第一章 概述

一、填空题

1 将模式识别任务看做 1 个输入输出系统 $y=f(x)$，则对于分类任务，y 通常为（离散量）类别的标签；对于回归任务，y 通常为（~~连续量~~）。待预测变量

2 利用一个线性分类器对 iris 数据库进行分类，其样本空间是（花的特征）的集合，维度是（4）；其假设空间是（　　　　）的集合，其参数空间维度是（4）；其类别标签集合是（　　　　）。

二、判断题

3 一个数据集的特征维度总是小于该数据集的样本数量。（√）特征≤ 特征数小于样本

4 模型容量大小与训练集样本数量无关。（X）√ 模型容量与模型的参数有关 表现为假设空间大小

5 明日降水概率预测属于回归问题。（√）

6 班级内自由分组讨论不属于分类问题。（√）无类

7 驯兽师用鞭打和喂食的方式教导老虎跳火圈属于有监督学习。（√）X 强化学习

三、选择题

8 处理二维数据的单个神经元模型的参数个数是：（A）

A. 3个　　B. 2个　　C. 1个　　D. 不确定

9 从当前考场的所有考生中，找出作弊的学生，该任务属于：（C）AC

A. 分类　　B. 聚类　聚类是指根据数据内部自行归纳　C. 异常检测　　D. 以上均是

10 以下方法中不属于主流机器学习方法的是：（C）

A. 有监督学习　　B. 无监督学习　　C. 随机学习　　D. 强化学习

四、简答题

11 什么是分类任务？分类任务有哪些实际的应用？ 根据数据判断一个对象的类别 人脸识别…

12 什么是回归任务？回归任务有哪些实际的应用？ 根据观测数据预测某个对象的某个连续属性的数值 天气预报

13 分类任务和回归任务的区别是什么？

14 什么是聚类任务？请列举几个聚类的实际应用。互联网客户分析 将数据观测对象分为由相似对象组成的多个类别

15 什么是异常检测任务？异常检测的实际应用是什么？ 找出不属于某个类别的数据对象 账号异常登录

16 人脸图像伪造属于分类任务还是回归任务，为什么？ 回归，因为人脸伪造算法生成的图像可以看做一个高维连续性变量 无明确类别信息

17 Iris 数据集有哪几个类别？每个样本都包含那些特征？ 3个类别 5个

18 模式识别方法可以分为哪三种思路？请分别举例说明！

基于知识的方法 → 动力学…兵乓球轨迹
基于经验的方法 → 专家系统
基于学习的方法 → 图像大数据人脸识别

19. 有监督学习 { 定义：利用训练数据中给出的"标准"答案学习
特点：
无监督学习 { 定义：无标准答案，从数据中寻找答案.
特点：
强化学习 { 定义：无标准答案，但有奖惩信号，从探索中寻找答案
特点：

~~19~~ 现代机器学习的三种主流方法是什么，各自的定义和特点是什么？

~~20~~ 如果一个模型在训练集上表现良好，但在测试集上表现不佳，试说明该模型的容量与问题复杂度之间的关系。

该模型的容量可能明显超出了问题的复杂度，导致了过拟合现象.

五、计算(画图)题

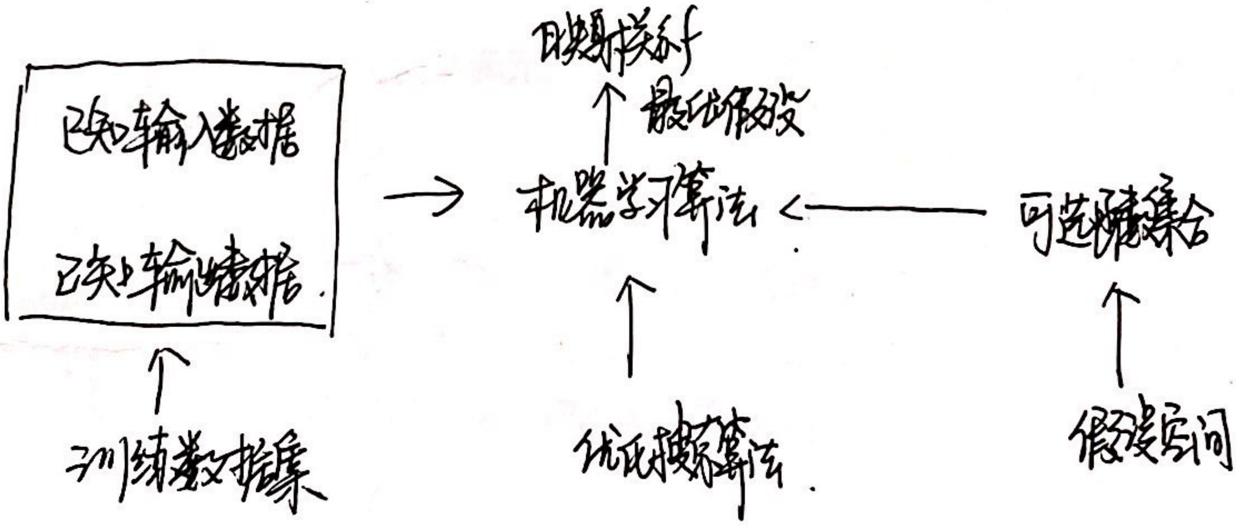
~~21~~ 下面是一个房屋价格预测数据集的部分数据，该任务是分类任务还是回归任务？是有监督还是无监督？试写出每个样本的特征向量？特征空间的维度是多少？ 回归，有监督. [占地面积，距市中心距离，地下室面积].

维度是3.

编号	占地面积/m²	距市中心距离/km	地下室面积/ m²	房屋价格/万元
1	100	12	3	81.5
2	120	5	6	126
3	90	18	5	76.2
4	150	3	10	165.3
5	200	20	0	161
6	160	9	8	152.6

一号样本特征向量为 [100, 12, 3]

22 请画出有监督机器学习系统的学习阶段的系统框图



第二章 贝叶斯决策论

一、填空题

后验

1. 已知样本x的具体数值，其从属于某个类别的概率是（ $P(\omega_i|x)$ ）；已知样本x属于某个类别，则该 x 取某个具体数值的概率称为（ $P(x|\omega_i)$ ）。⇒ 类条件
2. 利用一个线性分类器对 iris 数据库进行分类，其样本空间是（　　　　）的集合，维度是（　　　）；其假设空间是（　　　　　）的集合，其参数空间维度是（　　　）；其类别标签集合是（　　　　　）。

3. 面向n维数据的线性分类器决策面是n维空间中的一个(n-1)维的超平面

二、判断题

是关于离散类别状态ω的概率

~~3.~~ 在分类问题中，类条件概率是概率质量，后验概率是概率密度。（ ✗ ）

密度　质量

~~4.~~ 有些问题中，无需了解先验概率也可以直接估计后验概率。（ √ ）

是关于连续变量x的概率　判别问题　边缘分布

5. 最大后验概率分类准则和最小错误概率分类准则在任何情况下给出的分类结果都是一样的。（ ✗ ）
6. 最小错误概率分类准则给出的分类结果总是正确的。（ ）√

三、选择题

$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$ 协方差矩阵

$P(x|\omega_i)$ 密度 $= \frac{P(x|\omega_i)}{P(\omega_i)}$

~~7.~~ 假设样本$x \in \Re^3$，服从多元正态分布$\mathcal{N}(\mu, \Sigma)$，则类条件概率密度函数的有效参数数量为：（ B ）

$x \sim N(\mu, \Sigma)$

$\mu \in R^3$ (3个有效参数)　$\Sigma \in R^{3\times 3}$ (3+3)

A. 12 个　　B. 9 个　　C. 6 个　　D. 不确定　对称矩阵

8. 对于一维样本二分类问题，两类ω_1, ω_2均服从正态分布，假设按照最小平均错误概率准则设计的分类器为 C1，分类阈值为 T1，按照最小平均风险准则设计的分类器为 C2，T2，如果风险系数$\lambda_{12} > \lambda_{21}$，$\lambda_{11} = \lambda_{22}$则以下论断正确的是：（ ~~D~~ ）. A

$=1-P(\omega_i|x)$

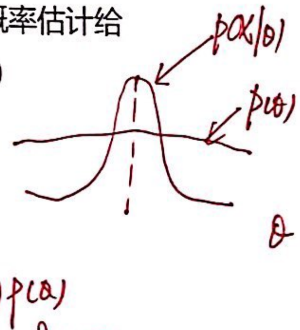
A. T1 比 T2 更靠近ω_1类的均值　　B. T2 比 T1 更靠近ω_1类的均值

C. T1 与 T2 相同　　D. 无法判断

$\Lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix}$ λ_{ij}表示将ω_i类分给ω_j类的概率

四、简答题

~~9.~~ 如果对于某个数据集的概率分布模型，最大似然估计和最大后验概率估计给出结果完全相同，试解释可能造成这一结果的原因（可画图说明）



10. 写出 K 个高斯成分的 GMM 模型的对数似然函数的数学形式

9. 最大似然估计的结果是似然函数的最大值解 $\arg\max_\theta p(x;\theta)$
最大后验估计…… 后验概率的最大值解 $\arg\max_\theta p(x;\theta)p(\theta)$
如果相同，说明$p(x;\theta)$与参数θ的先验概率分布可能存在相同的最大解。

$$\hat{\mu}_j = \frac{\sum_{i=1}^{N} y_{ji}^{t} x_i}{\sum_{i=1}^{N} y_{ji}^{t}} \Rightarrow \mu_j^{t+1} \quad j=1,2,\cdots K$$

$$\hat{\Sigma}_j = \frac{\sum_{i=1}^{N} y_{ji}^{t}(x_i-\mu_j)(x_i-\mu_j)^T}{\sum_{i=1}^{N} y_{ji}^{t}} \Rightarrow \Sigma_j^{t+1}.$$

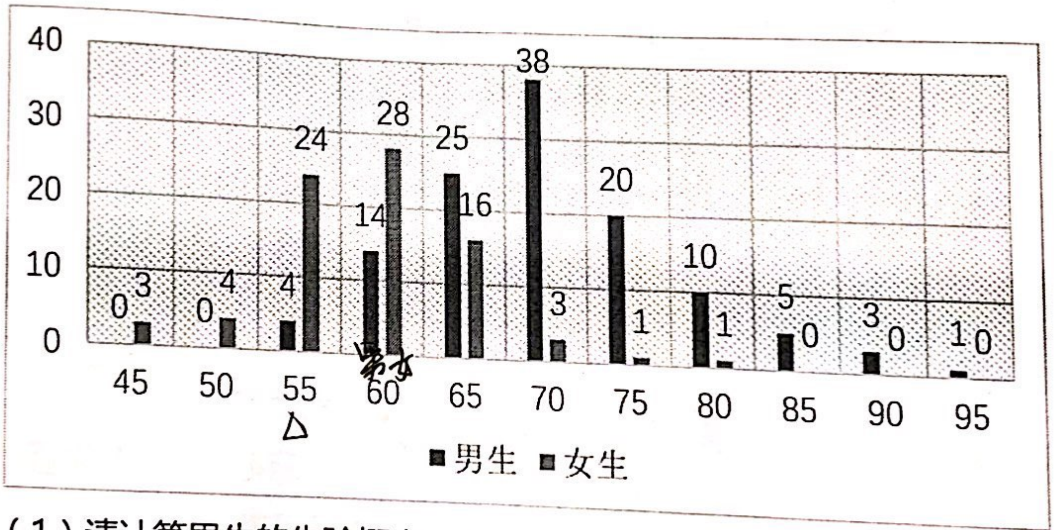
11. 请写出 EM 算法的 M 步骤中，各高斯成分的均值向量μ_j、协方差矩阵Σ_j和比例系数P_j在 t 时刻的迭代更新公式。

$$\hat{P}_j = \frac{\sum_{i=1}^{N} y_{ji}^{t}}{N} \Rightarrow P_j^{t+1}.$$

五、计算（画图）题

12. 假设一个学校里有 60%的男生和 40%的女生。女生点外卖的人数和吃食堂的人数相等，所有男生都吃外卖。一个人在校门口看到了一个拿着外卖的学生，那么这个学生时女生的概率是多少？

13. 一个班级中共有学生 200 人，其中男生 120 人，女生 80 人，他们的体重分布如下表所示：



（1）请计算男生的先验概率、女生体重处于 55 公斤区间段的类条件概率、男学生体重为 55 公斤的类条件概率密度函数估计值、任意学生体重处于 55 公斤区间段的概率、某个体重为 55 公斤的学生属于女生类别的后验概率，并写出计算过程。

（2）请根据最大后验概率预测某个体重为 55 公斤的学生的性别，并写出计算过程。

14. 数据如上题：设某学生体重为 55 公斤，问题：

（1）将该生预测为男生的错误概率为多少？

（2）将该生预测为女生的错误概率为多少？

（3）根据最小错误概率分类准则，预测结果为什么？

$$\Lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix} = \begin{bmatrix} 0 & 2 \\ 1 & 0 \end{bmatrix}$$

（1）预测该生为男生的风险为多少？

（2）预测该生为女生的风险为多少？

（3）根据最小风险分类准则，预测结果为什么？

16. 已知两个类别ω_1和ω_2，其先验概率相等，两类的类条件概率服从正态分布，有$p(x|w_1) = \mathcal{N}\left(\mu_1, \Sigma\right)$和$p(x|w_2) = \mathcal{N}\left(\mu_2, \Sigma\right)$，且$\mu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$，$\mu_2 = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$，$\Sigma = \begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix}$，试采用贝叶斯决策对样本$x = \begin{bmatrix} 1.0 \\ 2.2 \end{bmatrix}$进行分类。

17. 两个学生玩猜硬币的游戏；共猜了 30 次，其中正面 17 次，反面 13 次。假设一枚硬币扔出正面的概率为θ，单次扔硬币的结果服从伯努利分布：

$$P(x;\theta) = \theta^x(1-\theta)^{1-x}, x = 0\ or\ 1$$

其中，$x = 1$表示扔出正面，$x = 0$表示扔出反面；(1). $L_{ML}(X;\theta) = p(X;\theta)$

（1）根据题意，写出似然性函数的数学表达式。

（2）写出概率模型最大似然估计的最优化问题标准形式。 $= \prod_{i=1}^{N} p(x_i;\theta)$

（3）采用最大似然估计求取参数θ的值？

(2). $\hat{\theta}_{ML} = \arg\max_{\theta} L_{ML}(X;\theta)$　(3). $\hat{\theta}_{M} = \arg\max_{\theta} L_{ML}(X;\theta)$

18. 问题描述如上题所示

（1）假设θ的分布服从$\mathcal{N}(0.4, 0.2^2)$，写出最大后验概率估计？ $= \theta^{17}(1-\theta)^{13} = \frac{17}{4}$

（2）根据最大后验概率估计推导出θ的值？

$\hat{\theta} = \arg\max_{\theta} \left(\frac{p(x|\theta)p(\theta)}{p(x_0)} \right) =$

19. 已知 iris 数据中的 setosa 类的 10 个样本的部分特征数据如下，请依据下述数据估计 setosa 类的特征均值向量与协方差矩阵，写出计算过程。 $(x-\bar{x})^2$

序号	花萼长度/cm	花萼宽度/cm	类别
1	5.1	3.5	setosa
2	4.9	3	setosa
3	4.7	3.2	setosa
4	4.6	3.1	setosa
5	5	3.6	setosa
6	5.4	3.9	setosa
7	4.6	3.4	setosa
8	5	3.4	setosa
9	4.4	2.9	setosa

$u_1 = (5.1+4.9+4.7+4.6+5+5.4+4.6+5+4.4+4.9)\times\frac{1}{10} = 4.86$

$u_2 = (3.5+3+\cdots\cdots+3.1)\times\frac{1}{10} = 3.29$　$u = [4.86, 3.29]$

$\Sigma_i = E[(x-u_1)(x-u_2)^T] = \frac{1}{10}\times\left[([5.1\ 3.5]-[4.86\ 3.29])^T([5.1\ 3.5]-[4.86\ 3.29])+\cdots\right]$

10	4.9	3.1	setosa

20. 数据如上题所述，分别采用直方图法(bin 边长为 0.4cm)，KDE 法($\delta^2 = 1$) 和 KNN 法 (K=3) 估计花萼长度=4.8cm，花萼宽度=3.2cm 处的概率密度函数估计值。

21. 数据如上题所述，利用朴素贝叶斯分类器+直方图法估计花萼长度=4.8cm，花萼宽度=3.2cm 的样本的类条件概率密度函数值。

~~19.~~

20. $h=0.4cm$

$d=2$

$\hat{p}(x_1) = \frac{1}{(0.4)^2} \times \frac{8}{10} = 5$

w_2=女生

$w_1)=0.6$

$w_2)=0.4$.

$$w_2|x=斜棱) = \frac{P(x=斜棱|w_2)\,P(w_2)}{P(x=斜棱)} = \frac{\frac{1}{2}\times 0.4}{\cancel{0.2\times}\ 0.4\times\frac{1}{2}+0.6} = \frac{0.2}{0.2+0.6} = \frac{1}{4}$$

w_1=男生.

w_2=女生.

$$w_1) = \frac{120}{200} = \frac{3}{5}$$

$$=55|w_2) = \frac{24}{80} = \frac{3}{10}$$

$$=55|w_1) = \frac{4}{120} = \frac{1}{30}$$

$$x=55) = \frac{28}{200} = \frac{7}{50}$$

$$w_2|x=55) = \frac{P(x=55|w_2)\,P(w_2)}{P(x=55)} = \frac{\frac{3}{10}\times\frac{2}{5}}{\frac{7}{50}} = \frac{\frac{6}{50}}{\frac{7}{50}} = \frac{6}{7}$$

$$2|x=55) = \frac{6}{7}$$

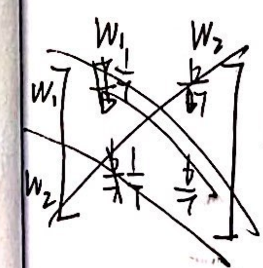
$$1|x=55) = \frac{P(x=55|w_1)\,P(w_1)}{P(x=55)} = \frac{\frac{1}{30}\times\frac{3}{5}}{\frac{7}{50}} = \frac{1}{7}$$

$2|x=55) > P(w_1|x=55)$ ∴女生.

$$P_e(w_2|x=55) = 1-P(w_2|x=55) = \cancel{\tfrac{6}{7}}\ \frac{1}{7}$$

$$P_e(w_1|x=55) = 1-P(w_1|x=55) = \cancel{\tfrac{1}{7}}\ \frac{6}{7}$$

是女生.



$$\begin{array}{ccc} & a_{11}\ \lambda_{12} & \\ \lambda = & \begin{bmatrix} 0 & 2 \\ 1 & 0 \end{bmatrix} & \begin{matrix}\lambda_{21}\\ \lambda_{22}\end{matrix} \end{array}$$

$$\cancel{P(w_1|x) = P(w_1)\ \frac{1}{7}\times 0+\frac{1}{7}\times 2 = \frac{2}{7}}$$

$$\cancel{P(w_2|x) = 2\times\frac{6}{7}+0 = \frac{12}{7}}$$

男生.

$$l_1 = a_{11}P(w_1|x)+\lambda_{21}P(w_2|x) = 0\times\frac{1}{7}+2\times\frac{6}{7} = \frac{12}{7}$$

$$l_2 = \lambda_{12}P(w_1|x)+\lambda_{22}P(w_2|x) = 1\times\frac{1}{7} = \frac{1}{7}$$

解：

$$w|x) = \frac{P(x|w)P(w)}{P(x)}$$. 贝叶斯 . 先验 $P(w)$.

$$1|x) = \frac{P(x|w)\cdot\frac{1}{2}}{\frac{1}{2}} = \left[\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_i|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\right]\right]$$ 多元正态分布.

$$= \frac{1}{(2\pi)\sqrt{1.1\times1.9-0.09}} \exp\left[-\frac{1}{2}\left(x\begin{bmatrix}1.0\\2.2\end{bmatrix}-\begin{bmatrix}0\\0\end{bmatrix}\right)^T\begin{bmatrix}1.1 & 0.3\\0.3 & 1.9\end{bmatrix}^{-1}\left(\begin{bmatrix}1\\2.2\end{bmatrix}\right)\right]$$

$$= \frac{1}{4\pi}\cdot\exp(\quad) = e^{-1.476} = 0.159$$

$$2|x) = 0.229$$

$$(w=2|x) > P(w=1|x).$$

第三章 线性模型

一、填空题

1. 生成模型通过估计（类条件概率）进行分类，判别模型通过估计（后验概率）进行分类。
2. 感知器算法通常采用（梯度下降）法求解目标函数的优化问题。
3. 线性回归模型的封闭解又称为（~~闭式~~）解。最小二乘（因为使用了均方误差目标函数）
4. 逻辑回归函数的取值区间是（[0，1]）。(0，1)

二、判断题

5. 当两个类别均服从正态分布时，根据贝叶斯决策理论计算出的决策面必然是一个线性决策面。（√）
6. 当两个类别均服从正态分布时，根据贝叶斯决策理论计算出的决策面必然是一个二次型决策面。（ ）
7. 线性回归模型如存在唯一解，必然可以令模型在训练集上的均方误差为 0.（√X）取唯一解仅表示均方误差为最小值
8. 4 个不具有共线性的三维样本，无法用线性回归模型得到唯一解。（√）N维
9. 逻辑回归模型无法用于多类分类问题。（√）X 将多类分类转化为多个二分类

三、选择题

10. 假设线性回归问题的训练集为：$x_i \in \Re^3, i = 1, \ldots, 100$，则该线性模型包含多少个参数：（B）3维100个

A. 3 个　　B. 4 个　　C. 100 个　　D. 101 个

$y = W^T x_i + b = [w_1\ w_2\ w_3]^T [x_{i1}\ x_{i2}\ x_{i3}]^T + b$

11. 上题中如果使用线性回归的最小二乘解的封闭解形式，则自相关矩阵 R_x 的大小为：（B）

A. 3×3　　B. 4×4　　C. 100×100　　D. 101×101

矩阵
$\hat{a} = X^T y$
$(X^T X)^{-1} X^T y$
$R^{(d+1)\times(d+1)}$

12. 逻辑回归模型中的逻辑回归函数可以看做是对以下哪种概率的描述：（B）

A. $p(x|\omega_1)$　　B. $p(\omega_1|x)$　　C. $p(x)$　　D. $p(\omega_1)$

后验
是逻辑分析对后验概率的描述

四、简答题

13. 请写出感知器算法的目标函数的标准数学形式，并解释其中每一个符号的意义与计算方法，说明其合理性。

$\delta_x = \begin{cases} -1 & x \in w_1 \\ 1 & x \in w_2 \end{cases}$

$\hat{w}^* = \arg\min_{\hat{w}} J(\hat{w}) = \cancel{\arg\min \delta_x \hat{w}^T \hat{x}}$

$\arg\min_{\hat{w}} \sum_{x \in \Omega(\hat{w})} \delta_x \hat{w}^T \hat{x}$

$\hat{x}$ 错分样本
$\hat{w}$ 参数模型

$(\hat{\omega})=E[(y-f(x))^2]$

$=E[(y-\hat{\omega}^T\hat{x})^2]$.

$\approx \frac{1}{N}\sum_{n=1}^{N}(y_n-\hat{\omega}^T\hat{x}_n)^2$.

14. 请写出线性回归模型的封闭解数学形式，及其推导过程，并标明符号意义。

15. 在线性回归模型中，假设输入样本记为$x_i, i=1,\dots,N$，相应的类别标签记为$y_i, i=1,\dots,N$。请给出自相关矩阵和互相关向量的定义（公式与符号表达）。如果样本集$X\in\Re^{N\times(d+1)}$中，$N<d+1$，应如何处理才能得到合理的模型参数向量唯一解。 $y_i \rightarrow x_i$

标签 样本

16. 请从求解线性方程组的角度说明线性回归模型时的无解情况，从矩阵运算角度说明线性回归模型时的唯一解情况，并比较两者之间的联系与区别。

归纳后的有效方程组

17. 标签y是一个随机变量，由函数$\hat{\omega}^T\hat{x}$加上一个随机噪声生成：$y=\hat{\omega}^T x+\epsilon$；其中噪声$\epsilon$服从正态分布$\mathcal{N}(0,\sigma 2\ )$，设当前训练样本集为$X=\{\hat{x}_i|i=1,\dots,N\}, Y=\{y_i|i=1,\dots,N\}$；

$X=\{\hat{x}_i|i=1\cdots N\}$

$Y=\{y_i|i=1\cdots N\}$

(1) 试推导$\hat{\omega}$的最大似然解的数学形式；

(2) 假设$\hat{\omega}$的先验分布服从$d+1$元正态分布$\mathcal{N}\left(0, I_{(d+1)\times(d+1)}\right)$,试采用最大后验概率估计法推导$\hat{\omega}$的最优解的数学形式.

18. 如何防止线性回归模型出现过拟合现象，请给出具体的方案

19. 请给出逻辑回归模型中 sigmoid 函数形式的推导过程，并说明其概率解释。

20. 请给出逻辑回归模型梯度下降法更新公式的推导过程。

$x+$

五、计算（画图）题

21. 已知两个类别ω_1和ω_2，其先验概率相等，两类的类条件概率密度服从正态分布，有$p(x|w_1)=\mathcal{N}\left(\mu_1,\Sigma\right)$和$p(x|w_2)=\mathcal{N}\left(\mu_2,\Sigma\right)$，且$\mu_1=\begin{bmatrix}0\\0\end{bmatrix}$，$\mu_2=\begin{bmatrix}3\\3\end{bmatrix}$，$\Sigma=\begin{bmatrix}1.1 & 0.3\\0.3 & 1.9\end{bmatrix}$，试写出上述问题每一类的判别函数$g_i(x), i=1,2$的线性形式以及整个分类问题的决策面方程的线性形式。并判断样本$x=\begin{bmatrix}1.2\\1.9\end{bmatrix}$属于哪一类，给出计算过程。

22. 已知决策面方程： $5x_1-x_2-1=0$，当前错分样本为：(0.4,0.6), (0.1−0.25)。

$5\times0.4-0.6-1=0.4$

(1) 写出参数向量w ，并设定合适的学习步长η；

$5\times0.1+0.25-1=-0.25$

23.

$\hat{w} = (X^T X)^{-1} X^T y$

$X = [$花萼长度 花萼宽度. 花瓣宽度$]$.

$X^T{}_{3\times 1} \cdot X_{1\times 3} = (X^T X)_{3\times 3}$.

$X^T{}_{3\times 1} \cdot y_{1\times 1} = (X^T y)_{3\times 1}$.

$\hat{w} = (X^T X)^{-1} X^T y = \hat{w}_{3\times 1}$.

(2) 列出迭代公式和计算结果

(3) 列出迭代后的决策面方程

(4) 画出相应的决策面（虚线）

12. 解： $5x_1-x_2-1=0$.

$(0.4, 0.6)$

$(0.1, -0.25)$

$$\begin{cases} 5\times0.4-0.6-1=0.4 \\ 5\times0.1+0.25-1=-0.25 \end{cases}$$

$$W_{t+1}=W_t-\rho_t\sum_{x\in\Omega(x)}\delta_x \hat{x}$$

$$W_1=W_0-0.7\times(1)\times\begin{bmatrix}0.4\\0.6\\1\end{bmatrix}-0.7\times(-1)\times\begin{bmatrix}0.1\\-0.25\\1\end{bmatrix}=\begin{bmatrix}4.79\\-1.245\\-1\end{bmatrix}$$

$$W_2=W_1-0.7\times(1)\times\begin{bmatrix}0.4\\0.6\\1\end{bmatrix}-0.7\times(-1)\times\begin{bmatrix}0.1\\-0.25\\1\end{bmatrix}=\begin{bmatrix}4.58\\-1.84\\-1\end{bmatrix}.$$

$$W_3=W_2-0.7\times\begin{bmatrix}0.4\\0.6\\1\end{bmatrix}+0.7\times\begin{bmatrix}0.1\\-0.25\\1\end{bmatrix}=\begin{bmatrix}4.37\\ \\ \end{bmatrix}$$

(2) 列出迭代公式和计算结果

(3) 列出迭代后的决策面方程

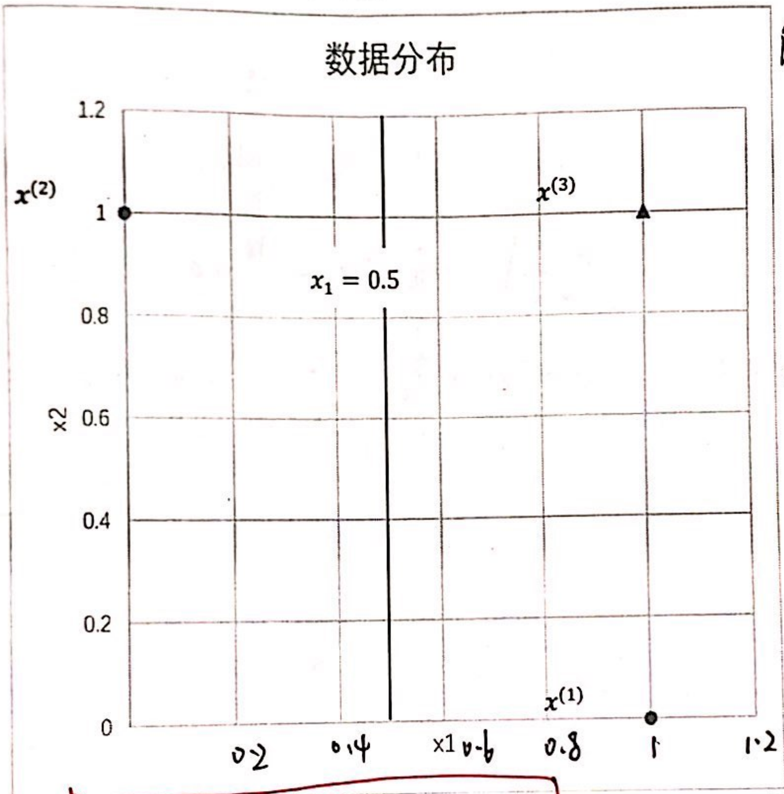
(4) 画出相应的决策面（虚线）

(5) 给出结论与分析

23. 以 Iris 数据库每类的 70%的样本为训练集，基于线性回归模型，使用花萼长度、花萼宽度和花瓣宽度特征来估计其花瓣长度特征，请写出线性回归模型参数的封闭解的数学形式，标明每个变量的矩阵大小或矢量维度。

~~24~~. 已知 3 个样本坐标为$x^{(1)} = [1,0], x^{(2)} = [0,1], x^{(3)} = [1,1]$，其中$x^{(1)}, x^{(2)} \in \omega_1, x^{(3)} \in \omega_2$，初始化逻辑回归模型对应的线性决策面为：$x_1 = 0.5$。设学习步长为 0.2。请计算基于逻辑回归梯度下降法进行一次迭代后的决策面方程，并在下图中画出相应的决策面直线。

$w_1x_1 + w_2x_2 + w_3 = 0$



$$\hat{w}^{(1)} = \hat{w}^{(0)} - \rho \cdot \frac{\partial J(\hat{w})}{\partial w}$$

$$= \begin{bmatrix} 1 \\ 0 \\ -0.5 \end{bmatrix} - 0.2 \times \begin{bmatrix} 0.082 \\ 0 \\ -0.13 \end{bmatrix}$$

$$= \begin{bmatrix} 0.98 \\ 0 \\ -0.47 \end{bmatrix}.$$

$0.98x_1 - 0 - 0.47 = 0$

$x_1 \doteq 0.48$.

解：步长 $\rho = 0.2$. $\boxed{\hat{w} = [w_1, w_2, w_3] = [1, 0, -0.5]}$. $y = [1, 1, 0]^T$.

$\hat{x}_1 = [1, 0, 1]$

$\hat{x}_2 = [0, 1, 1]$

$\hat{x}_3 = [1, 1, 1]$.

$$\frac{\partial J(\hat{w})}{\partial w} = \frac{1}{N}\sum_{i=1}^{N}\left(h(\hat{x}^{(i)}) - y^{(i)}\right)\hat{x}^{(i)}$$

$$= \frac{1}{3}\left\{\left(\frac{1}{1+e^{[1,0,1]^T\cdot(-1)\cdot[1,0,-0.5]}} - 1\right)\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} + \right.$$

$$\left(\frac{1}{1+e^{[0,1,1]^T\cdot(-1)\cdot[1,0,-0.5]}} - 1\right)\begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} +$$

$$\left.\left(\frac{1}{1+e^{[1,1,1]^T\cdot(-1)\cdot[1,0,-0.5]}} - 1\right)\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} 0\right\} = [0.082, 0, -0.1

方程组 ① 化简后的有效方程组的个数小于未知个数 ⇒ 无穷多个解.

② - - - - - 出现 $(0=d)$ 型不兼容方程. ⇒ 无解.

③ 不出现第②种情况且未知数等于方程组个数 ⇒ 唯一解.

$=16$.

矩阵 非齐次线性方程组 $Ax=b$

- $r(A) < r(A,b)$ 无解
- $r(A) = r(A,b) < n$ 多个解
- $r(A) = r(A,b) = n$ 唯一解.

齐次方程

- $\det A == 0$. 多解.
- $\det A \neq 0$ 唯一解

加入正则化, 一般有 L_1-normal 和 L_2-normal.

损失函数. $J(\theta) = \frac{1}{2}(h_\theta(x^{(i)}) - y^{(i)})^2$

normal $J(\theta) = \frac{1}{2}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \left|\sum_{i=0}^{n} \theta_j\right|$ $\lambda > 0$. $=> 18$

normal. $J(\theta) = \frac{1}{2}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{i=0}^{n} \theta_j^2$ $\lambda > 0$

是参数 $\boxed{\text{模型}}$.

问题中

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

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

$= \ln \frac{p(x|w_1)p(w_1)}{p(x|w_2)p(w_2)}$ (在 w_i 服从多元正态分布的情况下且 $p(w_1)=p(w_2)$, $\Sigma = \Sigma_1 = \Sigma_2$).

$= \ln p(x|w_1) - \ln p(x|w_2) = -\frac{1}{2}(x-u_1)^T \Sigma_1^{-1}(x-u_1) + \frac{1}{2}(x-u_2)^T \Sigma_2^{-1}(x-u_2)$ $=> 19$.

$= W^T x + W_0 = \hat{W}^T \hat{x}$

为 $\frac{y}{1-y} = e^{\hat{W}^T \hat{x}}$ $\frac{1}{y} - 1 = \frac{1}{e^{\hat{W}^T \hat{x}}}$ $y = \frac{1}{1+e^{-\hat{W}^T \hat{x}}}$

代价函数的推导 (二分类)

$$\hat{w}) = -\sum_{x\in X}\left(y\ln p(\omega_1|x) + (1-y)\ln(1-p(\omega_1|x))\right) \Rightarrow p(\omega|x) = \begin{cases} p(\omega_1|x) & y=1 \\ 1-p(\omega_1|x) & y=0. \end{cases}$$

$$= -\sum_{x\in X}\left(y\ln\frac{1}{1+\exp(-\hat{w}^T\hat{x})} + (1-y)\ln\frac{\exp(-\hat{w}^T\hat{x})}{1+\exp(-\hat{w}^T\hat{x})}\right)$$

$$= \sum_{x\in X}\left(-y\hat{w}^T\hat{x} + \ln(1+\exp(\hat{w}^T\hat{x}))\right).$$

$$\frac{(\hat{w})}{\hat{w}} = \sum_{x\in X}\left(-y\hat{x} + \frac{\exp(\hat{w}^T\hat{x})}{1+\exp(\hat{w}^T\hat{x})}\hat{x}\right)\Big|_{\hat{x}} = \sum_{x\in X}(h(x)-y)\hat{x}$$

$$\hat{w}(t+1) = \hat{w}(t) - \rho\frac{\partial J(\hat{w})}{\partial \hat{w}} = \hat{w}(t) - \rho\sum_{x\in X}(h(x)-y)\hat{x}$$

$$) = \mu_i^T\Sigma_i^{-1}x + C_i$$

$$) = \mu_1^T\Sigma_1^{-1}x + C_1 = [0\ 0]\begin{bmatrix}1.1 & 0.3\\ 0.3 & 1.9\end{bmatrix}x + (-\tfrac{1}{2})\times\ln\begin{vmatrix}1.1 & 0.3\\ 0.3 & 1.9\end{vmatrix} + \ln\tfrac{1}{2} = -\ln 2 - \tfrac{1}{2}\ln 2.1$$

$$x) = \mu_1^T\Sigma_1^{-1}x + C_2 = [3\ 3]\begin{bmatrix}1.1 & 0.3\\ 0.3 & 1.9\end{bmatrix}x + C_2 = [\cancel{0.8x + C_2}][4.2\ \ 6.6]x + C_2$$

$$(x) = g_1(x) - g_2(x) = \cancel{0.8x}[4.2\ \ 6.6]x = 0.$$

$\wedge\ x = \begin{bmatrix}1.2\\ 1.9\end{bmatrix}$ $\quad -2\ \ 4.2\ \ 6.6]\begin{bmatrix}1.2\\ 1.9\end{bmatrix} = -17.58 < 0$ $\quad \omega_2$类.

$[5\ -1\ -1]^T$ $\quad y=0.5$ (假定)

$$w_0 - y\frac{\partial J(\hat{w})}{\partial w}\Big|_{\hat{w}} = w_0 - y\sum_{x\in\Omega(\hat{w})}\delta x\hat{x} = \begin{bmatrix}5\\ -1\\ -1\end{bmatrix} - 0.5\times(-1)\begin{bmatrix}0.4\\ 0.6\\ 1\end{bmatrix} - 0.5\times(1)\times\begin{bmatrix}0.1\\ -0.25\\ 1\end{bmatrix} = \begin{bmatrix}5.15\\ -0.575\\ -1\end{bmatrix}.$$

设 >0 是 w_1，<0 是 w_2.

ω

$$= \begin{bmatrix}5\\ -1\\ -1\end{bmatrix} - 0.7\times(-1)\begin{bmatrix}0.4\\ 0.6\\ 1\end{bmatrix} - 0.7\times(1)\times\begin{bmatrix}0.1\\ -0.25\\ 1\end{bmatrix} = \begin{bmatrix}5.21\\ -0.755\\ -1\end{bmatrix}$$

$$= \begin{bmatrix}5.21\\ -0.755\\ -1\end{bmatrix} - 0.7\times(-1)\begin{bmatrix}0.4\\ 0.6\\ 1\end{bmatrix} - 0.7\times 1\times\begin{bmatrix}0.1\\ -0.25\\ -1\end{bmatrix} = \begin{bmatrix}5.42\\ -0.(?)\\ -1\end{bmatrix}.$$

w

解：自相关矩阵　　互相关向量

$R_X = X^T X$　　$X^T Y$

其中 $X = \begin{bmatrix} \hat{x}_1^T \\ \hat{x}_2^T \\ \vdots \\ \hat{x}_n^T \end{bmatrix}$　　$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$.

$N < d+1$ 时，可以通过增加正则化项 $\lambda \|w\|^2$ 得到新的最小二乘解.

$$\hat{w} = \left[(X^T X) + \lambda I\right]^{-1} X^T y$$

$\epsilon = y - \hat{w}^T \hat{x} \sim N(0, \delta^2)$.

$$(x; \delta^2, \hat{w}) = \frac{1}{\sqrt{2\pi\delta^2}} \exp\left(\frac{(y-\hat{w}^T \hat{x})^2}{2\delta^2}\right)$$

$$L(X; \delta^2, \hat{w}) = \sum_{i=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi\delta^2}} \exp\left(\frac{(y_i - \hat{w}^T \hat{x}_i)^2}{2\delta^2}\right)\right)$$

$$= \sum_{i}^{N} \left(\frac{(y - \hat{w}^T \hat{x}_i)^2}{2\delta^2} - \frac{1}{2}\ln 2\pi - \frac{1}{2}\ln \delta^2\right)$$

$\Rightarrow$ 当前数据集 X 的整体对数似然性函数.

回归参数 $\hat{w}$ 的估计问题　$\hat{w}^* = \underset{\hat{w}}{\arg\max}\,(LL(X; \delta^2, \hat{w}))$.

$$\frac{\partial LL(X; \delta^2, \hat{w})}{\partial \hat{w}} = \sum_{i=1}^{N} \left(\frac{(y_i - \hat{w}^T \hat{x}_i)\hat{x}_i}{\delta^2}\right) = 0$$

$$\hat{w} = (X^T X)^{-1} X^T y$$

最大后验概率估计

$$J(X; \delta^2, \hat{w}) = \sum_{i=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi\delta^2}} \exp\left(\frac{(y_i - \hat{w}^T \hat{x}_i)^2}{2\delta^2}\right)\right) + \ln\left(\frac{1}{\sqrt{2\pi}} \exp\left(\frac{\hat{w}^T \hat{w}}{2}\right)\right).$$

$$\frac{\partial J(X; \delta^2, \hat{w})}{\partial w} = 0 \Rightarrow \hat{w} = (X^T X + N\delta^2 I)^{-1} X^T y$$

令 $N\delta^2 = \lambda$

则 $\hat{w} = (X^T X + \lambda I)^{-1} X^T y$.

$\max\limits_{\alpha_i, i=1\cdots m} [\sum_{i=1}^m \alpha_i - \frac{1}{2}\sum_{i=1}^m\sum_{j=1}^m \alpha_i\alpha_j y_i y_j x_i^T x_j]$

第四章 支持向量机

一、填空题

1. 50 个样本训练集上的线性 SVM 原始优化问题，具有（ 50 ）个（不等式 ）约束条件. $y_i(w^Tx_i+r)-1\ge 0$, $i=1,2\cdots50$.
2. 线性 SVM 问题经过拉格朗日对偶处理后的优化问题的$\boxed{\text{自变量}}$是（x_i样本） 拉格朗日乘子 α_i.
3. 线性不可分问题可以使用（软间隔）SVM 或（非线性）SVM 进行求解。
4. 常见的核函数有（多项式）核函数、（ 高斯 ）核函数和（sigmoid）核函数.

二、判断题

5. 对于一组线性可分的训练集，所有能够将它们正确分类的线性分类器都具有相同的分类性能。（ X ）
6. 所有支撑向量均在间隔区域的边界上。（ √ ）那软间隔内部的是吗？ 支持向量可以在间隔中.
7. 间隔区域越宽意味着模型的经验风险越小。（ ）
8. 线性 SVM 求解必须依靠拉格朗日对偶技巧。（ X ）. 简化. 原理上直接撸也是可以的.

三、选择题

9. 对于包含 100 个样本的训练集，线性 SVM 原始问题的待优化变量数量为（ B ）D α. w. γ 样本的特征维度未知.

   A. 2 个　　B. 3 个　　C. 101 个　　D. 无法确定

10. 原始线性 SVM 问题经过拉格朗日对偶处理后转化为一个：（ C ）

    A. 无约束最大化问题　　B. 无约束最小化问题 min{max{ }}.

    C. 有约束最大化问题　　D. 有约束最小化问题

11. 利用 SMO 算法求解 SVM 问题时，每轮迭代选取（ B ）个系数变量进行更新

    A. 1 个　　B. 2 个　　C. 样本数量　　D. 无法确定

    选取两个违反KKT条件最严重的参数进行更新

四、简答题

支持向量 每类中距离决策面最近的样本.

1. 简述支持向量的定义，并画图说明。
2. 简述 SVM 优化问题的三要素，并辅助变量/表达式形式进行说明。
3. 写出样本空间中的点x到超平面(ω,γ)的距离公式，并对符号进行解释。
4. 推导线性 SVM 约束条件 $y_i(\omega^T x_i+\gamma)\geq 1,\ \forall x_i$。 $d=\frac{|w^Tx+r|}{\|w\|}$.
5. 推导线性 SVM 的间隔宽度为$\frac{2}{\|\omega\|}$。
6. 写出线性 SVM 原始问题的 KKT 条件。
7. 写出经过拉格朗日对偶变换后线性 SVM 的最优化问题的标准形式。

问题优化三要素
- 决策面方程（优化变量） $g(x)=w^Tx+r=0$
- 分类间隔（目标函数）max. $w\Rightarrow d\Rightarrow \frac{|w^Tx_s+r|}{\|w\|}$
- 所有训练样本被正确分类（约束条件）. $y_i(w^Tx_i+r)\ge 1$.

推导 $\frac{2}{\|w\|}$（间隔宽度）

$$W = 2d = 2 \times \frac{|w^T x_s + r|}{\|w\|}$$

又∵间隔宽度由支持向量决定．

当 x 是支持向量时．

约束条件 $y_i(w^T x_i + r) = 1$ 成立．

又∵ y_i 取 $+1$ 或 -1 所以 $w^T x_i + r = \pm 1$．

$$W = 2d = 2 \times \frac{1}{\|w\|} = \frac{2}{\|w\|}$$

$$\begin{cases} \alpha_i \geq 0 \\ y_i(w^T x_i + r) - 1 \geq 0 \\ \alpha_i(y_i(w^T x_i + r) - 1) = 0 \end{cases}, \quad i = 1, 2, \cdots m.$$

$$\max_{\alpha} \left[\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j \right].$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0.$$

$$\alpha_i \geq 0, \ i = 1, 2, \cdots, m.$$

第四章、支持向量机

4、

每个样本标签 $y_i = \begin{cases} +1 & x_i \in A类 \\ -1, & x_i \in B类 \end{cases}$

对于所有样本 x_i，正确的决策面应该满足.

$$\begin{cases} W^T x_i + r > 0 & , \forall y_i = 1 \\ W^T x_i + r < 0 & , \forall y_i = -1. \end{cases} \quad ①$$

决策面位于间隔区域的中心，所有支持向量到决策面距离为 d.

其他样本到决策面距离一定大于 d

$$\begin{cases} \dfrac{W^T x_i + r}{\|W\|} \geq d & , \forall y_i = 1. \\ \dfrac{W^T x_i + r}{\|W\|} \leq -d, & \forall y_i = -1. \end{cases}$$

令 $\begin{cases} W_d = \dfrac{W^T}{\|W\| d} \\ r_d = \dfrac{r}{\|W\| d}. \end{cases} \Rightarrow \begin{cases} W_d x_i + r_d \geq 1 & , \forall y_i = 1. \\ W_d x_i + r_d \leq -1, & \forall y_i = -1. \end{cases} \quad ②$

由于这个②与①是同一个超平面，所以又可以写作

$$\begin{cases} W x_i + r \geq 1 & , \forall y_i = 1 \\ W x_i + r \leq -1 & , \forall y_i = -1. \end{cases}$$

$\Downarrow$

统一表示法

$$y_i (W x_i + r) \geq 1 \quad , \forall x_i.$$

处理后． $\max\left[\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j\right]$．

$\sum_{i=1}^{m}\alpha_i y_i=0$

$\alpha_i\geq 0,\ i=1,2,\cdots,m$．

与约束条件；

$\alpha_i \geq 0$

3）试写出经过拉格朗日对偶处理后的 3 条 KKT 条件。⇒ 都隔开．对．

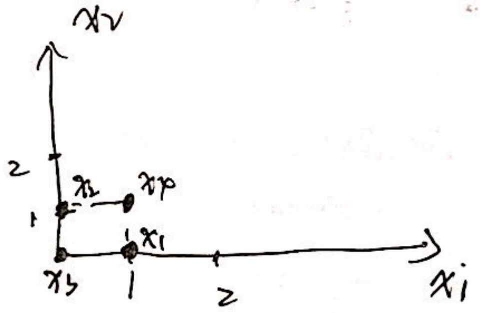
$y_i(w^T x_i+\gamma)-1\geq 0$，$i=1,2,\cdots,m$

4）对照图中的决策面和间隔区域，试推算样本$x_i, i=1,2,3$对应的系数$\alpha_i, i=1,2,3$。

$\alpha_1=\alpha_3=\frac{1}{4}$，$\alpha_2=0$．

$\alpha_i(y_i(w^T x_i+\gamma)-1)=0$

14. 对于 1 个标准的 XOR 问题，$x_1=[1,0], x_2=[0,1], x_3=[0,0], x_4=[1,1]$。其中$x_1, x_2\in\omega_1$类，$x_3, x_4\in\omega_2$类。请设计一个非线性矢量映射函数$\phi(x)$，将样本从 2 维空间映射到 3 位空间，使得数据在 3 维空间中可分，并给出对应的核函数。



化目标为 $\min_{w,r,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i$

约束条件. $y_i(w^T x_i + r) \geq 1-\xi_i$.

$\xi_i \geq 0 \quad \forall x_i$

8. 试分析，在软间隔 SVM 中，当松弛变量ξ_i对应的系数 C 增大时，最优解对应间隔宽度的变化趋势，并解释原因。

~~9.~~ 在一个松弛项系数$C = 0.3$的软间隔 SVM 模型中，如果一个样本x_i对应的系数$\alpha_i = 0.1$，试分析该样本与最优决策面和间隔区域的关系。

≠0

本x_i是个

量

$=0.3$

界上

10. 在非线性 SVM 的推导过程中，为什么要使用非线性映射的策略而不使用构造非线性决策面的策略呢？

非线性决策面的数学模型多种多样，我们不知道该采用哪种才能使问题有最优解，与之相比，非线性映射更为简单。

11. 试说明，如何判断一个函数是不是核函数

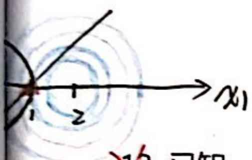
五、计算（画图）题

12. 求解有约束优化问题，并给出求解过程。

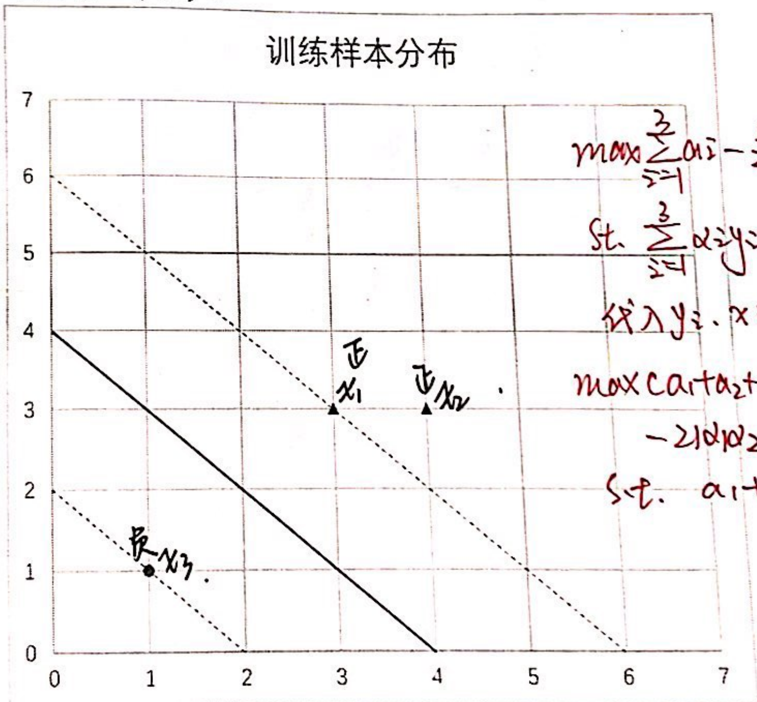
$$\min_{x}[(x_1-2)^2 + x_2^2]$$

$$\text{s.t. } x_2 - x_1 + 1 = 0$$
$$x_1{}^2 + x_2^2 - 1 \leq 0$$

x_2



~~13.~~ 已知一个如下图所示的训练数据集，其正类样本为$\boldsymbol{x_1} = (\mathbf{3,3})^T, \boldsymbol{x_2} = (\mathbf{4,3})^T$，负类样本是$\boldsymbol{x_3} = (\mathbf{1,1})^T$。



$\max \sum_{i=1}^{3}\alpha_i - \frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3}\alpha_i\alpha_j y_i y_j x_i^T x_j$

$\text{s.t. } \sum_{i=1}^{3}\alpha_i y_i = 0 ; \ \alpha_i \geq 0 \quad i=1,2,3$

代入y_i、x_i得.

$\max(\alpha_1+\alpha_2+\alpha_3 - 9\alpha_1^2 - 12.5\alpha_2^2 - \alpha_3^2 - 21\alpha_1\alpha_2 + 7\alpha_2\alpha_3 + 6\alpha_3\alpha_1)$

$\text{s.t. } \alpha_1+\alpha_2-\alpha_3 = 0 ; \ \alpha_i \geq 0 \quad i=1,2,3$

1）试写出上述问题的线性 SVM 原始优化问题的数学形式，包含目标函数与约束条件；

2）试写出经过拉格朗日对偶处理后的优化问题的数学形式，包含目标函数

处理前

1. $\min_{w,r} \frac{1}{2}\|w\|^2$

$\text{s.t. } y_i(w^T x_i + r) \geq 1 \quad i=1,2,\cdots,m.$

写开

2). $L(w,r,\alpha) = \frac{\|w\|^2}{2} + \sum_{i=1}^{m}\alpha_i(1-y_i(w^T x_i + r))$

$\max\{\min_{w,r} L(w,r,\alpha)\}$

化目标为 $\min_{w,r,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i$

束条件. $y_i(w^T x_i + r) \ge 1 - \xi_i$.

$\xi_i \ge 0 \quad \forall i$

8. 试分析，在软间隔 SVM 中，当松弛变量ξ_i对应的系数 C 增大时，最优解对应间隔宽度的变化趋势，并解释原因。

=0

9. 在一个松弛项系数$C = 0.3$的软间隔 SVM 模型中，如果一个样本x_i对应的系数$\alpha_i = 0.1$，试分析该样本与最优决策面和间隔区域的关系。

本x_i是个

量.

10. 在非线性 SVM 的推导过程中，为什么要使用非线性映射的策略而不使用构造非线性决策面的策略呢？

非线性决策面的数学模型多种多样，我们不知道该采用哪种才能使问题有最优解。与之相比，非线性映射更为简单.

=0.3.

11. 试说明，如何判断一个函数是不是核函数

界上.

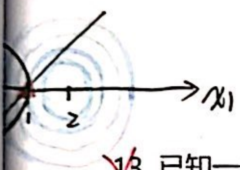
五、计算（画图）题

x_2

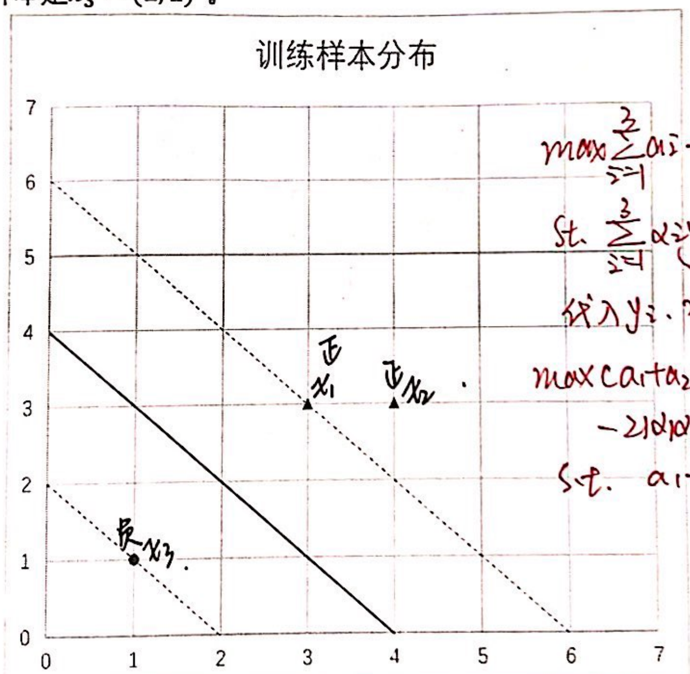
12. 求解有约束优化问题，并给出求解过程。

$$\min_{x}[(x_1-2)^2+x_2^2]$$

$$\text{s.t. } x_2 - x_1 + 1 = 0$$
$$x_1{}^2 + x_2^2 - 1 \le 0$$



13. 已知一个如下图所示的训练数据集，其正类样本为$\boldsymbol{x_1} = (\mathbf{3,3})^T$, $\boldsymbol{x_2} = (\mathbf{4,3})^T$，负类样本是$\boldsymbol{x_3} = (\mathbf{1,1})^T$。



max $\sum_{i=1}^{3}\alpha_i - \frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3}\alpha_i\alpha_j y_i y_j x_i^T x_j$

s.t. $\sum_{i=1}^{3}\alpha_i y_i = 0$; $\alpha_i \ge 0$ $i=1,2,3$

代入 y_i, x_i 得.

max $(\alpha_1+\alpha_2+\alpha_3 - 9\alpha_1^2 - 12.5\alpha_2^2 - \alpha_3^2 - 21\alpha_1\alpha_2 + 7\alpha_2\alpha_3 + 6\alpha_3\alpha_1)$

s.t. $\alpha_1+\alpha_2-\alpha_3=0$; $\alpha_i \ge 0$ $i=1,2,$

1）试写出上述问题的线性 SVM 原始优化问题的数学形式，包含目标函数与约束条件；

处理前

2）试写出经过拉格朗日对偶处理后的优化问题的数学形式，包含目标函数

$\min_{w,r} \frac{1}{2}\|w\|^2$

s.t. $y_i(w^T x_i + r) \ge 1$ $i=1,2,\cdots,m$.

写开

2). $L(w,r,\alpha) = \frac{\|w\|^2}{2} + \sum_{i=1}^{m}\alpha_i(1 - y_i(w^T x_i + r))$

$\max\{\min_{w,r} L(w,r,\alpha)\}$

样本x_i对应的系数$\alpha_i=0.1$、即有$\alpha_i>0$.

根据KKT条件$\alpha_i(1-\xi_i-y_if(x_i))=0$.

推出$1-\xi_i-y_if(x_i)=0$、因此x_i是一个支持向量.

$\alpha_i=0.1<C=0.3$. $\Rightarrow$ KKT条件$C=\alpha_i+\mu_i$得$\mu_i>0$.

又∵ KKT中$\mu_i\xi_i=0$、推得$\xi_i=0$. 即松弛变量为0.

⇓

x_i在间隔区域的边界上.

第五章 决策树

一、填空题

1. 大多数决策树的节点类型可以分为（内部节点）和（叶子节点）。根节点在类型上属于（内部）节点，剪枝节点减去的是（叶子）节点？

编程意义上.

二、判断题

2. 基于基尼指数的分裂结果与基于信息增益比的分裂结果总是相同的。（√）×

大多数

3. 信息增益越大的特征信息增益比也越大。（×）不一定

4. 信息增益比越大的特征不纯度越小。（√）

基尼指数

$Gini = 1-\sum_{j=1}^{M} p_j^2$

$Gini\_sorscore(l, f_l) = \sum_{k=1}^{K_l} \frac{N_l^k}{N} Gini(...)$

信息增益比 $R_l = \frac{G_l}{A(f_l(x))} = \frac{G_l}{-\sum_{k=1}^{K_l}\frac{1}{K_l}\log\frac{1}{K_l}}$

三、选择题

5. 假设离散特征 A 所有可能的取值为$\{a_1, a_2, a_3\}$，设当前节点样本集为S，且有$A(x_i) \neq a_2, \forall x_i \in S$，则利用特征 A 进行分裂后产生几个子节点：（B）

$l^* = \arg\min_l [Gini\_Score]$

A. 2个　训练样本无 但是未来测试可能会出现 需要保留　B. 3个　C. 无法确定　D. 无法分裂

6. 当前节点内样本集包含 5 个样本，特征数量为 3，其中两个离散特征，1 个连续特征，采用多叉树方案，则当前节点共有多少种可选的分裂方案。C

A. 1种　B. 2种　C. 3种　D. 无法确定

每个特征对应一种分裂方案

当前叶子节点的父节点对应的样本集中没有符合该叶子节点特征取值要求的样本，但分裂特征的所有可能取值在样本集中又包含了该叶子节点对应的特征取值

四、简答题

7. 简述如何利用"分而治之"策略解决复杂非线性分类问题

利用树形决策超平面将特征空间划分为多个局部区域，子区域内样本均来自同一类别，落入子区域S_t的样本均应该被分为l类.

8. 什么情况下一个叶子节点中会没有样本，此时该叶子节点返回的类别标签如何确定。把当前结点标为叶结点，将其类别设定为其父结点所含样本最多的类别.

9. 简述前剪枝与后剪枝的差别及各自的特点

前剪枝是一边生成决策树一边进行剪枝.

后剪枝是将完整的决策树生成后再进行剪枝的

特点：前剪枝训练、测试时间较短，过拟合风险低但欠拟合高，泛化一般

后剪枝训练时间较长，测试时间较少，过/欠拟合风险低，泛化好.

五、计算（画图）题

10. 一下关于 iris 数据库的某个特征增广版本包含 7 个样本，具体情况如下：

序号	香气	颜色	花萼长度	花萼宽度	花瓣长度	花瓣宽度	类别
1	有	红	5.1	3.5	1.4	0.2	setosa
2	有	红	4.9	3	1.4	0.2	setosa
3	有	粉	4.7	3.2	1.3	0.2	setosa
4	有	紫	5.3	3.7	1.5	0.2	setosa
5	无	粉	7	3.2	4.7	1.4	versicolor
6	无	紫	6.4	3.2	4.5	1.5	versicolor

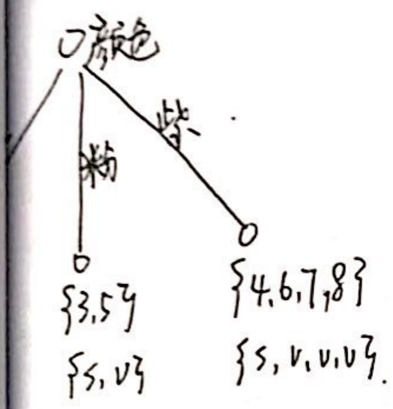
$=\{1,2,3,4,8\}$.

$=\{5,6,7\}$.

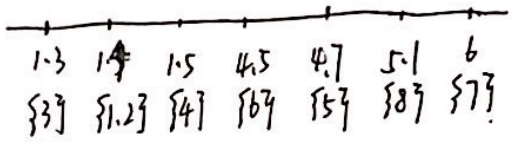
红 $=\{1,2\}$

粉 $=\{3,5\}$

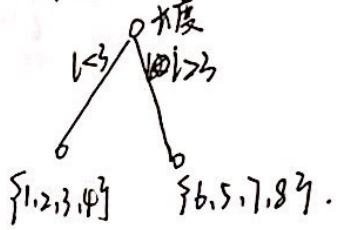
紫 $=\{4,6,7,8\}$.



(2).



随意取中值 $l=\frac{1}{2}(1.5+4.5)=3$.



增益.

$a=\{1,2,3,4\}$.

color $=\{5,6,7,8\}$.

$\hat{H}(Y)=(\frac{1}{2}\log_2\frac{1}{2}+\frac{1}{2}\log_2\frac{1}{2})\times(-1)=\log_2 2=1$.

颜色.

$\hat{H}(Y_2^1)=\cancel{\ }-(\frac{4}{5}\times\log_2\frac{4}{5}+\frac{1}{5}\log_2\frac{1}{5})$

$\hat{H}(Y_2^2)=-(0)=0$.

$=-(1\times\log_2 1)=0$.

$=-(\frac{1}{2}\log_2\frac{1}{2}+\frac{1}{2}\log_2\frac{1}{2})=\log_2 2=1$

$=-(\frac{1}{4}\log_2\frac{1}{4}+\frac{3}{4}\log_2\frac{3}{4})=-[\frac{1}{4}\times(-2)+(2+\log_2 3)\times\frac{3}{4}]=2-\frac{3}{4}\log_2 3$.

$\hat{H}_1=0\times\frac{3}{8}+\frac{5}{8}\times(-\frac{4}{5}\times\log_2\frac{4}{5}-\frac{1}{5}\log_2\frac{1}{5})\cancel{=\frac{5}{8}}$
$=\frac{5}{8}\log_2 5-1$

$0+1\times\frac{2}{8}+(2-\frac{3}{4}\log_2 3)\times\frac{4}{8}=\frac{5}{4}-\frac{3}{8}\log_2 3$.

$G_2=\hat{H}(Y)-\hat{H}_2=2-\frac{5}{8}\log_2 5$

$=\hat{H}(Y)-\hat{H}_1=\frac{3}{8}\log_2 3-\frac{1}{4}$

增益比.

$\frac{G_L}{H(f_L(x))}$ $H(f_L) = -\sum_{k=1}^{K_L} \frac{N_L^k}{N} \log \frac{N_L^k}{N}$

$) = -1 \times (\frac{5}{8}\log_2 \frac{5}{8} + \frac{3}{8}\log_2 \frac{3}{8}) = \frac{1}{8}(5\log_2 5 + 3\log_2 3) - 3.$

$\frac{G_2}{H(f_2)} = \frac{2 - \frac{5}{8}\log_2 5}{\frac{1}{8}(5\log_2 5 + 3\log_2 3) - 3}.$

$= -1 \times (\frac{2}{8}\log_2 \frac{2}{8} + \frac{2}{8}\log_2 \frac{2}{8} + \frac{4}{8}\log_2 \frac{4}{8}) = 1.5$

$\frac{G_1}{H(f_1)} = \frac{\frac{3}{8}\log_2 3 - \frac{1}{4}}{\frac{3}{2}} = \frac{1}{4}\log_2 3 - \frac{1}{6}$

指数.

~~$Y_1) = 1 - (1^2 + (\frac{1}{2})^2 +$~~

$(Y=红色) = 1 - (1^2) = 0.$

$(Y=粉) = 1 - [(\frac{1}{2})^2 + (\frac{1}{2})^2] = 1 - [\frac{1}{2}] = \frac{1}{2}$

$Y=紫) = 1 - [(\frac{1}{4})^2 + (\frac{3}{4})^2] = 1 - \frac{5}{8} = \frac{3}{8}$

$\_score(Y, f_1) = 0 + \frac{2}{8} \times \frac{1}{2} + \frac{4}{8} \times \frac{3}{8} = \frac{3+2}{16} = \frac{5}{16}$

看点:

$Gini = 1 - [(\frac{4}{5})^2 + (\frac{1}{5})^2] = \frac{8}{25}$

$Gini = 1 - [\phi] = 0.$

$Gini\text{-}Score(Y, f_2)$

$= \frac{8}{25} \times \frac{5}{8} = \frac{1}{5}$

选看点

7	无	紫	6.3	3.3	6	2.5	virginica
8	有	紫	5.8	2.7	5.1	1.9	virginica

1) 请给出采用颜色特征进行多叉树分裂的结果
2) 请给出采用花瓣长度进行二叉树分裂的任意一个结果
3) 对比香气和颜色两种离散特征，分别依据信息增益、信息增益比和基尼指数给出相应的分裂特征选择结果，并给出计算过程。

11. 根据表 5.1 计算各个特征的信息增益、信息增益比、基尼系数。

编号	天气	温度	湿度	有无风	是否出去玩
1	晴	热 4	高 7	无 8	否 5
2	晴	热	高	有 6	否
3	阴	热	高	无	是 9
4	雨	温和 6	高	无	是
5	雨	凉爽 4	正常 7	无	是
6	雨	凉爽	正常	有	否
7	阴	凉爽	正常	有	是
8	晴	温和	高	无	否
9	晴	凉爽	正常	无	是
10	雨	温和	正常	无	是
11	晴	温和	正常	有	是
12	阴	温和	高	有	是
13	阴	热	正常	无	是
14	雨	温和	高	有	否

晴5 雨5 热.

阴4

信息增益

1) 使用表 5.1 数据集，根据 ID3 决策树算法，手动生成一棵决策树，用于预测是否应该出去玩。请写出根节点的分裂依据计算过程
2) 使用表 5.1 数据集，根据 C4.5 决策树算法，在不考虑剪枝的情况下，手动生成一颗决策树，用于预测是否应该出去玩。请写出根节点的分裂依据计算过程

基尼

1). 信息增益.

$\hat{H}(Y) = -\sum_{j=1}^{2} P(y=w_j)\log_2 P(y=w_j) = -[\frac{5}{14}\times\log_2\frac{5}{14} + \frac{9}{14}\log_2\frac{9}{14}]$

$D_1=3.\ D_2=5\ \ D_3=4.$

天气

$\hat{H}(Y_1^1) = -(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5})$

$\hat{H}(Y_1^2) = -(\frac{4}{4}\log_2 1 + 0) = 0.$

$\hat{H}(Y_1^3) = -(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5})$

$\hat{H}_1 = \frac{5}{14}\times\hat{H}(Y_1^1) + \frac{4}{14}\hat{H}(Y_1^2) + \frac{5}{14}\hat{H}(Y_1^3).$

$\hat{H}(Y_2^1) = -(-$

第六章 集成学习

一、填空题

1. 如果将多项式看做一个集成学习系统，其个体学习器是（~~基~~ 多项式）函数，个体学习器的参数是（权重）。多项式的系数及常数项。
2. 当所有个体学习器类型相同，称它们是（~~基学习~~ 同质）的，此时的个体学习器又称为（基学习器）。
3. Boosting 方法采用（串行）方式学习个体分类器，Bagging 方法采用（并行）方式学习个体分类器。
4. Adaboost 算法中集成学习器$H(x)$应最小化（指数）损失函数。
5. 在标准 Adaboost 算法中个体学习器$h_t(x)$输出的结果为（+1）或（-1）
6. 在标准 Adaboost 算法中个体学习器$h_t(x)$的正确率为 75%，则α_t的值为（ ）$\frac{1}{2}\ln 3$.
   $\varepsilon_t = 25\%$ $\alpha_t = \frac{1}{2}\ln\frac{75\%}{25\%} = \frac{1}{2}\ln 3$.
7. 在标准 Adaboost 算法中，当训练集有 N 个样本时，每个样本的初始化权重为（ ）。$D_1(x) = [w_{1i}^{(t)} | i=1,\cdots N; t=1] = [\frac{1}{N}, \frac{1}{N} \cdots \frac{1}{N}]$.

二、判断题

8. 使用相对多数投票时，若没有任何一类得票数高于 50%，则拒绝分类。（√）
9. 采用前向分步优化策略得到的最优解就是集成学习问题的全局最优解。（X）不一定是全优
10. Adaboost 算法中，被个体学习器h_1错误分类的样本的权重在h_2的学习中会增加。（√）样本x_i在h_1中的权重为$w_{1i} = \frac{1}{N}$
    对于误错误x_i，$w_{2i} = \frac{w_{1i}}{Z_1} e^{-\alpha_1 f(x_i) h_1(x_i)}$

三、选择题

11. 设样本集为$X = \{1,3,5,7\}, Y = \{+1,-1,-1,+1\}$，权重为$\mathcal{D}_t = \{0.1,0.5,0.2,0.1\}$，个体分类器函数形式为$h_t(x) = \text{sign}(x - v)$。则当前$v$的最优取值为：（D）

A. 0　　B. 2　　C. 4　　D. 6

12. 当样本数趋近于无穷大时，Bootstrap 采样时样本被抽中的概率约为：（ ）

A. 0.368　　B. 0.612　　C. 1　　D. 0.5

四、简答题

13. 设训练集为$\{x_i, y_i | i = 1, \dots, N\}$，请将加法模型的学习问题转化为参数$\theta_t$与系数$\alpha_t$的优化问题，并写出目标函数表达式，并加以说明
14. 简述集成学习中个体学习器的设计与学习要满足什么样的规则，并加以解释。

15. 试说明 boosting 方法与 bagging 方法的主要异同之处。
16. AdaBoost 的算法采用什么措施使个体分类器“不同”。
17. 随机森林算法采用哪些措施使个体分类器“不同”。

五、计算（画图）题

18. 给定如表所示训练数据。假设弱分类器由$x < v$或$x > v$产生，试使用 AdaBoost 算法求解。

序号	1	2	3	4	5	6
x	0	1	2	3	4	5
y	1	1	-1	-1	1	1

1) 给出前两个个体分类器的训练过程和结果
2) 采用 Bootstrap 方法对上述训练集进行采样，从统计效果上看，单个样本在某一轮采样中未被选中的概率是多少？

19. 某集成学习框架下，5 个个体分类器对于一个二分类问题各自给出的结果及相应的系数如下表所示。请分别根据绝对多数投票，相对多数投票和加权投票的策略给出集成学习框架的最终识别结果。

分类器序号	1	2	3	4	5
权重	0.1	0.2	0.3	0.3	0.1
分类结果	是	否	是	否	是

第七章 模型评估

一、填空题

1 在当前训练集上经验误差为 0 的假设集合称为（版本空间；符合参数化模型函数形式的所有映射关系的集合称为（假设空间

2 当两个假设在训练集上具有完全相同的经验误差时，应依据（~~奥卡姆剃刀~~ 归纳偏好）来进行假设选择。

3 函数$h(x;a,b) = \mathrm{sign}(ax-b), a,b \in \mathfrak{R}$支撑的假设空间为$\mathcal{H}$，则增长函数$\Pi_{\mathcal{H}}(3) =$（~~4~~）

等价于对[illegible]函数.

4 线性分类器$h(x) = w_1x_1 + w_2x_2 + w_0$的 VC 维是（3 ）

5 使用 10-折交叉验证法处理 Iris 数据集，则每一次验证时训练集中的 setosa 类样本数量是（~~15×9~~），测试集的样本数量是（15 ）

$\frac{1}{3}$ 45 150/10 $\frac{150}{10} = 15$

二、判断题

6 一个高效 PAC 可学问题一定是 PAC 可学问题。（√）

7 模型在独立同分布采样获得的训练集上的经验误差总是大于该模型在该分布上的泛化误差。（X）不一定，泛化误差[illegible]大于经验误差，但[illegible]不一定

8 若存在假设$h \in \mathcal{H}$在数据集 D 上的经验误差为零，则有真实映射$f \in \mathcal{H}$。（X）

泛化

三、选择题

9 当样本数为N，精度为ϵ，置信度为δ时，下述那种学习时间T对应的学习算法是 PAC 学习算法

A. $T = N!$ X

B. $T = N\log N + 2^{1/\epsilon}$

C. $T = N\log N + \left(\frac{1-\delta}{0.05}\right)^5$

D. $T = \frac{2^N}{\delta^2}$ X

四、简答题

10 请简述"奥卡姆剃刀"原理，并说明其在模型评估中的应用。

11 请简述"没有免费午餐"定理，并说明其对于机器学习研究的意义。

12 请简要阐述经验误差与泛化误差之间的关系。

13 请简述 PAC 学习理论中 P，A 的概念及其对应于机器学习结果的哪些性能。

14 请简述不可知 PAC 学习与可知 PAC 学习的最主要差别

15 请结合本课程讲授的分类算法，说明三种具体的正则化方法。

五、计算(画图)题

16 考虑一个二分类问题，$x \in \Re^2, y \in \{+1,-1\}$. 给定一个假设空间如下：

$$\mathcal{H} = \{h(x;a,b) = \mathrm{sign}[a(x^T x - b)] \mid a \in \{+1,-1\}, b \in [0,+\infty)\}$$

其中，函数$\mathrm{sign}(x) = \begin{cases} -1, & x < 0 \\ +1, & x \geq 0 \end{cases}$. 请通过画图法找出假设空间$\mathcal{H}$的 VC 维 d 的具体取值（提示：画出$\mathcal{H}$打散 d 个样本但无法打散 d+1 个样本的情况）

17 某二分类任务训练样本集为：$x^{(i)} = \left[x_1^{(i)}, x_2^{(i)}, x_3^{(i)}\right]^T, i = 1,\ldots,100$，其中特征$x_1$有 3 个取值，特征$x_2$有 2 个取值，特征$x_3$有 2 个取值。某个集成分类器算法$\mathcal{L}_1$的假设空间$\mathcal{H}_1$包含该问题上所有可能的映射。

1) 请计算出假设空间$\mathcal{H}_1$的大小。 $3\times2\times2=12$? $2^3+2^2+2^2$?
2) 请判断算法$\mathcal{L}_1$能否在当前数据集下以不低于 80%的概率获得泛化误差不超过 0.1 的识别模型，并给出计算过程。
3) 如果存在一个算法$\mathcal{L}_2$，可以生成 512 种不同的假设，且在当前数据机上的误差为 0.05。试计算$\mathcal{L}_2$在该问题上的泛化误差上界。

18 已知测试集真实标签 Y 和分类器分类结果 H 如下

样本序号	1	2	3	4	5	6	7	8	9	10
Y	P	N	N	P	P	P	N	P	N	P
H	P	N	P	P	N	P	P	P	N	N

请计算以下性能指标，并给出计算方法。

1) 真阳率
2) 假阳率
3) 查准率
4) 查全率
5) F1 度量

$\frac{8.5}{200}$

1). 真阳率

$$TPR = \frac{TP}{TP+FN} = \frac{①+④+⑥+⑧}{①+④+⑥+⑧+⑤+⑩} = \frac{4}{6} = \frac{2}{3}$$

2). $FPR = \frac{FP}{FP+FN} = \frac{③+⑦}{4} = \frac{1}{2}$

3). $\mathrm{precision} = \frac{TP}{TP+FP} = \frac{4}{4+2} = \frac{2}{3}$

4). $\mathrm{recall} = TPR = \frac{2}{3}$

5). $F_1 = \frac{2TP}{2TP+FP+FN} = \frac{2\times4}{2\times4+2+2} = \frac{8}{12} = \frac{2}{3}$

奥卡姆剃刀原理：如非必要，勿增实体。

在两种模型都能使经验误差为零时，我们选择更加简单的一种

该两种机器学习算法LA和LB在所有任务上的平均性质是相同的

不存在某种机器学习算法适用于所有任务。

在数据量趋于无穷大时，泛化误差与经验误差相等 $E(h)=\lim_{n\to\infty}\hat{E}(h)$

样本 经验误差的数学期望等于泛化误差

利用经验误差与模型参数可以描述泛化误差的边界。

→ probably → 置信度 $1-\delta$

→ Approximately → 错误概率边界 ϵ

真实映射属于假设空间时是可知PAC

不属于时是不可知PAC学习

回归 ⇒ 正则项是 $\lambda\|\omega\|^2$，在损失函数中加入，得到岭回归系数的解。

回归 ⇒ 由于采用的是正态分布假设，模型已知。

树 ⇒ (1) 加入剪枝操作，能降低决策树的复杂度。软间隔SVM，带有结构风险项

SVM在线性SVM基础上加入松弛变量

$=\mathrm{sign}C\ \mathrm{sign}[a(x^Tx-b)]$ 以

17.

① . $H = \cancel{\quad} 3 \times 2 \times 2 = 12$.

② . $P(E(h) \le 0.1) \ge 80\% = 1 - 0.2$

$\varepsilon = 0.1$

$\delta = 0.2$

$N \ge \frac{1}{\varepsilon}(\ln(H) + \ln\frac{1}{\delta}) = 10 \times (\ln 12 + \ln 5) = 10 \times \ln 60 \approx 40.9$

可以

③ . $H_2 = 512$

$\varepsilon_2 = \sqrt{\frac{1}{2N}(\ln(H) + \ln\frac{2}{\delta})} = \sqrt{\frac{1}{200} \times (\ln 512 + \ln 10)} \doteq \sqrt{\frac{1}{200} \times (6.2 + 2.3)} \doteq 0.21$

第八章 特征选择与学习

一、填空题

1 数据的均值、方差、直方图等特征属于（通用）特征；工业生产中的温度、压力、速度等特征属于（结构化）特征；

2 比较常见的线性子空间特征包括（主成分分析（独立成分分析和（线性判别分析 请用中文回答）

3 比较常见的特征搜索策略包括（完全搜索）和（启发式搜索、随机搜索、

4 MDS 和 IsoMap 的相同之处是（~~将高维映射到低维~~）不同之处是前者使用（欧氏距离）距离，后者使用（测地线）距离 ~~映射方式非线性~~ 降维前后保持

5 10 维空间中的 5 个样本，最多需要（4）维的子空间可以保证投影损失为 0。 2维空间中的2个样本，可以放入一条直线上，样本间距离保持不变。

6 为了保证编码的稀疏性，标准的稀疏编码算法使用（重构误差 为0）作为约束条件。

二、判断题

7 表情特征可以用于身份识别。（ ✗ ）

8 稀疏编码方法中的编码数值是通过将原始数据向字典上投影得到的。（ ）

9 IsoMap 流形学习算法无法直接实现训练集以外的新样本的特征提取。（ √ ）

10 稀疏编码算法中的重构损失和稀疏性均可作为目标函数或约束条件。（ ）

11 稀疏编码算法的字典中的特征向量必须正交。（ ✗ ）一般不正交，正交太多无法满足编码稀疏性。

12 PCA 算法中的基向量必须正交。（ ）

三、简答题

13 请简要列举几种可以用于身份识别的生物特征，并分析其特点

14 给出 PCA 中方差最大化和投影损失最大化的等价性证明。

13. 指纹特征：性能高、采集成本低、采集成本中等

人脸特征：——高 ——中等 ——低

基因 ：——高 ——高 ——高

投影方差的目标函数

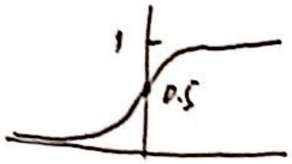
$$J_1(b_1)=E(y_1^2)=E\{(b_1^Tx)^2\}=b_1^TE\{xx^T\}b_1=b_1^TC_xb_1.$$

投影损失的目标函数

$$J_2(b_1)=E\{\|x-y_1b_1\|^2\}=C_x-b_1^TC_xb_1$$

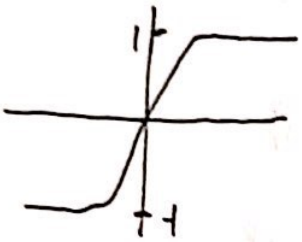
由于 C_x 是一个常数 $b^*=\arg\max\limits_{b_1} J_1(b_1)=\arg\min\limits_{b_1} J_2(b_1)\ s.t.\ \|b_1\|=1.$

sigmoid 激活函数.

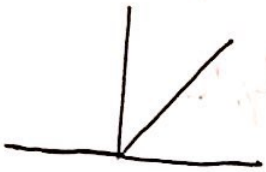


$f(z)=\frac{1}{1+e^{-z}}$

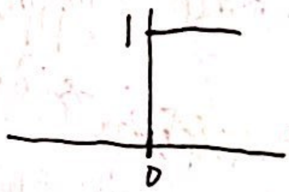
$\tanh(x)=\frac{e^x-e^{-x}}{e^x+e^{-x}}$



$Relu=\max(0,x)$



$df(x)/dx$



第九章 神经网络

一、填空题

1 单层感知器无法解决的代表性问题是（线性不可分问题

2 从传统神经网络向深度学习转变过程中，具有里程碑式意义的新激活函数是（ relu ）函数。

3 ReLU 函数在$x=1$处的梯度值为（ 1 ）。

二、选择题

4 以下函数中不是常见的激活函数的有 A.B.D.

A. $f(x) = x$	B. $f(x) = \text{sign}(x)$
C. $f(x) = \tanh(x)$	D. $f(x) = \sin(x)$

三、简答题

5 请从集成学习的角度解释 MLP（多层感知机）为何能解决 XOR（异或）问题。

单层感知机无法解决非线性问题。

6 请简述传统 BP 网络的瓶颈问题及成因。

7 试分析等激活函数为线性函数时，无论网络多深其总体效果均为线性映射。

8 请在下图中标出输入门、输出门和遗忘门，并对其功能加以简要说明

输入门：决定让多少新的信息加入cell状态.

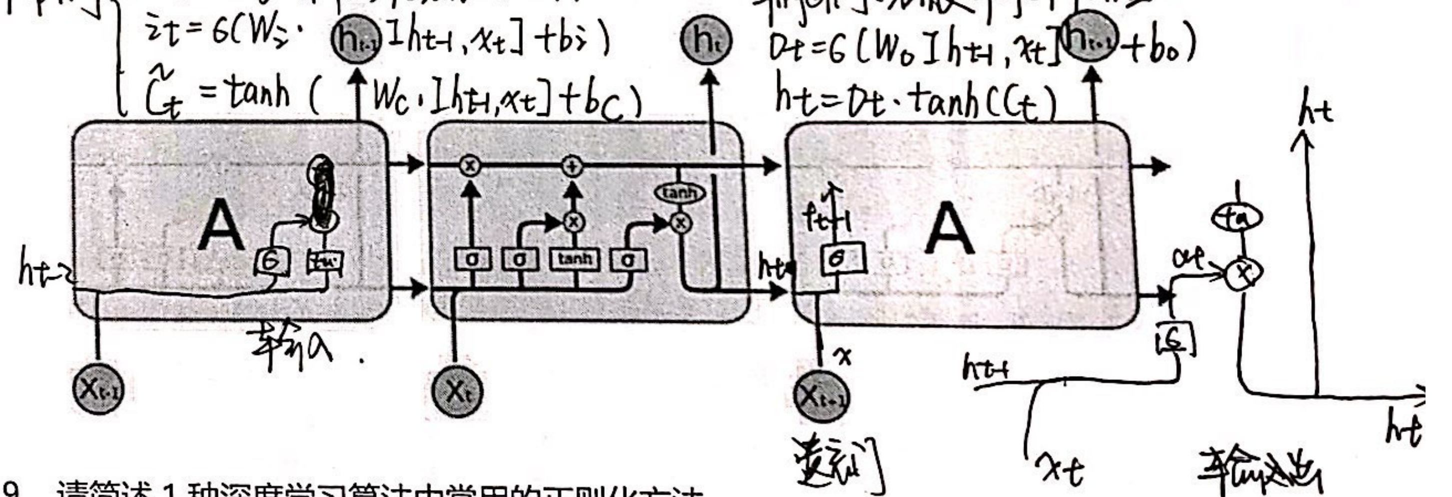
$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$

$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$

输出门：确定输出什么值.

$o_t = \sigma(W_o [h_{t-1}, x_t] + b_o)$

$h_t = o_t \cdot \tanh(C_t)$



遗忘门：决定从细胞状态丢弃信息.

$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$

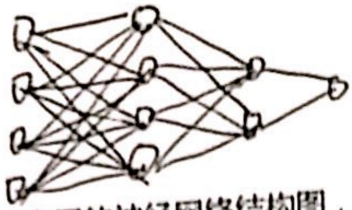
1表示完全保留

0表示完全舍弃

9 请简述 1 种深度学习算法中常用的正则化方法

四、计算（画图）题

正则化.



10 手绘一副包含输入层、隐层和输出层的神经网络结构图，用于解决 Iris 数据库的分类问题。设训练样本标记为$\{x_i \in \Re^4, y_i \in \{+1,-1\} | i = 1,\ldots,150\}$,请给出权重、信号、激活函数与损失函数的符号表达与计算公式。

11 请结合上题给出的网络结构和损失函数及相关符号表达，采用 BP 算法，写出损失函数对隐藏第 2 个神经元连接输出层第 1 个神经元的权重的导数的数学表达式。

12 对于一个 5 输入的 MP 神经元模型，采用 Logistic Sigmoidal 激活函数，对应连接权值为[0.2,0.3,−0.1,−0.5,0.9]，输入样本为[0.2,0.5,−0.3,−0,7,0.6]，偏置为-1，请计算该 MP 神经元模型的输出，并给出计算过程。

$f(x)=\frac{1}{1+e^{-0.47}}$

$y = x^T \cdot w = [0.2, 0.5, -0.3, -0.7, 0.6]\begin{bmatrix}0.2\\0.3\\-0.1\\-0.5\\0.9\end{bmatrix} + (-1) = 0.47$

13 在下图的神经网络中，假设输入$x_1 = 1$，$x_2 = 0.5$，网络的权值分别为：$w_1 = 1$，$w_2 = 2$，$w_3 = 3$，$w_4 = 4$，$w_5 = 0.5$，$x_6 = 0.6$ (w_6)，激活函数使用的是 sigmoid 核，计算网络前向传播之后的输出 y

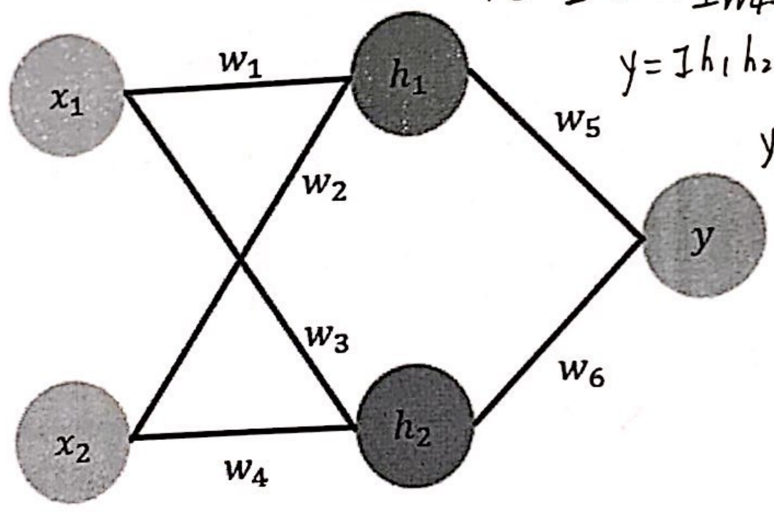
$f(x)=\frac{1}{1+e^{x}}$

$h_1=[x_1\ x_2]\begin{bmatrix}w_1\\w_2\end{bmatrix}=[1\ 0.5]\begin{bmatrix}1\\2\end{bmatrix}=2.$

$h_2=[x_1\ x_2]\begin{bmatrix}w_3\\w_4\end{bmatrix}=[1\ 0.5]\begin{bmatrix}3\\4\end{bmatrix}=5$

$y=[h_1\ h_2]\begin{bmatrix}w_5\\w_6\end{bmatrix}=[2\ 5]\begin{bmatrix}0.5\\0.6\end{bmatrix}=4$

$y=\frac{1}{1+e^{-4}}$



14 假设$f(z) = z^2$，$z = y^3 + 2y^2$，$y = 3x + 1$，使用链式求导法则求$\frac{df(z)}{x}$

$\frac{df(z)}{x} = \frac{df(z)}{dz}\cdot\frac{dz}{dy}\cdot\frac{dy}{dx} = 2z\cdot(3y^2+4y)\times 3 = 6z(3y^2+4y)$

15 假设当前单通道图像可以用 1 个 4×4 大小的矩阵 X 表示，两个不同的卷积核分别为 conv1，conv2，请计算经过卷积（stride=1，边界用 0 补齐）和

max pooling（2×2）后的结果，给出计算过程。

X:

-2	4	-1	-1
-2	-2	4	-1
-2	-2	4	4
-2	-2	4	-2

Conv1:

1	-0.5	-0.5
-0.5	1	-0.5
-0.5	-0.5	1

Conv2:

0.5	0.5	-1
0.5	0.5	-1
0.5	0.5	-1

stride 为1.

$$\begin{matrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -2 & 4 & -1 & -1 & 0 \\ 0 & -2 & -2 & 4 & -1 & 0 \\ 0 & -2 & -2 & 4 & 4 & 0 \\ 0 & -2 & -2 & 4 & -2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

$$X \cdot conv1 = \begin{bmatrix} -5 & 11.5 & -4.5 & -2 \\ -3 & 0.5 & 11.5 & -7.5 \\ 7 & 0 & -3.85 & 5.5 \\ 1 & 6 & 0 & -2 \end{bmatrix}$$

$$\cancel{X \cdot conv1 \cdot pool = \begin{bmatrix} 11.5 & 11.5 \\ 7 & 5.5 \end{bmatrix}}$$

$$X \cdot conv2 = \begin{bmatrix} -4 & -4 & 0.5 & 0.5 \\ -3 & -10 & 1.5 & 3.5 \\ 3 & 6 & 2 & 6.5 \\ 2 & 4 & 0 & 7 \end{bmatrix}$$

$$\cancel{X \cdot conv2 \cdot pool = \begin{bmatrix} -3 & 3.5 \\ 6 & 7 \end{bmatrix}}$$

$$X \cdot conv1 \cdot pool = \begin{bmatrix} 11.5 & 11.5 & 11.5 \\ 7 & 11.5 & 11.5 \\ 7 & 0 & 5.5 \end{bmatrix}.$$