Question 1

Not yet answered | Marked out of 1.00 | Flag question

Which of the following is true about K-means?

B

- A. K-means clustering aims to classify the objects into groups that with k objects in each group. 是分成k个簇，不是每个簇有k个样本
- B. K-means is guaranteed to converge to a local minimum. 属于硬EM算法，确保收敛，但不确保得到全局最优
- C. The initialization of the centroids has no affect on the final result of K-means. k-means对初始聚类中心敏感，故有k-means++改进算法
- D. K-means is guaranteed to converge to a global minimum.

目标函数的复杂程度直接影响EM算法是否能找到全局最优。
对初始参数选择敏感是因为不同的初始值可能导致算法收敛到不同的局部最优解。

Question 2

Not yet answered | Marked out of 1.00 | Flag question

Which of the following statement is NOT true?

D

- A. Naïve Bayes is a supervised learning method. 朴素贝叶斯是监督学习
- B. Naïve Bayes assumes that all the features in a dataset are independent. 贝叶斯假设每个条件（属性）之间相互独立
- C. Naïve Bayes assumes that all the features in a dataset are equally important. 也没上过权重
- D. Naïve Bayes can be only used for binary classification problems. 显然可以多分类

Question 3

Not yet answered | Marked out of 1.00 | Flag question

垃圾邮件分类

You've just finished training a decision tree for spam classification, and it is getting abnormally bad performance on both of your training and test sets. Suppose that your implementation has no bugs, so which of the following could be the problem?

决策树过浅，欠拟合　　训练集&&测试集性能低→欠拟合

A

- A. The decision trees are too shallow.
- B. The learning rate is too small. 决策树没有学习率
- C. The model is overfitting. 还没过拟合
- D. None of the above.

Question 4

Not yet answered | Marked out of 1.00 | Flag question

Which of the following statement is INCORRECT about Random Forest?

A. Random Forest can not handle binary features. 随机森林可以处理二元特征

B. Random Forest can be used for classification task. 随机森林可以用来分类

C. Random Forest can be used for regression task. 随机森林可以用来回归

D. Random Forest has multiple decision trees as base learning models.

随机森林由多个决策树组成,每个决策树都是一个弱学习器，最后投票决定。

Question 5

Not yet answered | Marked out of 1.00 | Flag question

Which of the following statements is INCORRECT about PCA?

A. We must standardize the data first. PCA之前应当先对数据标准化，消除特征规模不同导致的偏差

B. We should select the principal components which explain the highest variance. 选择能解释最大方差（特征值）的方向（特征向量）从而完成降维

C. PCA components are not always orthogonal. 特征向量总是两两正交的

D. We can use PCA for visualizing the data in lower dimensions. 可以使用PCA降维

Question 6

Not yet answered | Marked out of 1.00 | Flag question

Which is not the application of language processing?

（自然）语言处理的应用

比较模糊

A. Textual Entailment 文本蕴含，指两段文本有指向性关系，类似推理：如果一个人读了t能够推论h非常可能是真实的，那么t 蕴涵 h(t=>h)

B. Spelling Correction 拼写检查

C. Information Retrieval 信息/情报检索

D. Latent Space Learning 隐空间学习，学习数据的潜在模式

Question 7

Not yet answered | Marked out of 1.00 | Flag question

梯度下降法中使用不合适的学习率会导致

Inappropriate selection of learning rate value in gradient descent gives rise to:

D

- A. Local Minima 只能找到局部最小，找极值找不到最值是梯度下降的通病
- B. Oscillations ✓ 震荡，即无法收敛，在学习率过大的时候会产生
- C. Slow convergence ✓ 收敛速度慢，常在学习率过小时发生
- D. All of the above

Question 8

Not yet answered | Marked out of 1.00 | Flag question

特征脸（Eigenface）是指用于机器视觉领域中的人脸识别问题的一组特征向量，任意一张人脸图像都可以被认为是这些标准脸的组合

Which of the following techniques is "eigenfaces" build on? 原理上是用PCA做，那么SVD也能做

D

- A. Non-negative matrix factorization 非负矩阵分解
- B. Independent Component Analysis
- C. Singular Value Decomposition 奇异值分解，scikit-learn内部的PCA也是用SVD做的
- D. Support Vector Machine 完全不相干

NMF与PCA、ICA和SVD都是降维找特征的方法容易混淆，本题需要记概念！

Question 9

Not yet answered | Marked out of 1.00 | Flag question

零假设

The point where the Null Hypothesis gets rejected is called as?

b

- A. Rejection Value
- B. Critical Value 假设检验中，如果观察到的统计量大于或小于临界值（Critical Value）则认为该统计量具有显著性差异，从而拒绝零假设（Null Hypothesis）。
- C. Significant Value
- D. Acceptance Value

零假设可以是“这个药物对治疗癌症没有作用”；
于是有对立的备择假设（Alternative Hypothesis）如“这个药物对治疗癌症有一定的有效性”
然后构造一个小概率事件，并基于抽取的样本数据来检验这个小概率事件是否发生。
如果小概率事件发生了，则拒绝零假设；如果小概率事件没有发生，则接受零假设。

Question 10

Not yet answered | Marked out of 1.00 | Flag question

Consider a communication system that consists of three messages with probabilities of P_1= 0.25, P_2= 0.25 and P_3= 0.25, P_4= 0.25. The entropy of the system is (the base of the logarithm is 2):

对有k种类别的样本X，Entropy(X)=-sum^{k}_{i=1} P_i*log2(P_i)

信息熵的取值范围是[0, log2(k)]

- A. 1
- B. 2
- C. 1.5
- D. 3

Question 11

Not yet answered | Marked out of 1.00 | Flag question

Which of the following does NOT use machine learning/AI?

- A. Walkman (tape player) 什么人工智能随身听
- B. Self-driving cars
- C. SIRI/Alexa
- D. Facial recognition on App on your phone

Question 12

Not yet answered | Marked out of 1.00 | Flag question

监督学习

Which of the following is the supervised learning method?

- A. Expectation Maximization EM方法是典型的非监督学习，如GMM，k-means
- B. K-means 非监督聚类
- C. Decision Tree 必须根据分类后的效果决定每个节点选择什么特征分类
- D. Hierarchical clustering 非监督聚类，只根据簇和样本间的距离决定，与label无关

Question 13

Not yet answered | Marked out of 1.00 | Flag question

数据清洗，包括解决丢失的数据，平滑或删去噪声和错误值

Which of the following method is used in data cleaning?

D

- A. Data munging 数据整理
- B. Handling missing data 解决丢失的数据
- C. Smooth noisy data 平滑噪声
- D. All of the above

Question 14

Not yet answered | Marked out of 1.00 | Flag question

声音信号处理

Which of the following is NOT considered as an acoustic signal processing task?

A

- A. Abnormality Detection 异常检测
- B. Audio Captioning 语音字幕
- C. Automatic Music Generation (Symbolic Data) 音乐生成，你的算法有些松弛
- D. Acoustic Event Detection 声音事件检测

Question 15

Not yet answered | Marked out of 1.00 | Flag question

数据没有标签的情况下，没有答案只能使用非监督学习

If our data has no labels or true values associated with them, what type of learning may be used?

A

- A. Unsupervised learning
- B. Supervised learning
- C. Semi-supervised
- D. All of the above

Question 16

Not yet answered | Marked out of 1.00 | Flag question

Which of the following statement is true for k-fold cross-validation?

增加交叉验证折数一般会提升模型泛化能力，降低方差，但不保证

C

B

A. Higher values of k will result in higher variance on the cross-validation result.

模型的准确率不是错误率

B. The overall accuracy of the model is the average error across all k trials. X

C. The number of data points must be larger than k. k-fold中应当是采样k-1次

D. Every data point has the chance to be in the training set exactly once.

同一个样本可以重复采样

Question 17

Not yet answered | Marked out of 1.00 | Flag question

Consider the following data

X	-2	-1	1	2
Y	0	2	4	6

A linear regression model

B

$$f(x) = \omega_0 + \omega_1 x$$

最小二乘法做线性回归

is fit to the data using the least square method. What are the optimal parameters?

A. $\omega_0 = 3$ and $\omega_1 = 1.4$ $f(x) = 3 + 1.4x$ $Z = a2^2 + a\varphi^2 + a\varphi^2 + a2^2 = a4$

B. $\omega_0 = 2$ and $\omega_1 = 1$ $f(x) = 2 + x$ $Z = 0 + 1^2 + 1^2 + 2^2 = 6$

C. $\omega_0 = 2.4$ and $\omega_1 = 2$ $f(x) = 2.4 + 2x$ $Z = 1.6^2 + 1.6^2 + a4^2 +$ X

D. $\omega_0 = 3$ and $\omega_1 = 2.4$ $f(x) = 3 + 2.4x$ $Z = 1.8^2 + 1.\varphi^2 +$ X

选择题专用做法就是看谁的残差平方和最小

E=min(sum(f(x)-y))^2

Question 18

Not yet answered | Marked out of 1.00 | Flag question

既不能拟合测试集也不能拟合训练集属于欠拟合（Underfitting）

____refers to a model that can neither model the training data nor generalize to new data.

很好拟合训练集不拟合测试集属于过拟合（Overfitting）
在训练集和测试集上有均衡表现属于良好拟合（Good fitting）

C

- A. Good fitting
- B. Overfitting
- C. Underfitting
- D. None of the above

1b 2d 3a 4a 5c
6b 7d 8d 9b 10b
11a 12c 13d 14a 15a
16b 17a 18c