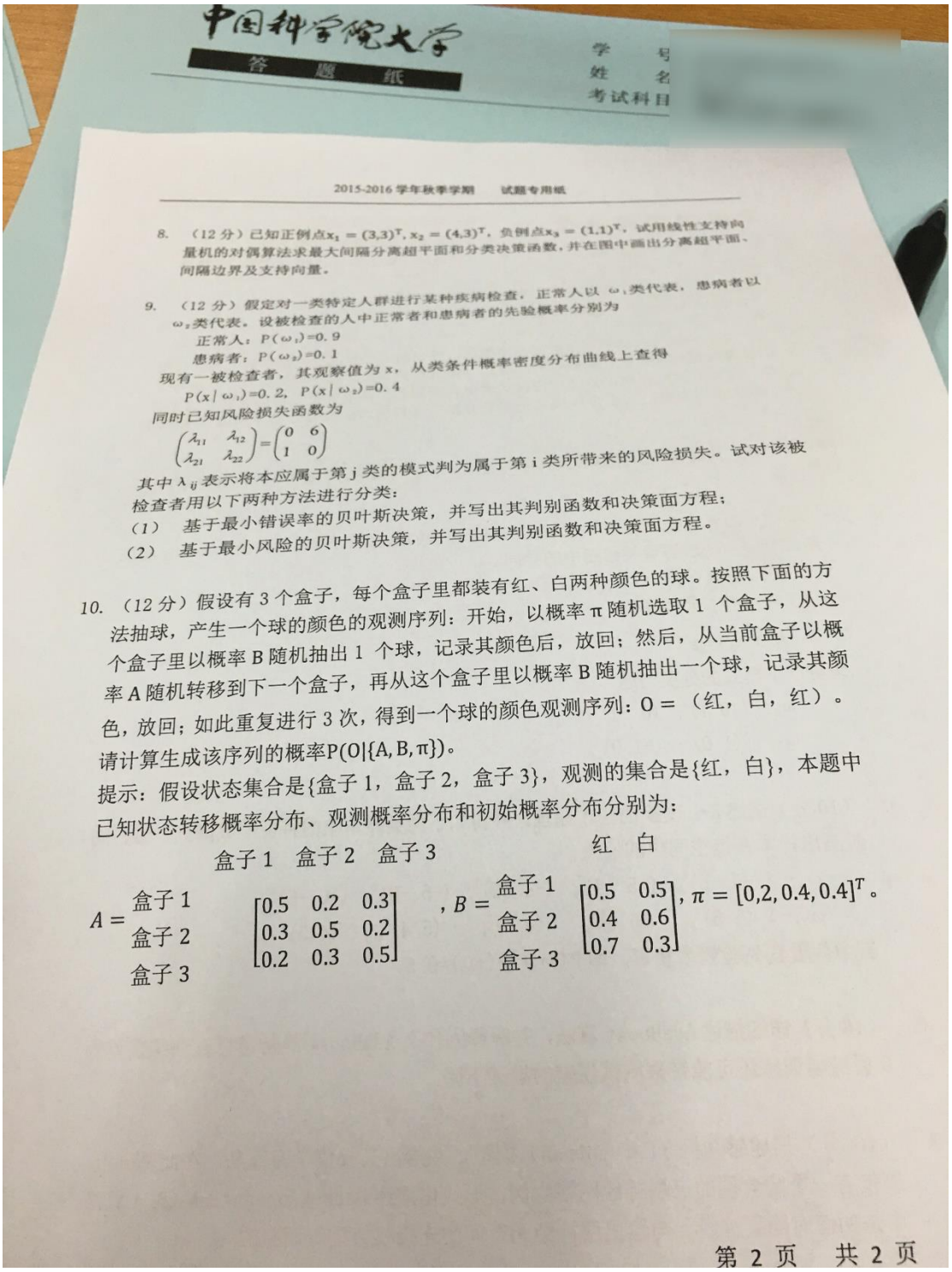
中国科学院大学 黄庆明老师 模式识别与机器学习 2016 试卷 拍摄版 画质清晰 仅供参考



中国科学院大学
答 题 纸

学 号
姓 名
考试科目

2015-2016 学年秋季学期 试题专用纸

8. （12 分）已知正例点$x_1=(3,3)^T$, $x_2=(4,3)^T$，负例点$x_3=(1,1)^T$，试用线性支持向量机的对偶算法求最大间隔分离超平面和分类决策函数，并在图中画出分离超平面、间隔边界及支持向量。

9. （12 分）假定对一类特定人群进行某种疾病检查，正常人以 $\omega_1$ 类代表，患病者以 $\omega_2$ 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 $\lambda_{ij}$ 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。

10. （12 分）假设有 3 个盒子，每个盒子里都装有红、白两种颜色的球。按照下面的方法抽球，产生一个球的颜色的观测序列：开始，以概率 π 随机选取 1 个盒子，从这个盒子里以概率 B 随机抽出 1 个球，记录其颜色后，放回；然后，从当前盒子以概率 A 随机转移到下一个盒子，再从这个盒子里以概率 B 随机抽出一个球，记录其颜色，放回；如此重复进行 3 次，得到一个球的颜色观测序列：O =（红，白，红）。

请计算生成该序列的概率$P(O|\{A,B,\pi\})$。

提示：假设状态集合是{盒子 1，盒子 2，盒子 3}，观测的集合是{红，白}，本题中已知状态转移概率分布、观测概率分布和初始概率分布分别为：

$$A=\begin{array}{c}\\ \text{盒子 1}\\ \text{盒子 2}\\ \text{盒子 3}\end{array}\begin{array}{c}\begin{array}{ccc}\text{盒子 1} & \text{盒子 2} & \text{盒子 3}\end{array}\\ \begin{bmatrix}0.5 & 0.2 & 0.3\\ 0.3 & 0.5 & 0.2\\ 0.2 & 0.3 & 0.5\end{bmatrix}\end{array},\quad B=\begin{array}{c}\\ \text{盒子 1}\\ \text{盒子 2}\\ \text{盒子 3}\end{array}\begin{array}{c}\begin{array}{cc}\text{红} & \text{白}\end{array}\\ \begin{bmatrix}0.5 & 0.5\\ 0.4 & 0.6\\ 0.7 & 0.3\end{bmatrix}\end{array},\quad \pi=[0.2,0.4,0.4]^T。$$

第 2 页 共 2 页

学　号
姓　名
考试科目

2015-2016 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：091M4043H
课程名称：模式识别与机器学习
任课教师：黄庆明、山世光、常虹、兰艳艳

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. （8 分）试阐述线性判别函数的基本概念，并说明既然有线性判别函数，为什么还需要非线性判别函数？假设有两类模式，每类包括 5 个 3 维不同的模式，且良好分布。如果它们是线性可分的，问权向量至少需要几个系数分量？假如要建立二次的多项式判别函数，又至少需要几个系数分量？（设模式的良好分布不因模式变化而改变）

2. （8 分）简述偏差方差分解及其推导过程，并说明偏差、方差和噪声三部分的内在含义。

3. （8 分）试描述用 EM 算法求解高斯混合模型的思想和过程，并分析 k-means 和高斯混合模型在求解聚类问题中的异同。

4. （10 分）用下列势函数

$$K(x,x_k)=e^{-\|x-x_k\|^2}$$

求解以下模式的分类问题

$\omega_1$: $\{(0\ 1)^T,\ (0\ -1)^T\}$

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

5. （10 分）试述 K-L 变换的基本原理，并将如下两类样本集的特征维数降到一维，同时画出样本在该空间中的位置。

$\omega_1$: $\{(-5\ -5)^T,\ (-5\ -4)^T,\ (-4\ -5)^T,\ (-5\ -6)^T,\ (-6\ -5)^T\}$

$\omega_2$: $\{(5\ 5)^T,\ (5\ 6)^T,\ (6\ 5)^T,\ (5\ 4)^T,\ (4\ 5)^T\}$，

其中假设其先验概率相等，即 $P(\omega_1)=P(\omega_2)=0.5$。

6. （10 分）详细描述 AdaBoost 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。

7. （10 分）描述感知机（Perceptron）模型，并给出其权值学习算法。在此基础上，以仅有一个隐含层的三层神经网络为例，形式化描述 Back-Propagation（BP）算法中是如何对隐层神经元与输出层神经元之间的连接权值进行调整的。

# 中科院（国科大）黄庆明模式识别与机器学习 2016-2017 试卷带答案（兰艳艳、郭嘉丰、山世光）

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

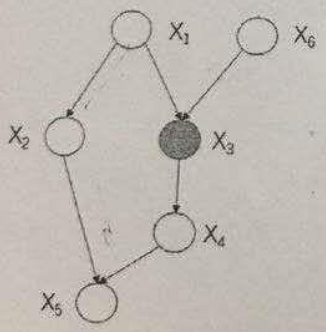
任课教师：黄庆明、山世光、兰艳艳、郭嘉丰

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）
   $\omega_1$: $\{(0\ 0\ 0)^T, (1\ 0\ 0)^T, (1\ 0\ 1)^T, (1\ 1\ 0)^T\}$
   $\omega_2$: $\{(0\ 0\ 1)^T, (0\ 1\ 1)^T, (0\ 1\ 0)^T, (1\ 1\ 1)^T\}$

7. （12 分）设以下模式类别具有正态概率密度函数：

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

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

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$\hat{y}$和$\beta$是n维向量，X是一个$m \times n$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$D_l = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}$和未标记样本$D_u = \{(x_{l+1}, y_{l+1}), (x_{l+2}, y_{l+2}), \ldots, (x_{l+u}, y_{l+u})\}$，$l \ll u, l+u=m$，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 $\omega_1$类代表，患病者以 $\omega_2$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 $\lambda_{ij}$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 $y=\sin x$ $(x\in[0,2\pi])$。训练

使用不同的多项式函数拟合，三次的效果最佳。在同等错误率的条件下，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 $h_i$

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

通过多个分类器的组合，得到强分类器。每次训练一个分类器后，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，以此来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。

2018-1-19 22:14

6. $W(1)=(0,0,0,0)$，对 $\omega_2$ 的样本取相反数

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

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

更新规则：$W(K+1)=\begin{cases}W(K) & W(K)^TX(K)>0\\ W(K)+X(K+1) & \text{其它}\end{cases}$ （W不变即不更新的步骤）

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

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

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

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

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

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

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

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

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

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

$X(11)=(0,0,0,1)^T$，$W(11)^TX(11)=0$，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 $\omega$ 的情况下，$x$ 的各分量相互独立。

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

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

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

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

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

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

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

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

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

$\max \log p(\vec{\beta}|\vec{y})$ 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，$D_u \{(x_i,y_i)\}_{i=l+1}^{m}$ 未知

令 $z_{ik} = \mathbb{I}(x_i,y_i=k)$，$\forall i$ $l+1 \le i \le m$，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：~~ $\sum_{i=1}^{l}[\log p(x_i,y_i) + \sum_{i=l+1}^{m}\sum_{j=1}^{n} z_{ij}(\log p(x_i|y_i=j) + \log p(y_i=j))]$

~~$= \sum_{i=1}^{l}[\log p(x_i,y_i) + \sum_{i=l+1}^{m}\sum_{j=1}^{n} z_{ij}(\frac{1}{2}(x_i-\mu_j)^T\Sigma_j^{-1}(x_i-\mu_j) + \log p(y_i=j))]$ + constant~~

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

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

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

10.

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

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

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

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

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

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

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

$\therefore x \in w_2$.

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

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

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

$= 0.24$

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

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

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

$= 0.18$

$r_2 = 0.18 < 0.24 = r_1$

$\therefore x \in w_2$

2018-1-19 22:14

https://imlogm.github.io/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0/occam-razor-NFL/

2017-2018 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：091M4044H
课程名称：模式识别与机器学习
任课教师：黄庆明、兰艳艳、鲁鹏丰、山世光

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一并交回。

1. （8 分）试阐述线性判别函数的基本概念，并说明既然有线性判别函数，为什么还需要非线性判别函数？假设有两类模式，每类包括 6 个 4 维不同的模式，且良好分布。如果它们是线性可分的，问权向量至少需要几个系数分量？假如要建立二次的多项式判别函数，又至少需要几个系数分量？（设模式的良好分布不因模式变化而改变）

2. （8 分）简述 SVM 算法的原理。如果使用 SVM 做二分类问题得到如下结果，分别应该采取什么措施以取得更好的结果？并说明原因。
(1) 训练集的分类准确率 90%，验证集的分类准确率 90%，测试集的分类准确率 88%；
(2) 训练集的分类准确率 98%，验证集的分类准确率 90%，测试集的分类准确率 88%。

3. （8 分）请从两种角度解释主成分分析（PCA）的优化目标。

4. （8 分）请给出卷积神经网络 CNN 中卷积、Pooling、RELU 等基本层操作的含义。然后从提取特征的角度分析 CNN 与传统特征提取方法（例如 Gabor 小波滤波器）的异同。

5. （10 分）用线性判别函数的感知器赏罚训练算法求下列模式分类的解向量，并给出相应的判别函数。

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

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

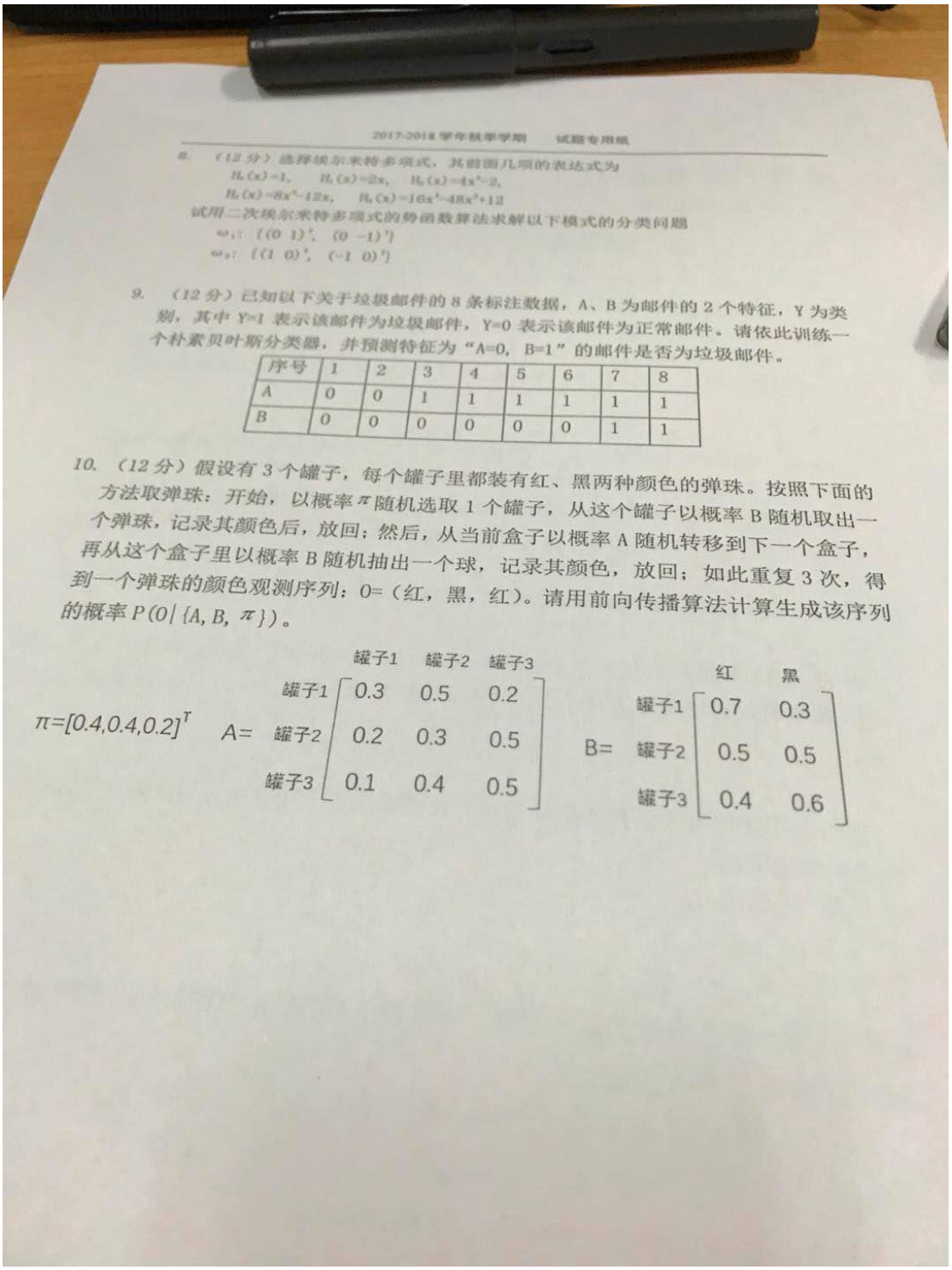
6. （10 分）试述 K-L 变换的基本原理，并将如下两类样本集的特征维数降到一维，时画出样本在该空间中的位置。

$\omega_1$: $\{(-5\ -5)^T, (-5\ -4)^T, (-4\ -5)^T, (-5\ -6)^T, (-6\ -5)^T\}$

$\omega_2$: $\{(5\ 5)^T, (5\ 6)^T, (6\ 5)^T, (5\ 4)^T, (4\ 5)^T\}$，

其中假设其先验概率相等，即 $P(\omega_1)=P(\omega_2)=0.5$。

. （12 分）请解释 AdaBoost 的基本思想和工作原理，写出 AdaBoost 算法



2017-2018 学年秋季学期　　试题专用纸

8. （12 分）选择埃尔米特多项式，其前面几项的表达式为

$H_0(x)=1$，　$H_1(x)=2x$，　$H_2(x)=4x^2-2$，

$H_3(x)=8x^3-12x$，　$H_4(x)=16x^4-48x^2+12$

试用二次埃尔米特多项式的势函数算法求解以下模式的分类问题

$\omega_1$：$\{(0\ 1)^T,\ (0\ -1)^T\}$

$\omega_2$：$\{(1\ 0)^T,\ (-1\ 0)^T\}$

9. （12 分）已知以下关于垃圾邮件的 8 条标注数据，A、B 为邮件的 2 个特征，Y 为类别，其中 Y=1 表示该邮件为垃圾邮件，Y=0 表示该邮件为正常邮件。请依此训练一个朴素贝叶斯分类器，并预测特征为“A=0，B=1”的邮件是否为垃圾邮件。

| 序号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| A | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| B | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

10. （12 分）假设有 3 个罐子，每个罐子里都装有红、黑两种颜色的弹珠。按照下面的方法取弹珠：开始，以概率 $\pi$ 随机选取 1 个罐子，从这个罐子以概率 B 随机取出一个弹珠，记录其颜色后，放回；然后，从当前盒子以概率 A 随机转移到下一个盒子，再从这个盒子里以概率 B 随机抽出一个球，记录其颜色，放回；如此重复 3 次，得到一个弹珠的颜色观测序列：$O=$（红，黑，红）。请用前向传播算法计算生成该序列的概率 $P(O|\{A, B, \pi\})$。

$\pi=[0.4,0.4,0.2]^T$

$$A=\begin{array}{c} \\ \text{罐子1} \\ \text{罐子2} \\ \text{罐子3} \end{array}\begin{array}{c} \begin{array}{ccc}\text{罐子1} & \text{罐子2} & \text{罐子3}\end{array} \\ \begin{bmatrix} 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \\ 0.1 & 0.4 & 0.5 \end{bmatrix} \end{array}$$

$$B=\begin{array}{c} \\ \text{罐子1} \\ \text{罐子2} \\ \text{罐子3} \end{array}\begin{array}{c} \begin{array}{cc}\text{红} & \text{黑}\end{array} \\ \begin{bmatrix} 0.7 & 0.3 \\ 0.5 & 0.5 \\ 0.4 & 0.6 \end{bmatrix} \end{array}$$

2. 根据结果来看，SVM 在进行二分类时采用的是软间隔，其中 1 和 2 的训练集分类准确度不同，是因为 C 值设置的问题，C 表示对分类错误的惩罚程度，C 越大分类器就越不会允许出现分类错误现象，此时对应 2，C 越小分类器就越不会在乎训练集上的分类错误，此时对应 1，所以应该采取对 1 来说增大 C 值，对 2 来说减小 C 值。
另外 C 值和间隔宽度有着互斥关系，C 越大导致间隔宽度表小。

# 中国科学院大学

试 题 专 用 纸

考试日期：2019 年 1 月 14 日

课程编号：091M4042H-1、091M4042H-2

课程名称：模式识别与机器学习

任课教师：黄庆明、常虹、郭嘉丰、山世光、李国荣、卿来云

姓名_____ 学号_____ 成绩_____

1. （10 分）简述 Fisher 线性判别方法的基本思路，写出准则函数及对应的解。

2. （12 分）假设某个地区细胞识别中正常（$w_1$）和异常（$w_2$）两类的先验概率分别为：正常状态：$P(w_1)=0.95$，异常状态 $P(w_2)=0.05$。现有一待识别的细胞，其观察值为 $x$，已知 $p(x|w_1)=0.2$，$p(x|w_2)=0.5$。同时已知风险损失函数为：$\begin{pmatrix}\lambda_{11} & \lambda_{12}\\ \lambda_{22} & \lambda_{21}\end{pmatrix}=\begin{pmatrix}0 & 1\\ 8 & 0\end{pmatrix}$

其中 $\lambda_{ij}$ 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。试对该待识别细胞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。

3. （10 分）SVM 可以借助核函数（kernel function）在特征空间（feature space）学习一个具有最大间隔的超平面。对于两类的分类问题，任意输入$x$的分类结果取决于下式：

$$< \hat{w}, \phi(x) > + \hat{w}_0 = f(x; \alpha, \hat{w}_0)$$

其中，$\hat{w}$和$\hat{w}_0$是分类超平面的参数，$\alpha = [\alpha_1, \ldots \alpha_{|SV|}]$表示支持向量（support vector）的系数，$SV$ 表示支持向量集合。使用径向基函数（radial basis function）定义核函数$K(\cdot,\cdot)$，即$K(x,x') = \exp(-\frac{D(x,x')}{2s^2})$。假设训练数据在特征空间线性可分，SVM 可以完全正确地划分这些训练数据。给定一个测试样本$x_{far}$，它距离所有训练样本都非常远。

试写出$f(x; \alpha, \hat{w}_0)$在核特征空间的表达形式，进而证明：$f(x_{far}; \alpha, \hat{w}_0) \approx \hat{w}_0$

4. （10 分）K-L 变换属于有监督学习（supervised learning）还是无监督学习（unsupervised learning）？试利用 K-L 变换将以下样本集的特征维数降到一维，同时画出样本在该空间的位置。

$$\{(-5\ -5)^T, (-5\ -4)^T, (-4\ -5)^T, (-5\ -6)^T, (-6\ -5)^T, (5\ 5)^T, (5\ 6)^T, (6\ 5)^T, (5\ 4)^T, (4\ 5)^T\}$$

5. （12 分）过拟合与欠拟合。

1) 什么是过拟合？什么是欠拟合？
2) 如何判断一个模型处在过拟合状态还是欠拟合状态？
3) 请给出 3 种减轻模型过拟合的方法。

6. （12 分）用逻辑回归模型（logistic regression model）解决$K$类分类问题，假设每个输入样本$x \in \mathbb{R}^d$的后验概率可以表示为：

$$P(Y=k|X=x)=\frac{\exp(w_k^T x)}{1+\sum_{l=1}^{K-1}\exp(w_l^T x)} \quad , k=1,\dots,K-1 \qquad (1)$$

$$P(Y=K|X=x)=\frac{1}{1+\sum_{l=1}^{K-1}\exp(w_l^T x)} \qquad (2)$$

其中$w_k^T$表示向量$w_k$的转置。通过引入$w_K=\vec{0}$，上式也可以合并为一个表达式。

1) 该模型的参数是什么？数量有多少？

2) 给定 $n$ 个训练样本$\{(x_1,y_1),(x_2,y_2),\dots,(x_n,y_n)\}$，请写出对数似然函数（log likelihood function）$L$的表达形式，并尽量化简。

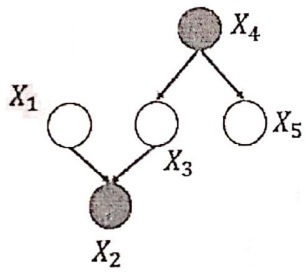
$$L(w_1,\dots,w_{K-1})=\sum_{i=1}^{n}\ln P(Y=y_i|X=x_i)$$

3) 如果加入正则化项（regularization term），定义新的目标函数为：

$$J(w_1,\dots,w_{K-1})=L(w_1,\dots,w_{K-1})-\frac{\lambda}{2}\sum_{l=1}^{K}\|w_l\|_2^2$$

请计算 $J$ 相对于每个$w_k$的梯度。

7. （10 分）给定如下概率图模型，其中变量 $X_2$、$X_4$ 为已观测变量，请问变量 $X_1$ 和 $X_5$ 是否独立？并用概率推导证明之。




8. （12 分）假设有 2 枚硬币，分别记为 A 和 B，以π的概率选择 A，以 1-$\pi$的概率选择 B，这些硬币正面出现的概率分别是$p$和$q$。掷选出的硬币，记正面出现为 1，反面出现为 0，独立地重复进行 4 次试验，观测结果如下：1，1，0，1。给定模型参数$\pi=0.4, p=0.6, q=0.5$，请计算生成该序列的概率，并给出该观测结果的最优状态序列。

9. （12 分）基于 AdaBoost 的目标检测需要稠密的扫描窗口并判断每个窗口是否为目标，请描述基于深度学习的目标检测方法，如 SSD 或 YOLO，如何做到不需要稠密扫描窗口而能发现并定位目标位置？

# 中国科学院大学

考试日期：2020.01.11

试 题 专 用 纸

课程编号：721004H

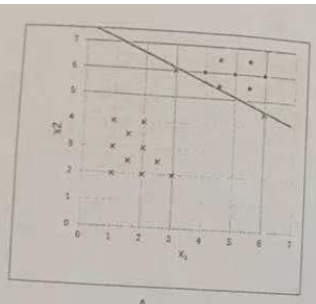
课程名称：模式识别与机器学习

任课教师：黄庆明等

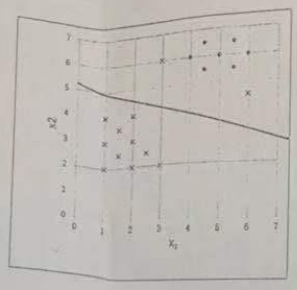
姓名_____ 学号_____ 成绩_____

一、（16 分）选择题。（每个选项 2 分，请将答案写在答题纸上）

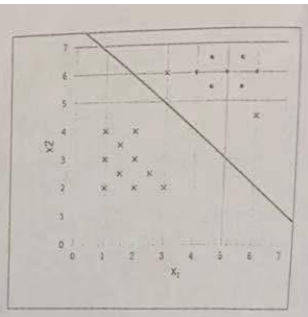
1. 基于二次准则函数的 H-K 算法较之于感知器算法的优点是哪个？
   A. 计算量小
   B. 可以判别问题是否线性可分
   C. 其解完全适用于非线性可分的情况
2. 在逻辑回归中，如果正则项取 L1 正则，会产生什么效果？
   A. 可以做特征选择，一定程度上防止过拟合
   B. 能加快计算速度
   C. 在训练数据上获得更准确的结果
3. 如果模型的偏差较高，我们如何降低偏差？
   A. 在特征空间中减少特征
   B. 在特征空间中增加特征
   C. 增加数据点
4. 假设采用正态分布模式的贝叶斯分类器完成一个两类分类任务，则下列说法正确的是哪个。
   A. 假设两类的协方差矩阵均为对角矩阵，则判别界面为超平面。
   B. 假设两类的协方差矩阵相等，则判别界面为超平面。
   C. 不管两类的协方差矩阵为何种形式，判别界面均为超平面。
5. 下列方法中，哪种方法不能用于选择 PCA 降维（K-L 变换）中主成分的数目$K$？
   A. 训练集上残差平方和随$K$发生剧烈变化的地方（肘部法）
   B. 通过监督学习中验证集上的性能选择$K$
   C. 训练集上残差平方和最小的 $K$
6. 考虑某个具体问题时，你可能只有少量数据来解决这个问题。不过幸运的是你有一个针对类似问题已经预先训练好的神经网络，请问可以用下面哪种方法来利用这个预先训练好的网络？
   A. 把除了最后一层外所有的层都冻住，重新训练最后一层
   B. 对新数据重新训练整个模型
   C. 只对最后几层进行调参(fine tune)
7. 如下图所示，假设该数据集中包含一些线性可分的数据点。训练 Soft margin SVM 分类器，其松弛项的系数为$C$。请问当$C \to 0$时，分类边界为下图中的哪个？



A

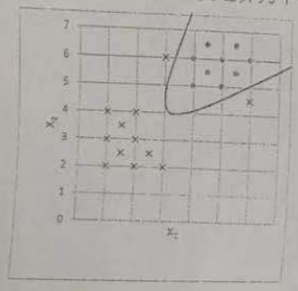


B

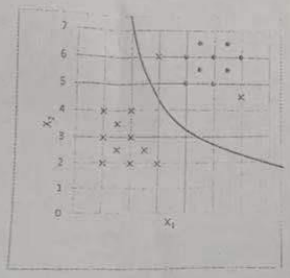


C

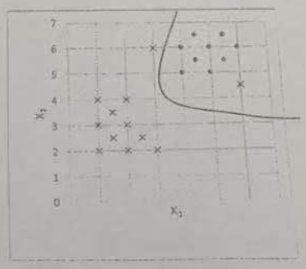
8. 如下图所示，假设该数据集中包含线性不可分的数据点。采用二次核函数训练 Soft margin SVM 分类器，请问当$C \to \infty$时，分类边界为下图中的哪个？



A



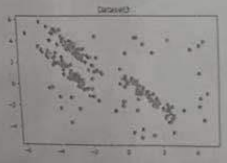
B



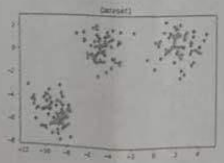
C

二、（6 分）请列举半监督学习对数据样本的三种基本假设。

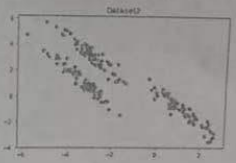
三、（8 分）针对下图所示的三种数据分布，从 K 均值、GMM 和 DBSCAN 中分别选择最合适的聚类算法，并简述理由。



(a)



(b)



(c)

四、（12 分）对于具有类别标签的数据，采用 K-L 变换和 Fisher 线性判别分析两种方法对数据降维。

（1）简述这两种数据降维方法的基本过程。（8 分）

（2）这两种方法中哪种方法对分类更有效？并简述原因。（4 分）

五、（10 分）逻辑回归

（1）简述逻辑回归算法的原理。（4 分）

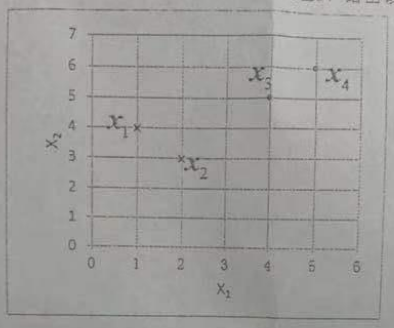
（2）如果使用逻辑回归算法做二分类问题得到如下结果，分别应该采取什么措施以取得更好的结果？并说明理由。（6 分）

（a）训练集的分类准确率 85%，验证集的分类准确率 80%，测试集的分类准确率 75%；
（b）训练集的分类准确率 99%，验证集的分类准确率 80%，测试集的分类准确率 78%；

六、（10 分）解释 AdaBoost 算法的基本思想和工作原理，并给出 AdaBoost 算法的伪代码。

七、（10 分）从特征提取的角度，分析深度卷积神经网络与传统特征提取方法（例如 Gabor 小波滤波器）的异同，并给出深度学习优于传统方法的原因。

八、（8 分）硬间隔支持向量机（Hard margin SVM）
如下图所示，一个数据集包含来自 2 个类别的 4 个数据点，在此集合上训练一个线性 Hard margin SVM 分类器。请写出 SVM 的形式化模型，并计算出该分类器的权重向量$w$和偏差$b$，给出该分类器的支持向量。




九、（10 分）拟利用贝叶斯判别方法检测 SNS 社区中不真实账号。设$Y=0$表示真实账号，$Y=1$表示不真实账号。每个用户有三个属性：$X_1$表示日志数量/注册天数，$X_2$表示好友数量/注册天数，$X_3$表示是否使用真实头像。已知$P(Y=0)=0.89$，$P(X_3=0|Y=0)=0.2$，$P(X_3=0|Y=1)=0.9$，且给定$Y$的情况下$X_1$、$X_2$的分布如下：

| $P(X_1\|Y)$ | $X_1 \le 0.05$ | $0.05 < X_1 \le 0.2$ | $X_1 \ge 0.2$ |
|---|---|---|---|
| $Y=1$ | 0.8 | 0.1 | 0.1 |
| $Y=0$ | 0.3 | 0.5 | 0.2 |
| $P(X_2\|Y)$ | $X_2 \le 0.1$ | $0.1 < X_2 \le 0.8$ | $X_2 \ge 0.8$ |
| $Y=1$ | 0.7 | 0.2 | 0.1 |
| $Y=0$ | 0.1 | 0.7 | 0.2 |

若一个账号使用非真实头像，日志数量与注册天数的比率为 0.1，好友数与注册天数的比率为 0.2，判断该账号是不是虚假账号。

十、（10 分）现装有红色球和白色球的两个盒子，盒子 1 中红球的比例为$p$，盒子 2 中红球的比例为$q$。我们以概率π选择盒子 1,概率$1-\pi$选择盒子 2，然后从盒子中有放回地取出一个小球，独立地重复进行 4 次试验，观测结果为：红，红，白，红。

假定模型的参数初始值为$\pi^{(0)}=0.4$，$p^{(0)}=0.4$，$q^{(0)}=0.5$，请写出 EM 算法迭代一次后$p$和$q$的值。（计算结果保留两位小数）

# 中国科学院研究生院

课程编号： 712008Z

试 题 专 用 纸

课程名称：机器学习

任课教师：卿来云

---

姓名_____ 学号_____ 成绩_____

1. 判断题（ 20 分，每小题 2 分）

（1）给定 n 个数据点，如果其中一半用于训练，另一半用于测试，则训练误差和测试误差之间的差别会随着 n 的增加而减小。 （T）

（2）当训练数据较少时更容易发生过拟合。 （T）

（3）回归函数 A 和 B，如果 A 比 B 更简单，则 A 几乎一定会比 B 在测试集上表现更好。（F）

（4）在核回归中，最影响回归的过拟合性和欠拟合之间平衡的参数为核函数的宽度。 （T）

（5）在 AdaBoost算法中，所有被错分的样本的权重更新比例相同。 （T）

（6）Boosting 的一个优点是不会过拟合。 （F）

（7）梯度下降有时会陷于局部极小值，但 EM 算法不会。（F）

（8）SVM 对噪声（如来自其他分布的噪声样本）鲁棒。 （F）

（9）Boosting 和 Bagging 都是组合多个分类器投票的方法， 二者都是根据单个分类器的正确率决定其权重。 （F）

（10）在回归分析中，最佳子集选择可以做特征选择，当特征数目较多时计算量大；岭回归和 Lasso 模型计算量小，且 Lasso 也可以实现特征选择。 （T）

2、 logistic 回归模型。（20 分，每小题 10 分）

我们对如图 1(a)所示的数据采用简化的线性 logistic回归模型 进行两类分类，即

$$P(Y=1|x,w_1,w_2)=g(w_1x_1+w_2x_2)=\frac{1}{1+\exp(-w_1x_1-w_2x_2)}。$$

（为了简化，我们不采用偏差 $w_0$ 。）

训练数据可以被完全分开（训练误差为 0，如图 1(b)所示的 $L_1$）。

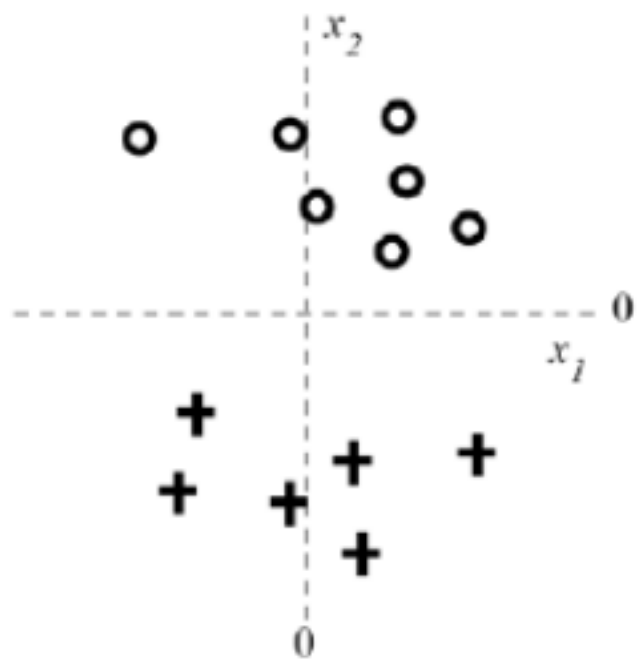



图 1(a) 2 维训练数据。

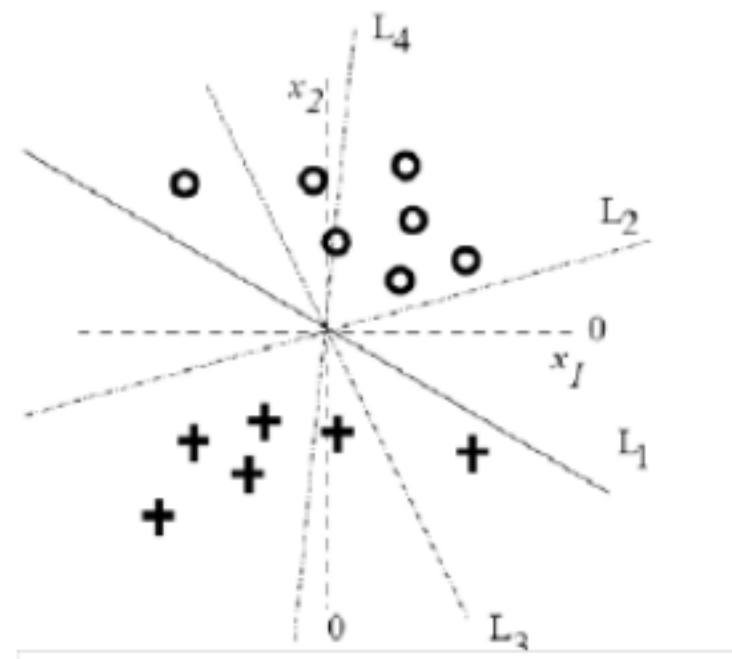



图1(b) 数据点可以被 $L_1$（实线）。$L_2$、$L_3$和$L_4$是另外几个可能的决策

（1） 考虑一个正则化的方法，即最大化

$$\sum_{i=1}^{N} \log P\left(y_i \mid x_i, w_1, w_2\right) - \frac{C}{2} w_2^2 。$$

注意只有 $w_2$ 被惩罚。则当 C 很大时，如图 1(b)所示的 4 个决策边界中，哪条线可能是有该正则方法得到的？

$L_2$、$L_3$ 和 $L_4$ 可以通过正则 $w_2$ 得到吗？

答： L2 不可以。当正则 w2 时，决策边界对 x2 的依赖越少，因此决策边界变得更垂直。而图中的 L2 看起来不正则的结果更水平，因此不可能为惩罚 w2 得到；

L3 可以。 w2^2 相对 w1^2 更小（表现为斜率更大），虽然该决策对训练数据的 log 概率变小（有被错分的样本）；

L4 不可以。当 C 足够大时，我们会得到完成垂直的决策边界（线 x1 = 0 或 x2 轴）。L4 跑到了 x2 轴的另一边使得其结果比其对边的结果更差。当中等程度的正则时，我们会得到最佳结果（ w2 较小）。 图中的 L4 不是最佳结果因此不可能为惩罚 w2 得到；

（2） 如果我们将正则项给出 L1 范式，即最大化

$$\sum_{i=1}^{N} \log P\left(y_i \mid x_i, w_1, w_2\right) - \frac{C}{2}\left(\left|w_1\right| + \left|w_2\right|\right)。$$

则随着 C 增大，下面哪种情形可能出现（单选）？

(a) $w_1$ 将变成 0，然后 $w_2$ 也将变成 0。(T)

(b) $w_1$ 和 $w_2$ 将同时变成 0。

(c) $w_2$ 将变成 0，然后 $w_1$ 也将变成 0。

(d) 两个权重都不会变成 0，只是随着 C 的增大而减小 0。

该数据可以被完全正确分类（训练误差为 0），且仅看 x2的值（w1 = 0）就可以得到。虽然最佳分类器 w1可能非 0，但随着正则量增大 w1会很快接近 0。L1 正则会使得 w1完全为 0。随着 C的增大，最终 w2 会变成 0。

3、产生式模型和判别式模型。 （16 分，每小题 8 分）

考虑两个分类器： 1) 核函数取二次多项式的 SVM 分类器 和 2) 没有约束的高斯混合模型（每个类别为一个高斯模型）。我们对 $R^2$ 空间上的点进行两类分类。假设数据完全可分， SVM 分类器中不加松弛惩罚项，并且假设有足够多的训练数据来训练高斯模型的协方差。

（1） 这两个分类器的 VC 维相同。（判断正误，并给出简短理由） （T）

因此两个分类器的决策边界都为二次函数，复杂度相同。

（2） 假设我们估计两个分类器的结构风险值， 该值为预测误差的上界。 则这连个分类器中哪个的结构风险值更小一些？给出简短理由。

SVM 可能会得到更好的结果。虽然两个分类器的复杂度相同，但 SVM对训练误差做优化从而得到更低（或相同）的

值。

4、 SVM 。（16 分，每小题 8 分）

我们采用两个 SVM 分类器对 $R^2$ 空间上的点进行两类分类，这两个分类器的不同在于核函数不同。其中分类器 1 采用的核函数为 $K_1(x,x') = x^T x'$，分类器 2 采用的核函数为 $K_2(x,x') = p(x)p(x')$，其中 $p(x)$ 为根据其他方法估计得到的概率密度函数。

（1） 采用核函数 $K_2$ 的分类器 2 的 VC 维是多少？

特征空间为 1 维（将任意点 x 映射成非负数 p(x)），因此 VC 维是 2。.

（2） 如果两个分类器都嫩对 N 个训练数据得到 0 训练误差，则哪个分类器会有较好的推广性能？给出简短理由。

分类器 1 的 VC 维为 3，而分类器 2 的 VC 维为 2，因此分类器 1 更复杂。当训练误差相同时，分类器 2 得到的预测误差的界更小，从而其推广性更好。



5、 Boosting 。（28分，每小题 7分）

考虑如下图 2所示的训练样本，其中 ’X和 ’O分别表示正样本和负样本。我们采用 AdaBoost 算法对上述样本进行分类。在 Boosting 的每次迭代中，我们选择加权错误率最小的弱分类器。假设采用的弱分类器为平行两个坐标轴的线性分类器。

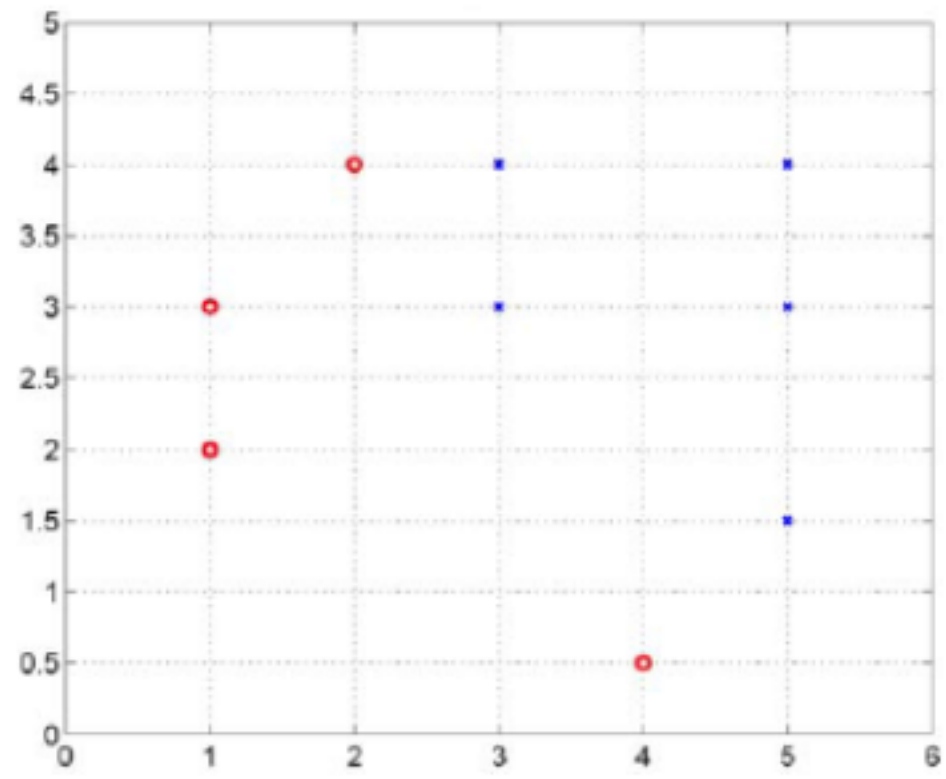



图 2：训练样本

（1） 在图 2 中标出第一次迭代选择的弱分类器（ $L_1$），并给出决策面的‘ + ’和‘ - ’面。

（2） 在图 2 中用圆圈标出在第一次迭代后权重最大的样本，其权重是多少？

（3） 第一次迭代后权重最大的样本在经过第二次迭代后权重变为多少？

（4） 强分类器为弱分类器的加权组合。 则在这些点中， 存在被经过第二次迭代后的强分类器错分的样本吗？给出简短理由。