-1)^T$，$W(9)^T X(9) = -1$，更新：$W(10) = (2,-1,-2,1)^T$

$X(10) = (0,-1,0,-1)^T$，$W(10)^T X(10) = 0$，更新：$W(11) = (2,-2,-2,0)^T$

$X(11) = (0,0,0,1)^T$，$W(11)^T X(11) = 0$，更新：$W(12) = (2,-2,-2,1)^T$

$W = W(12) = (2,-2,-2,1)$

7. $p(\omega_1|x) = \dfrac{p(x|\omega_1)p(\omega_1)}{p(x)}$

$p(\omega_2|x) = \dfrac{p(x|\omega_2)p(\omega_2)}{p(x)}$

$p(\omega_1|x) > p(\omega_2|x)$

$p(x|\omega_1)p(\omega_1) > p(x|\omega_2)p(\omega_2)$

$\dfrac{p(x|\omega_1)}{p(x|\omega_2)} - \dfrac{p(\omega_2)}{p(\omega_1)} > 0$

$d(x) = \dfrac{p(x|\omega_1)}{p(x|\omega_2)} - \dfrac{p(\omega_2)}{p(\omega_1)} = \dfrac{p(x|\omega_1)}{p(x|\omega_2)} - 1 = 0$

假设给定 ω 的情况下，x 的各分量相互独立.

$p(x|\omega) = p(x_1, x_2, x_3|\omega) = p(x_1|\omega)p(x_2|\omega)p(x_3|\omega)$

$dx = \dfrac{\prod_{i=1}^{3} p(x_i|\omega_1)}{\prod_{i=1}^{3} p(x_i|\omega_2)} - 1 = 0$

2018-1-19 22:14

8.

$$p(\vec{\beta}|\vec{y}) = \frac{p(\vec{y}|\vec{\beta},\vec{x},\sigma)\cdot p(\vec{\beta}|\tau)}{p(\vec{y})}$$

对数似然：$\log p(\vec{\beta}|\vec{y}) = \log p(\vec{y}|\vec{\beta},\vec{x},\sigma) + \log p(\vec{\beta}|\tau) - \log p(\vec{y})$

$$= \log \prod_{i=1}^{n} p(y_i|\vec{\beta},\vec{x}_i,\sigma) + \log \frac{1}{\sqrt{2\pi|\tau I|}} \cdot e^{-\frac{1}{2}(\vec{\beta}-\vec{0})^T(\tau I)^{-1}(\vec{\beta}-\vec{0})} - \log p(\vec{y})$$

$$= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi|\sigma I|}} e^{-\frac{1}{2}(y_i-\vec{x}_i^T\vec{\beta})(\sigma I)^{-1}(y_i-\vec{x}_i^T\vec{\beta})} + \left(-\frac{1}{2}\vec{\beta}^T(\tau I)^{-1}\vec{\beta}\right) + \text{constant}$$

$$= \sum_{i=1}^{n} -\frac{1}{2}(y_i-\vec{x}_i^T\vec{\beta})^T(\sigma I)^{-1}(y_i-\vec{x}_i^T\vec{\beta}) - \frac{1}{2}\vec{\beta}^T(\tau I)^{-1}\vec{\beta} + \text{constant}$$

$$= \sum_{i=1}^{n} -\frac{1}{2}\cdot\frac{1}{\sigma}(y_i-\vec{x}_i^T\vec{\beta})^T(y_i-\vec{x}_i^T\vec{\beta}) - \frac{1}{2}\cdot\frac{1}{\tau}\vec{\beta}^T\vec{\beta} + C.$$

$$= -\frac{1}{2\sigma}(y-X\beta)^2 - \frac{1}{2\tau}\|\beta\|^2 + C$$

$$= -\frac{1}{2\sigma}\left[(y-X\beta)^2 + \frac{\sigma}{\tau}\|\beta\|^2\right] + C$$

$\max \log p(\vec{\beta}|\vec{y})$ 等价于 $\min (y-X\beta)^2 + \frac{\sigma}{\tau}\|\beta\|^2$ $\therefore \lambda = \frac{\sigma}{\tau}$.

9. $D_1 = \{(x_i,y_i)\}_{i=1}^{l}$ 已知，$D_2 \{(x_i,y_i)\}_{i=l+1}^{m}$ 未知

令 $z_{ik} = I(x_i, y_i = k)$. $\forall i$ $l+1 \le i \le m$，易知 $z_{ik} \sim B(1, p(y_i=k|x_i))$。

E-step：$E(z_{ik}) = p(y_i=k|x_i) = \frac{p(x_i|y_i=k)p(y_i=k)}{p(x)} = \frac{p(x_i|y_i=k)p(y_i=k)}{\sum_{j=1}^{n} p(x_i|y_i=j)p(y_i=j)}$

M-step：$\prod_{i=1}^{m} p(x_i,y_i) = \prod_{i=1}^{l} p(x_i,y_i)\cdot\prod_{i=l+1}^{m} p(x_i,y_i)$

$\forall i$, $l+1 \le i \le m$ $p(x_i,y_i) = \prod_{j=1}^{n} p(x_i,y_i=j)^{z_{ij}}$

$$\prod_{i=1}^{m} p(x_i,y_i) = \prod_{i=1}^{l} p(x_i,y_i)\cdot\prod_{i=l+1}^{m}\prod_{j=1}^{n} p(x_i,y_i=j)^{z_{ij}}$$

~~对数似然：~~ $\cancel{\sum_{i=1}^{l}\left[\log p(x_i,y_i) + \sum_{i=l+1}^{m}\sum_{j=1}^{n} z_{ij}(\log p(x_i|y_i=j) + \log p(y_i=j))\right]}$

$$\mu_i = \frac{\sum_{j=1}^{l} x_j I(y_j=i)}{\sum_{j=1}^{l} I(y_j=i)} + \frac{\sum_{j=l+1}^{m} x_j\cdot z_{ij}}{\sum_{j=l+1}^{m} z_{ij}} + \cdots$$

$$\Sigma_i = \frac{1}{\sum_{j=1}^{l} I(y_j=i)}\sum_{j=1}^{l}(x_j-\mu_i)^2 I(y_j=i) + \frac{1}{\sum_{j=l+1}^{m} z_{ij}}\sum_{j=l+1}^{m}(x_j-\mu_i)^2\cdot z_{ij}$$

$$p(y=i) = \frac{\sum_{j=1}^{l} I(y_j=i)}{l} + \frac{\sum_{j=l+1}^{m} z_{ij}}{m-l}$$

2018-1-19 22:14

10. (1) $p(w_1|x) = \frac{p(x|w_1)\cdot p(w_1)}{p(x)}$

$p(w_2|x) = \frac{p(x|w_2)\, p(w_2)}{p(x)}$

$p(w_1|x) > p(w_2|x)$

$p(x|w_1)\,p(w_1) > p(x|w_2)\,p(w_2)$

$\frac{p(x|w_1)}{p(x|w_2)} - \frac{p(w_2)}{p(w_1)} > 0$

$d(x) = \frac{p(x|w_1)}{p(x|w_2)} - \frac{p(w_2)}{p(w_1)} = 0$

$d(x) = \frac{0.2}{0.4} - \frac{0.9}{0.1} = -7 < 0$

$\therefore x \in w_2$.

(2) $r_1 = L_{11}\cdot p(x, w_1) + L_{12}\cdot p(x, w_2)$

$= 0\cdot p(x|w_1)\,p(w_1) + 6\cdot p(x|w_2)\,p(w_2)$

$= 0 + 6\times 0.4\times 0.1$

$= 0.24$

$r_2 = L_{22}\cdot p(x, w_2) + L_{21}\, p(x, w_1)$

$= 0\cdot p(x|w_2)\,p(w_2) + 1\times p(x|w_1)\,p(w_1)$

$= 0 + 1\times 0.2\times 0.9$

$= 0.18$

$r_2 = 0.18 < 0.24 = r_1$

$\therefore x \in w_2$