

1-机器学习概述

一、单选题

- 关于 L1 正则和 L2 正则 下面的说法正确的是 **B**
 - L2 范数可以防止过拟合，提升模型的泛化能力。但 L1 正则做不到这一点
 - L2 正则化标识各个参数的平方的和的开方值
 - L2 正则化有个名称叫“Lasso regularization”
 - L1 范数会使权值平滑
- 下列有关机器学习中 L1 正则化和 L2 正则化说法正确的是? **A D**
 - 使用 L1 可以得到稀疏的权值
 - 使用 L2 可以得到稀疏的权值
 - 使用 L1 可以得到平滑的权值
 - 使用 L2 可以得到平滑的权值
- 以下关于正则化的描述错误的是 **D**
 - 正则化可以防止过拟合
 - L1 正则化能得到稀疏解
 - L2 正则化约束了解空间
 - Dropout 不是一种正则化方法
- 下列模型属于生成模型的是 **D**
 - SVM
 - 逻辑回归
 - CRF
 - HMM
- 在机器学习中，L1 正则化和 L2 正则化的引入为了解决什么问题 **C**
 - 数据量不充分
 - 训练数据不匹配
 - 训练过拟合
 - 训练速度太慢

二、填空题

- 向量 $x=[1,2,3,4,-9,0]$ 的 L1 范数是 **19** **有监督 无监督**
- 根据是否需要标注数据，机器学习方法可以分为 监督 学习和 无监督 学习
- 监督学习中的 训练集，用于估算模型
- 过拟合 是由于模型本身不能对训练集进行拟合或者训练迭代次数太少
- 正则化 是一种抑制模型复杂度的常用方法

三、计算题

- 测试集中 1000 个样本，600 个是 A 类，400 个 B 类，模型预测结果 700 个判断为 A 类，其中正确的有 500 个，300 个判断为 B 类，其中正确的有 200 个。请计算 B 类的精确率(Precision)和召回率(Recall). **$200/300$ $200/400$**
- 预测 20 个西瓜中哪些是好瓜，这 20 个西瓜中实际有 15 个好瓜，5 个坏瓜。某个模型

预测的结果是: 16 个好瓜, 4 个坏瓜。其中, 预测的 16 个好瓜中有 14 个确实是好瓜, 预测的 4 个坏瓜中有 3 个确实是坏瓜。(注:好瓜为正例, 坏瓜为反例。)

- (1) 画出混淆矩阵 (4 分)
- (2) 什么是精确率(Precision) , 计算准确率 P (3 分)
- (3) 什么是召回率(Recall), 计算召回率 R (3 分)

答案:

一、单选题

1. B
2. A D
3. D
4. D
5. C

二、填空题

1. 19
2. 有监督 无监督
3. 训练
4. 欠拟合
5. 正则化

三、计算题

1. 精确率: $200/300=0.66$;召回率: $200/400=0.5$
- 2.

(1)

		预测值	
		0	1
真实值	0	3	2
	1	1	14

- (2) 精确率表示被预测为正的样本实际也为正的样本概率。
 $P=TP/(TP+FP)=87.5\%$
- (3) 召回率表示实际为正的样本中被预测为正样本的概率。
 $R=TP/(TP+FN)=93.3\%$

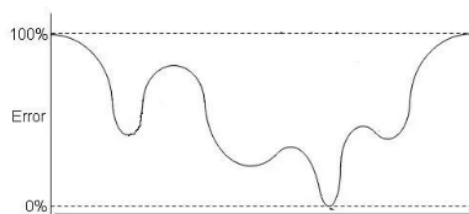
2-逻辑回归与最大熵

一、选择题

1. 以下关于逻辑回归说法错误的是:
- A. 特征归一化有助于模型效果
- B. 逻辑回归是一种广义线性模型



- C. 逻辑回归相比最小二乘法分类器对异常值更敏感
- D. 逻辑回归可以看成是只有输入层和输出层且输出层为单一神经元的神经网络
2. 以下关于逻辑回归的说法不正确的是？ **C**
- A. 逻辑回归必须对缺失值做预处理；
- B. 逻辑回归要求自变量和目标变量是线性关系；
- C. 逻辑回归比决策树，更容易过度拟合；
- D. 逻辑回归只能做 2 值分类，不能直接做多值分类；
3. 下列哪种方法在逻辑回归上最适合数据？ **B**
- A. 最小二乘法误差
- B. 极大似然估计
- C. 杰卡德距离
- D. A 和 B
4. 一个分类器在测试集上的混淆矩阵如下表所示，该分类器对类别 3 的召回率为 **A.**
- A. 80%
- B. 70%
- C. 60%
- D. 50%
5. 假设，下图是逻辑回归的代价函数，图中有多少个局部最小值 **D**



- A. 1
- B. 2
- C. 3
- D. 4

二、填空题

1. 使用逻辑回归算法对样本进行分类，得到训练样本的准确率和测试样本的准确率。现在，在数据中增加一个新的特征，其它特征保持不变。然后重新训练测试，那么训练样本的准确率会 **↑** 或者不变
2. 有数据集 正样本 120 个，负样本 80 个，模型 F 对样本进行预估 预测为正样本的有 80 个（其中真的是正样本的是 60 个），请问该模型的召回率是 **50%**
3. 计算下列混淆矩阵的 真正例率 TPR: **0.9**

		真实值	
		真	假
预测值	真	110	12
	假	31	501

4. 逻辑回归通过极大似然估计进行参数学习，其目标函数等价于最小化

负对数似然函数。

5. 当随机变量呈__分布时，熵值最大

答案：

均匀。

一、选择题

1. C
2. C
3. B
4. A
5. D

二、填空题

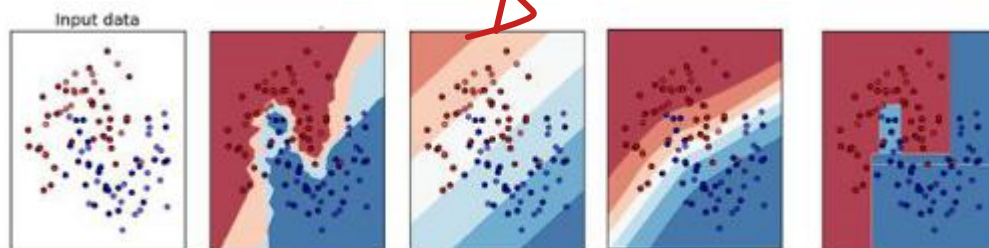
1. 增加
2. 50%
3. 0.9
4. 负对数似然函数（或：交叉熵损失）
5. 均匀

3-k 邻近算法

一、选择题

1. 下面关于 KNN 算法说法正确是（ C ）。

- A. KNN 算法的时间复杂度是 $O(n*k*t)$ ，其中 k 为类别数， t 为迭代次数
 - B. KNN 算法是一种非监督学习算法
 - C. 使用 KNN 算法进行训练时，训练数据集中含有标签
 - D. K 值确定后，使用 KNN 算法进行样本训练时，每次所形成的结果可能不同
2. 以下哪个图是 KNN 算法的训练边界



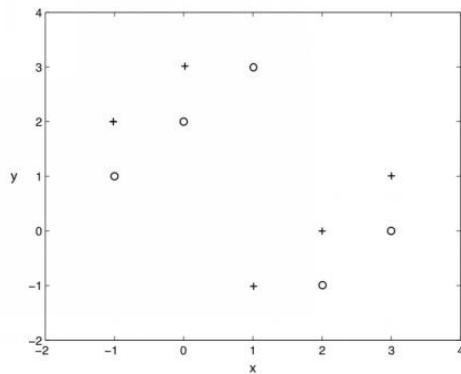
A. B

- B. A
- C. D
- D. C

3. 以下关于 KNN 的描述，不正确的是 (A).

- A. KNN 算法只适用于数值型的数据分类。
- B. KNN 算法对异常值不敏感。
- C. KNN 算法无数据输入假定。
- D. 其他说法都正确

4. 使用 $k=1$ 的 knn 算法，下图二类分类问题，“+” 和 “o” 分别代表两个类，那么，用仅拿出一个测试样本的交叉验证方法，交叉验证的错误率是多少： B



- A. 0%
 - B. 100%
 - C. 0%到 100
 - D. 以上都不是
5. 一般情况下，KNN 最近邻方法在 (D) 情况下效果最好
- A. 样本呈现团状分布
 - B. 样本呈现链状分布
 - C. 样本较多但典型性不好
 - D. 样本较少但典型性好

二、填空题

1. 在 K 近邻(KNN)模型中，超参数 K 的选择对模型的表现有较大的影响。一般而言，对比 1 近邻模型和 3 近邻模型，1 近邻模型的 Bias 更 ↓ Variance 更 ↑ .
2. KNN 是 有监督 学习算法
3. KNN 算法对 异常值 不敏感
4. KNN 算法可以用来解决 回归 问题
5. KNN 中若 $k=1$ ，样本 x 的预测结果只与由训练集中与其 距离最近 的那个样本决定

答案：

一、选择题

- 1. C
- 2. B
- 3. A
- 4. B

5. D

二、填空题

1. 小 大
2. 有监督
3. 异常值
4. 回归
5. 距离最近

4-决策树

一、选择题

1. 哪项不是决策树中属性选择的方法?

A

- A. 信息值
- B. 信息增益
- C. 信息增益率
- D. GINI 系数

2. 以下哪种方式通常不能帮助解决决策树过拟合:

D

- A. 限制最大树深度
- B. 后剪枝
- C. 样本抽样
- D. 增加新特征

3. 对于下面的超参数来说, 更高的值对于决策树算法更好吗?

D

1 用于拆分的样本量 2 树深 3 树叶样本

- A. 1 和 2
- B. 2 和 3
- C. 1 和 3
- D. 无法分辨

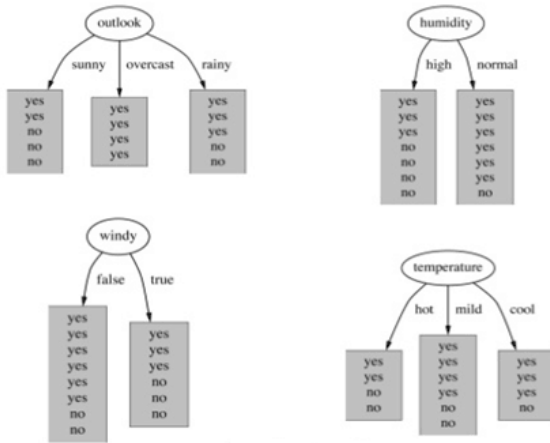
4. 我们想在大数据集上训练决策树, 为了使用较少时间, 我们可以:

B

- A. 增加树的深度
- B. 减少树的深度
- C. 增加学习率 (learning rate)
- D. 减少树的数量

5. 训练决策树模型, 属性节点的分裂, 具有最大信息增益的图是下图的哪一个

A .



- Outlook
- Humidity
- Windy
- Temperature

二、填空题

- 决策树学习算法中对付“过拟合”的主要手段是 **剪枝**。
- 决策树算法 **不能** 解决回归问题。
- 决策树中父节点的熵 **大于** 子节点的熵。
- 基尼指数越大样本的不确定性越 **大**。
- 逻辑回归分析需要对离散值做预处理，决策树 **不需要** 对离散值做预处理。

三、计算题

- 按照颜色这一特征划分计算信息增益和信息增益

	1	2	3	4	5	6	7	8
瓜色	深绿	深绿	深绿	深绿	深绿	浅绿	浅绿	浅绿
标签	生瓜	生瓜	生瓜	生瓜	生瓜	熟瓜	熟瓜	熟瓜

- 考虑题表中数据集（前 8 条为训练集，后 2 条测试集），并回答以下问题。

Ca+浓度	Mg+浓度	Na+浓度	Cl-浓度	类
低	高	高	高	冰川水
高	低	高	高	冰川水
低	高	低	低	冰川水
高	高	低	低	冰川水
低	低	低	低	湖泊水
高	低	低	低	湖泊水
低	高	高	低	湖泊水
高	低	高	低	湖泊水
低	高	高	低	①
高	高	低	高	②

- 整个训练集关于类属性的熵是多少？
- 训练集中属性 **Ca+浓度** 的信息增益是多少？

(3) 使用 ID3 决策树算法对两个未知类型的样本进行分类。

答案:

一、选择题

1. A
2. D
3. D
4. B
5. A

二、填空题

1. 剪枝
2. CART
3. 大于
4. 大
5. 不需要

三、计算题

$$1. H(D) = -\left(\frac{5}{8}\log_2 \frac{5}{8} + \frac{3}{8}\log_2 \frac{3}{8}\right) = 0.954375;$$

$$H(D|A) = -\left(\frac{3}{8}\left(\frac{2}{3}\log_2 \frac{2}{3} + \frac{1}{3}\log_2 \frac{1}{3}\right) + 0 + 0\right) = 0.344375;$$

$$g(D,A) = H(D) - H(D|A) = 0.61; \quad g_R(D,A) = g(D,A)/H(D) = 0.639.$$

2.

$$(1) \text{info}(D) = -4/8 \times \log(4/8) - 4/8 \times \log(4/8) = 1$$

$$(2) \text{info}_{Ca^+}(D_{\text{高}}) = -2/4 \times \log(2/4) - 2/4 \times \log(2/4) = 1$$

$$\text{info}_{Ca^+}(D_{\text{低}}) = -2/4 \times \log(2/4) - 2/4 \times \log(2/4) = 1$$

$$\text{info}_{Ca^+}(D) = 4/8 \times \text{info}_{Ca^+}(D_{\text{高}}) + 4/8 \times \text{info}_{Ca^+}(D_{\text{低}}) = 1$$

$$\text{Gain}(Ca^+ \text{浓度}) = \text{info}(D) - \text{info}_{Ca^+}(D) = 0$$

(3) ①湖泊水 ②冰川水

首先计算各属性的信息增益

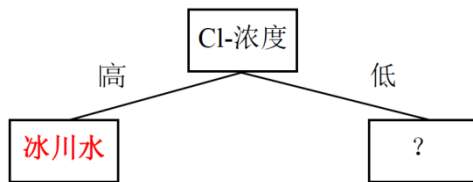
$$\text{Gain}(Ca^+ \text{浓度}) = 0$$

$$\text{Gain}(Mg^+ \text{浓度}) = 0.185$$

$$\text{Gain}(Na^+ \text{浓度}) = 0$$

$$\text{Gain}(Cl^- \text{浓度}) = 0.32$$

选择 Cl^- 浓度作为根节点



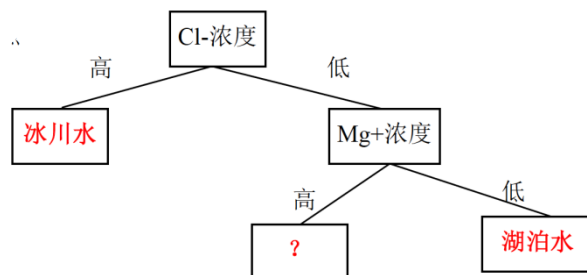
计算各属性的信息增益

$\text{Gain}(\text{Ca+浓度})=0$

$\text{Gain}(\text{Mg+浓度})=0.45$

$\text{Gain}(\text{Na+浓度})=0.24$

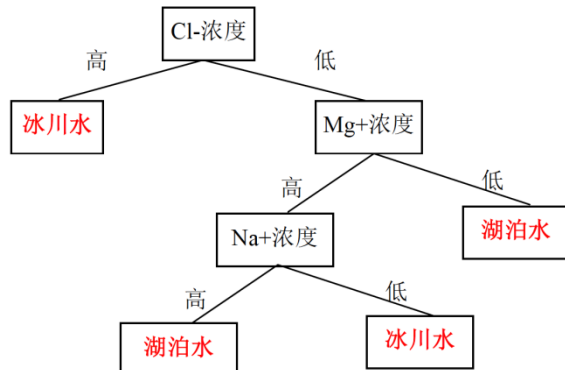
选择 **Mg+浓度** 作为节点



计算各属性的信息增益

$\text{Gain}(\text{Ca+浓度})=0.24$

$\text{Gain}(\text{Na+浓度})=0.91$



故①是湖泊水，②是冰川水。

5-朴素贝叶斯分类器

一、选择题

- 朴素贝叶斯分类器的训练过程是基于训练集 D 来估计 () **A.**
 - 先验概率
 - 后验概率
 - 概率分布函数
 - 概率密度函数

2. 下列哪种情况不能用朴素贝叶斯分类器 ()
- 训练数据集较大
 - 实例具有几个属性
 - 给定分类参数，描述实例的属性应该是条件独立的
 - 要求有较高的分类精度
3. 贝叶斯分类器的训练中，最大似然法估计参数的过程包括以下哪些步骤 ()
- 求导数，令偏导数为 0，得到似然方程组
 - 对似然函数取对数，并整理
 - 解似然方程组，得到所有参数即为所求
 - 以上所有
4. (朴素贝叶斯) 是一种特殊的 Bayes 分类器，特征变量是 X ，类别标签是 Y ，它的一个假定是 ()
- 各类别的先验概率 $P(Y)$ 是相等的
 - 特征变量 X 的各个维度是类别条件独立随机变量
 - 以 0 为均值， $\text{sqr}(2)/2$ 为标准差的正态分布
 - $P(X|Y)$ 是高斯分布
5. 下面的表格中列出了 14 个日期中天气、温度、湿度和风力四个因素和小明是否攀岩的关系。基于这 14 个观测数据，采用朴素贝叶斯分类方法计算出实例 $\langle \text{天气}=\text{晴天}, \text{温度}=\text{凉爽}, \text{湿度}=\text{高}, \text{风力}=\text{强} \rangle$ 时“休息”的概率为

日期	天气	温度	湿度	风力	攀岩
D1	晴天	热	高	强	休息
D2	晴天	热	高	强	休息
D3	阴天	热	高	弱	攀岩
D4	下雨	温和	高	弱	攀岩
D5	下雨	凉爽	正常	弱	攀岩
D6	下雨	凉爽	正常	强	休息
D7	阴天	凉爽	正常	强	攀岩
D8	晴天	温和	高	弱	休息
D9	晴天	凉爽	正常	弱	攀岩
D10	下雨	温和	正常	弱	攀岩
D11	晴天	温和	正常	强	攀岩
D12	阴天	温和	高	强	攀岩
D13	阴天	热	正常	弱	攀岩
D14	下雨	温和	高	强	休息

- 0.0795
- 0.0205
- 0.64
- 0.33

二、填空题

- 朴素贝叶斯分类算法假设属性之间相互
- 贝叶斯分类器可以解决__学习的问题
- 贝叶斯分类器是基于__概率，推导出__概率
- 假定某同学使用贝叶斯分类模型时，由于失误操作，致使训练数据中两个维度重复表示，那么模型效果精度会
- 贝叶斯定理中，如果描述随机事件 A 和 B 的条件概率的定理，表达式是__

三、计算题

- 已知样本的属性和标签如下表所示，当某样本属性为 (a_2, b_2, c_2) 时，采用朴素贝叶斯方法，非归一化的 $P(L_3|a_2, b_2, c_2)P(L_3|a_2, b_2, c_2)$ 值为

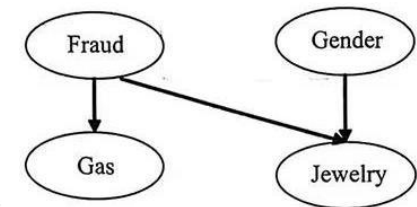
属性 1	属性 2	属性 3	标签
------	------	------	----

a2	b1	c3	L2
a1	b1	c2	L3
a1	b1	c1	L1
a3	b3	c1	L3
a1	b3	c2	L3
a3	b1	c3	L1
a2	b2	c1	L3
a1	b2	c1	L3
a2	b3	c3	L3
a2	b2	c3	L1

2、题表 2-1 中给出的数据集对应的贝叶斯信念网如题图 2-2 所示 (假设所有属性都是二元的)。

题表 2-1

ID	Fraud	Gas	Jewery	Gender
1	No	No	No	Female
2	No	No	No	Male
3	Yes	Yes	Yes	Male
4	No	No	No	Male
5	No	Yes	No	Female
6	No	No	No	Female
7	No	No	No	Male
8	Yes	No	Yes	Female
9	No	Yes	No	Male
10	No	No	No	Female



题图 2-2

- (1) 画出网络中每个节点对应的概率表。
- (2) 使用贝叶斯信念网计算 $P(\text{Gas}=\text{No}, \text{Jewery}=\text{No})$ 。
- (3) 使用贝叶斯信念网计算 $P(\text{Jewery} = \text{Yes} | \text{Fraud}=\text{No} , \text{Gender}=\text{Female})$ 。

3、考虑题表中数据集，并回答以下问题。

记录	A	B	类
1	0	0	F
2	1	0	T
3	0	1	F
4	1	0	F
5	1	0	T
6	0	0	T
7	1	1	F
8	0	0	F
9	0	1	T
10	1	1	T

- (1) 估计条件概率 $P(A=1 | T), P(A=1 | F)$

- (2) 根据 (1) 中的条件概率，使用朴素贝叶斯方法预测测试样本 ($A=1$ 、 $B=1$) 的类标号
 (3) 计算属性 A 的信息增益

答案：

一、选择题

1. A
2. D
3. D
4. B
5. A

二、填空题

1. 独立
2. 有监督
3. 先验 后验
4. 降低
5. $P(A|B) = P(B|A)P(A) / P(B)$

三、计算题

1. 对于非归一化的朴素贝叶斯算法，
 $P(L3|a2, b2, c2) = P(L3) * P(a2|L3) * P(b2|L3) * P(c2|L3) = (3/5) * (1/3) * (1/3) * (1/3) = 1/45$
- 2.

(1)

Fraud:

Yes	No
0.2	0.8

Gender:

Female	Male
0.5	0.5

Gas:

	Gas	
Fraud	Yes	No
Yes	0.5	0.5
No	0.25	0.75

Jewelry:

		Jewelry
--	--	---------

Fraud	Gender	Yes	No
Yes	Female	1	0
Yes	Male	1	0
No	Female	0	1
No	Male	0	1

(2)

$$\begin{aligned} P(\text{Gas} = \text{No}, \text{Jewery} = \text{No}) &= P(\text{Gas} = \text{No}, \text{Jewery} = \text{No} | \text{Fraud} = \text{Yes}, \text{Gender} = \text{Female}) \\ &\quad * P(\text{Fraud} = \text{Yes}) * P(\text{Gender} = \text{Female}) \\ &\quad + P(\text{Gas} = \text{No}, \text{Jewery} = \text{No} | \text{Fraud} = \text{Yes}, \text{Gender} = \text{Male}) \\ &\quad * P(\text{Fraud} = \text{Yes}) * P(\text{Gender} = \text{Male}) \\ &\quad + P(\text{Gas} = \text{No}, \text{Jewery} = \text{No} | \text{Fraud} = \text{No}, \text{Gender} = \text{Female}) \\ &\quad * P(\text{Fraud} = \text{No}) * P(\text{Gender} = \text{Female}) \\ &\quad + P(\text{Gas} = \text{No}, \text{Jewery} = \text{No} | \text{Fraud} = \text{No}, \text{Gender} = \text{Male}) \\ &\quad * P(\text{Fraud} = \text{No}) * P(\text{Gender} = \text{Male}) \\ &= 0 + 0 + \frac{3}{4} * \frac{4}{5} * \frac{1}{2} + \frac{3}{4} * \frac{4}{5} * \frac{1}{2} = 0.6 \end{aligned}$$

(3)

$$P(\text{Jewery} = \text{Yes} | \text{Fraud} = \text{No}, \text{Gender} = \text{Female}) = 0$$

$$3. (1) P(A=1|T)=3/5, P(A=1|F)=2/5$$

$$(2) P(T|A=1, B=1) = P(A=1|T) \cdot P(B=1|T) \cdot P(T)$$

$$= (3/5) * (2/5) * (5/10) = 3/25 = 0.12$$

$$P(F|A=1, B=1) = P(A=1|F) \cdot P(B=1|F) \cdot P(F)$$

$$= (2/5) * (2/5) * (5/10) = 2/25 = 0.08$$

$$P(T|A=1, B=1) > P(F|A=1, B=1)$$

因此预测样本的类标号为 T。

$$(3) \text{info}(D) = -1/2 \times \log(1/2) - 1/2 \times \log(1/82) = 1$$

$$\text{info}_A(D_0) = -3/5 \times \log(3/5) - 2/5 \times \log(2/5) = 0.673$$

$$\text{Gain}(A) = \text{info}(D) - \text{info}_A(D_0) = 0.327$$

6-支持向量机

一、选择题

1. 对于给定 1000 个训练样本的二分类问题，关于支持向量机的说法，正确的有 **A** .
- A. 需要构造 1000 个辅助变量，计算它们的非零值对应着支撑向量。
 - B. 如果使用高斯核函数，不需要构造 1000 个辅助变量，只需要 100 个。
 - C. 如果使用多项式核函数，不需要构造 1000 个辅助变量，只需要 100 个。
 - D. 在当前普通计算机上需要约 1 小时才能得到训练模型。

2. 对于 K 类线性分类问题，下列哪种 SVM 分类策略可以避免“投票机制”（ C ）。

- A. 一对其余
- B. 一对一
- C. 逐步二分类
- D. 直接多类 SVM 分类

3. SVM（支持向量机）与 LR（逻辑回归）的数学本质上的区别是什么？（ A ）。

- A. 损失函数
- B. 是否有核技巧
- C. 是否支持多分类
- D. 其余选项皆错

4. 下列有关 SVM 说法不正确的是：（ D ）。

- A. SVM 使用核函数的过程实质是进行特征转换的过程
- B. SVM 对线性不可分的数据有较好的分类性能
- C. SVM 因为使用了核函数，因此它没有过拟合的风险
- D. SVM 的支持向量是少数的几个数据点向量

5. 下列哪项不是 SVM 的优势（ C ）。

- A. 可以和核函数结合
- B. 通过调参可以往往可以得到很好的分类效果
- C. 训练速度快
- D. 泛化能力好

二、填空题

1. 支持向量机中当参数 C 越 小 时，分类间隔越大，分类错误越多，趋于欠学习

2. 支持向量机是一种 有监督 学习方法

3. LR 是分类模型与判别式模型，SVM 是 支持向量 模型与 判别式 模型

4. SVM 分类的依据是 支持向量

5. 核函数 可以使得 SVM 实现非线性分类

答案：

一、选择题

- 1. A
- 2. C
- 3. A
- 4. C
- 5. C

二、填空题

- 1. 小
- 2. 有监督
- 3. 分类 判别式
- 4. 支持向量
- 5. 核函数

7-集成学习

一、选择题

1. 在 Bagging 集成学习中，多样性是通过 () 实现的。

- A. 数据样本扰动
- B. 输入属性扰动
- C. 输出表示扰动
- D. 算法参数扰动

2. 以下哪些算法未使用集成学习的思想

- A. 随机森林
- B. GBDT
- C. SVM
- D. AdaBoost

3. 数据科学家经常使用多个算法进行预测，并将多个机器学习算法的输出（称为“集成学习”）结合起来，以获得比所有个体模型都更好的更健壮的输出。则下列说法正确的是？

- A. 基本模型之间相关性高
- B. 基本模型之间相关性低
- C. 集成方法中，使用加权平均代替投票方法
- D. 基本模型都来自于同一算法

4. 关于随机森林的训练过程下列描述正确的是：

- A. 样本扰动
- B. 属性扰动
- C. 样本扰动并且属性扰动
- D. 不存在扰动现象

5. 下列关于随机森林和 GBDT 的说法正确的是：

- A. 都是一种 Boosting 方法
- B. 组成随机森林的树可以分类树也可以是回归树，而 GBDT 只由回归树组成
- C. 随机森林的训练属于属性扰动，不属于样本扰动
- D. 随机森林对异常值敏感，而 GBDT 对异常值不敏感

二、填空题

1. Adaboost 相对于单个弱分类器而言通过 Boosting 了模型的方差，了偏差

2. 随机森林相对于单个决策树而言通过 Bagging 了模型的 Variance

3. 为了防止过拟合，随机森林应该对每一颗子树进行

4. Xgboost 模型的损失函数由__得到

5. 在 AdaBoost 算法中，所有被分错的样本的权重更新比例

答案：

一、选择题

- 1. A
- 2. C
- 3. B
- 4. C

5. B

二、填空题

1. 增大 降低
2. 降低
3. 剪枝
4. 泰勒展开
5. 相同

8-降维

一、选择题

1. 对于 PCA 说法错误的是: **A.**

- A. 变量表示成各公因子的线性组合
- B. 可以解决多重共线性
- C. 业务不可解释性
- D. 基于主元方向的样本方差最大

2. 数字图像处理中常使用主成分分析 (PCA) 来对数据进行降维, 下列关于 PCA 算法错误的是: **C**

- A. PCA 算法是用较少数量的特征对样本进行描述以达到降低特征空间维数的方法;
- B. PCA 本质是 KL-变换;
- C. PCA 是最小绝对值误差意义下的最优正交变换;
- D. PCA 算法通过对协方差矩阵做特征分解获得最优投影子空间, 来消除模式特征之间的相关性、突出差异性;

3. 应用 PCA 后, 以下哪项可以是前两个主成分? (D) **D.**

- (1). (0.5, 0.5, 0.5, 0.5) 和 (0.71, 0.71, 0, 0)
- (2). (0.5, 0.5, 0.5, 0.5) 和 (0, 0, -0.71, 0.71)
- (3). (0.5, 0.5, 0.5, 0.5) 和 (0.5, 0.5, -0.5, -0.5)
- (4). (0.5, 0.5, 0.5, 0.5) 和 (-0.5, -0.5, 0.5, 0.5)

- A. 1 和 2
- B. 1 和 3
- C. 2 和 4
- D. 3 和 4

4. 为了得到和 SVD 一样的投射 (projection), 你需要在 PCA 中怎样做? **A.**

- A. 将数据转换成零均值
- B. 将数据转换成零中位数
- C. 无法做到
- D. 以上方法不行

5. 主成分分析 PCA 和数据矩阵的奇异值分解 SVD **B**

- A. 完全等价, 可以得到相同的特征空间
- B. 数据均值归零后, SVD 与 PCA 等价, 可以得到相同的特征空间
- C. 无法得到不等价形式
- D. 以上都不对

二、填空题

1. PCA 主成分的最大数量__特征能数量

2. PCA 所有主成分彼此__

正交

3. 利用 PCA 将下列数据进行降维，得到的特征值为__

$\frac{22}{5}$

$$\begin{pmatrix} -2 & 0 & 0 & 1 & 1 \\ -1 & -1 & 0 & 2 & 0 \end{pmatrix}$$

4. PCA 中原始数据在第一主成分上的投影方差最__

大

5. SVD 经常作为特征降维的一种有效方法，对于以下四个样本， $x_1=\{6,6\}$, $x_2=\{0,1\}$, $x_3=\{4,0\}$, $x_4=\{0,6\}$ ，如果采用 SVD 的特征处理方式后，只保留最大特征值，则 SVD 后的样本向量的均方差误差为__

25

答案：

一、选择题

1. A

2. C

3. D

4. A

5. B

二、填空题

1. 小于等于

2. 正交

3. $2\frac{2}{5}$

4. 大

5. 25

9-聚类

一、选择题

1. 以下哪几种聚类算法在训练的时候需要设定聚类个数：

A.

A. K-means

B. AP 聚类

C. DBSCAN

D. 层次聚类

2. 在有监督学习中，我们如何使用聚类方法？

B

1 我们可以先创建聚类类别，然后在每个类别上用监督学习分别进行学习

2 我们可以使用聚类“类别 id”作为一个新的特征项，然后再用监督学习分别进行学习

3 在进行监督学习之前，我们不能新建聚类类别

4 我们不可以使用聚类“类别 id”作为一个新的特征项，然后再用监督学习分别进行学习

A. 2 和 4

B. 1 和 2

C. 3 和 4

D. 1 和 3

3. 下面哪种情况不会影响 K-means 聚类的效果? (B)

A. 数据点密度分布不均

B. 数据点呈圆形状分布

C. 数据中有异常点存在

D. 数据点呈非凸形状分布

4. 下列有关聚类说法正确的是 (C)

A. 聚类过程: 它找出描述并区分数据类或概念的模型(或函数), 以便能够使用模型预测类标记未知的对象类

B. 在聚类分析当中, 簇内的相似性越大, 簇间的差别越大, 聚类的效果就越差

C. 聚类分析可以看作是一种非监督的分类

D. K 均值是一种产生划分聚类的基于密度的聚类算法, 簇的个数由算法自动地确定

二、填空题

1. 请给出三种常见的聚类算法__、__、__

2. K-Means 算法常使用__距离作为其中心距离的度量

3. K-Means 算法__适合\不适合) 对以下螺旋分布的样本进行聚类

4. 马氏距离是不受__影响的欧式距离

答案:

一、选择题

1. A

2. B

3. B

4. C

二、填空题

1. K-means 聚类 EM 聚类 DBSCAN

2. 曼哈顿

3. 不适合

4. 量纲

10-神经网络

一、选择题

1. 对于神经网络的说法, 下面正确的是: (A)

1. 增加神经网络层数, 可能会增加测试数据集的分类错误率

2. 减少神经网络层数, 总是能减小测试数据集的分类错误率

3. 增加神经网络层数, 总是能减小训练数据集的分类错误率

- A. 1
 - B. 1 和 3
 - C. 1 和 2
 - D. 2
2. 神经网络模型是受人脑的结构启发发明的。神经网络模型由很多的神经元组成，每个神经元都接受输入，进行计算并输出结果，那么以下选项描述正确的是 **D.**
- A. 每个神经元只有一个单一的输入和单一的输出
 - B. 每个神经元有多个输入而只有一个单一的输出
 - C. 每个神经元只有一个单一的输入而有多输出
 - D. 每个神经元有多个输入和多个输出
3. 梯度下降算法中，哪个因素会影响到最终的效果？ **A.**
- A. 步长
 - B. 初始值
 - C. 样本集规模
 - D. 样本数据方差
4. 关于 CNN，以下说法错误的是 **B**
- A. CNN 可用于解决图像的分类及回归问题
 - B. CNN 最初是由 Hinton 教授提出的
 - C. CNN 是一种判别模型
 - D. 第一个经典 CNN 模型是 LeNet
5. 当在卷积神经网络中加入池化层(pooling layer)时，变换的不变性会被保留吗？ **C**
- A. 不能确定
 - B. 看情况
 - C. 是
 - D. 否

二、填空题

- 1. 如果增加多层感知机 (Multilayer Perceptron) 的隐藏层层数，分类误差会 **↓**
- 1. 使用 Tanh 作为激活函数可能会导致梯度 **消失**
- 2. 在 CNN 网络中，图 A 经过核为 3×3 ，步长为 2 的卷积层，ReLU 激活函数层，BN 层，以及一个步长为 2，核为 2×2 的池化层后，再经过一个 3×3 的卷积层，步长为 1，此时的感受野是 **13**
- 3. **RNN** 更适合对序列数据进行建模
- 4. GAN 中包含两个部分，分别是 **生成器** 和 **判别器**

三、问答题

- 1. 假设你有 5 个大小为 7×7 、边界值为 0 的卷积核，同时卷积神经网络第一层的步长为 1。此时如果你向这一层传入一个维度为 $224 \times 224 \times 3$ 的数据，那么神经网络下一层所接收到的数据维度是多少

答案：

一、选择题

- 1. A
- 2. D

- 3. A
- 4. B
- 5. C

二、填空题

- 1. 减小
- 2. 消失
- 3. 13
- 4. RNN
- 5. 生成器 判别器

三、问答题

- 1. 输出的大小为 $(W-F+2P)/S + 1$ ，输出的通道数为层内卷积核数量
其中 W 为输入， F 为卷积核， P 为 padding 值， S 为步长
因此结果为 $218 \times 218 \times 5$

11-EM 算法

一、选择题

- 1. EM 算法是 (**B**)
 - A. 有监督
 - B. 无监督
 - C. 半监督
 - D. 都不是
- 2. 关于 EM 算法的优缺点，下列说法错误的是 (**C**)
 - A. EM 算法简单而且稳定
 - B. EM 算法迭代速度受数据规模的影响
 - C. EM 算法的收敛速度，并不依赖于初始值
 - D. EM 算法求导函数找到的极值点不一定是最优解
- 3. 高斯混合模型(GMM)的极大似然估计(MLE) (**A**)
 - A. 只能得到非退化（协方差矩阵正定）的局部最优解
 - B. 存在非退化的全局最优解
 - C. 解情况依赖于数据分布
 - D. A 与 C 正确
- 4. 在 HMM 中,如果已知观察序列和产生观察序列的状态序列,那么可用以下哪种方法直接进行参数估计 (**D**)
 - A. EM 算法
 - B. 维特比算法
 - C. 前向后向算法
 - D. 极大似然估计
- 5. 如下数据集中，不适合使用隐马尔科夫模型(HMM)建模的有? (**B**)
 - A. 基因序列集合

- B. 电影影评数据集合
- C. 股票市场数据集合
- D. 北京气温数据集合

二、问答题

假设有 3 个盒子，编号为 1,2,3，每个盒子都装有红白两种颜色的小球，数目如下：

盒子号	1	2	3
红球数	5	4	7
白球数	5	6	3

然后按照下面的方法抽取小球，来得到球颜色的观测序列：

- 按照 $\pi=(0.2, 0.4, 0.4)$ 的概率选择 1 个盒子，从盒子随机抽出 1 个球，记录颜色后放回盒子；
- 按照下图 A 选择新的盒子，按照下图 B 抽取球，重复上述过程；

PS:

A 的第 i 行是选择到第 i 号盒子，第 j 列是转移到 j 号盒子，如：第一行第二列的 0.2 代表：上一次选择 1 号盒子后这次选择 2 号盒子的概率是 0.2。

B 的第 i 行是选择到第 i 号盒子，第 j 列是抽取到 j 号球，如：第二行第一列的 0.4 代表：选择了 2 号盒子后抽取红球的概率是 0.4。

$$\pi = \begin{pmatrix} 0.2 \\ 0.4 \\ 0.4 \end{pmatrix} \quad A = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{bmatrix} \quad B = \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \\ 0.7 & 0.3 \end{bmatrix}$$

求：得到观测序列“红白红”的概率是多少？

答案：

一、选择题

- B
- C
- A
- D
- B

二、问答题

- 根据前向算法，第一步计算 $\alpha_i(1)$ ，这很简单：

$\alpha_{i=1}(t=1)$ 即时刻 1 时位于状态“选中 1 号盒子”的前向概率，所以：

$$\alpha_1(1) = \text{“选中 1 号盒子”} * \text{“选中红球”} = \pi_0 * B_{10} = 0.2 * 0.5 = 0.1$$

$$\text{同理：} \alpha_2(1) = 0.4 * 0.4 = 0.16, \quad \alpha_3(1) = 0.4 * 0.7 = 0.28。$$

计算 $\alpha_i(2)$ ：

根据公式：

$$\begin{aligned} \alpha_1(2) &= (\sum_j \alpha_j(1) a_{j1}) b_{1y2} \\ &= (0.1 * 0.5 + 0.16 * 0.3 + 0.28 * 0.2) * 0.5 \\ &= 0.077 \end{aligned}$$

$$\alpha_2(2) = 0.1104$$

$$\alpha_3(2) = 0.0606$$

同 $\alpha_i(2)$ ，计算 $\alpha_i(3)$ ：

$$\alpha_1(3) = 0.04187$$

$$\alpha_2(3) = 0.03551$$

$$\alpha_3(3) = 0.05284$$

最终

$$\begin{aligned} P(0 \mid \lambda) &= \sum_i \alpha_i(3) \\ &= 0.04187 + 0.03551 + 0.05284 \\ &= 0.13022 \end{aligned}$$