**机器学习考试题**

**注意：**

**1、考核题发送邮箱：340234085@qq.com抄送：**[**834822435@qq.com**](mailto:834822435@qq.com)

**（注意：两个邮箱需要同步发送。如有漏，会导致考核成绩缺失）2、邮件主题：xxx（姓名）+ 机器学习考核题目**

**3、需要在规定的截止时间内进行提交，以邮件发送的时间为准。第一部分需要在20240128之前提交（提交文件类型为word格式文件）；第二部分需要在20240204前提交（提交文件类型需要为.py格式文件）。**

**第一部分（闭卷）**

1. **选择题(单选)**
2. L1正则和L2正则的共同点是什么？\_\_B\_\_\_\_\_\_\_\_\_

A.都会让数据集中的特征数量减少

B.都会增大模型的偏差

C.都会增大模型方差

D.以上选项皆错

1. 关于欠拟合（under-fitting），下面哪个说法是正确的？\_\_\_C\_\_\_\_\_\_\_\_

A. 训练误差较大，测试误差较小

   B. 训练误差较小，测试误差较大

   C. 训练误差较大，测试误差较大

1. SVM中的核函数的作用是\_\_B\_\_\_\_？
2. 特征升维
3. 特征降维
4. 防止过拟合
5. 在数据预处理阶段，我们常常对数值特征进行归一化或标准化处理。这种处理方式理论上不会对下列哪个模型产生很大影响？\_\_\_C\_\_\_\_\_\_

A. k-Means

  B. k-NN

  C. 决策树

1. 关于K均值和DBSCAN的比较，以下说法不正确的是\_\_\_\_A\_\_\_\_\_\_\_

A、K均值丢弃被它识别为噪声的对象，而DBSCAN一般聚类所有对象

B、K均值使用簇的基于原型的概念，而DBSCAN使用基于密度的概念

C、K均值很难处理非球形的簇和不同大小的簇，DBSCAN可以处理不同大小和不同形状的簇

D、K均值可以发现不是明显分离的簇，即便簇有重叠也可以发现，但是DBSCAN会合并有重叠的簇

1. 以下哪个集成学习算法的基学习器之间不存在强依赖关系?\_\_\_C\_\_\_\_\_\_
2. LightGBM
3. AdaBoost
4. 随机森林
5. 闵可夫斯基距离表示为曼哈顿距离时p为\_\_\_A\_\_\_\_\_\_
6. 1
7. 2
8. 3
9. 在市场营销中，聚类最后可能帮助经营者\_\_\_A\_\_\_
10. 对客户群进行划分
11. 进行商品推荐
12. 识别优质客户
13. 主成分分析中各主成分之间是\_\_\_\_C\_\_\_\_\_
14. 彼此相关
15. 存在线性关系
16. 互不相关
17. 以下不属于无监督学习的算法是\_\_\_D\_\_\_\_\_\_
18. KMeans
19. DBSCAN
20. PCA
21. KNN
22. 混淆矩阵中的假正是指\_\_C\_\_\_\_
23. 模型预测为负的负样本
24. 模型预测为负的正样本
25. 模型预测为正的负样本
26. 简答题（以下是机器学习的常见面试题，供参考）
27. 什么是有监督算法，什么是无监督算法？

监督学习：根据输入的特征数据集和输出的标签数据集来训练模型。

无监督学习：根据特征数据自身特点进行的学习，将特征数据集映射到不同的簇中。

1. 什么是超参数，怎样确定超参数的值？

训练前设置的、而不是通过训练得到的参数, 可以通过手工调参、交叉验证、随机搜索等方法进行优化。

1. 什么是欠拟合，什么是过拟合？怎样解决欠拟合问题？怎样解决过拟合问题？

欠拟合：由于模型过于简单，无法捕捉到特征数据集和标签数据之间的关系，造成对训练数据集和测试数据集出现较大的预测误差。

解决方法：增加训练次数、增加特征，减少对损失函数的正则化、增大梯度下降的学习率（learning rate）等。

过拟合：模型过于复杂，训练时学习了过多的噪音，造成对未训练过的样本产生较大的预测误差。

解决方法：降低模型复杂度、删除冗余特征、控制训练的迭代次数、对损失函数进行正则化、使用多个算法集成的机器学习等。

1. 什么是损失函数？它的作用是什么？有哪些常用的损失函数？

损失函数是用来计算机器学习算法的估计值和真实值之间差距的函数。通过计算最小损失可获得最佳模型的参数。

包括：

1. 均方差/L2损失函数（回归问题）

2. 绝对差/L1损失函数（回归问题）

3. 交叉熵损失函数（逻辑回归）

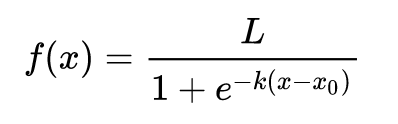
4. 指数损失函数（Adaboost）

5. Huber损失函数(GBDT）

6. Hinge损失函数(SVM）等

1. 什么是Logistic函数，它有什么特点？

Logistic函数可用下式表示：



其中：

x0为S形曲线中点的x值；

L为曲线的最大值

k为逻辑斯谛增长率或曲线的陡度。

特点是：Logistic函数呈S形，范围在(0,L)之间。当x趋向于正无穷时，*f(x)*的值逼近L，而x趋向于负无穷时，*f(x)*的值逼近0。

1. 什么是多项式扩展？它的作用是什么？

多项式扩展是将低纬度的数据扩展为高纬度的数据。作用是将二维空间中线性不可分的数据映射到高维空间后，变成线性可分的数据，从而可以使用线性回归模型来解决问题。

1. 什么是 核函数？它的作用是什么？

核函数是在低维空间到高维空间的映射，这个映射等价于对低维空间的向量先进行映射后再做内积。

作用：用低维空间中小的内积的计算量来让模型具有高维特征空间中线性可分的优点。

1. 什么是拉普拉斯平滑？

在朴素贝叶斯算法中，如果某个样本类别在训练集中没有出现过，会导致该事件的概率为零，影响后验概率的计算结果。为解决这个问题，则在每个随机变量的取值上加上一个正数λ ，λ =1时称为拉普拉斯平滑。

1. 请写出朴素贝叶斯算法中三门问题的解题过程。

p(奖=A)=1/3

p(奖=B)=1/3

p(奖=C)=1/3

p(指B/奖=A)=1/2

p(指C/奖=A)=1/3

p（指B/奖=B)=0

p(指C/奖=A)=1

p(指B/奖=C)=1

p(指C/奖=C)=0

P(奖=A/指B)=p(指B/奖=A).P(奖=A)/p(指B)=1/2\*1/3/?=1/6/?

p(奖=C/指B)=p(指B/奖=C).P(奖=C)/p(指B)=1\*1/3/?=1/3/?

1/3>1/6，所以换门后中奖的比例更大。

1. 什么是高斯朴素贝叶斯？

如果在计算后验概率时，每个特征属性为连续值，并在每个Y\_k下符合正态分布，则可以用高斯分布的概率密度公式来计算特征属性X\_i在Y\_k下的后验概率。

1. 什么是随机森林？

随机森林是在基于Bagging的树类算法，从原始样本集中用有放回重采样的方法选出n个样本，从这些样本的所有特征中随机选择k个特征，然后选择最佳分割特征作为当前节点的划分特征来训练决策树。训练m个决策树后，通过投票表决的方法决定数据属于哪一类别。这m个决策树就是随机森林。

1. 什么是boosting?主要的boosting算法有那些？

Boosting是一种集成算法，通过训练一系列弱学习器，加权累加后得到一个最终的模型，目的是减少训练简单学习器的误差。

1. bagging思想和boosting思想有哪些区别？

Bagging是采用有放回抽样的方法训练多个学习器，而Boosting是采用逐步改进一系列弱学习器的方法，得到最终模型。

1. 什么是KD树，它解决什么问题？

KD树是KNN算法中的而一种，是用所有样本的最大方差特征的中位数作为样本的划分点，以递归的方式对所有样本构建二叉树。

样本量比较大的时候，计算所有样本的距离工作量比较大，用KD树的方法可以减少计算量,快速找到数据的近邻样本。

1. 什么是数据不均衡问题，怎样解决数据不均衡问题？

数据不均衡一般指在监督学习中，样本标签值的分布不均匀。这将使得模型更倾向于将结果预测为样本标签分布较多的类别，而使少数样本的标签值预测性能下降。

解决方法包括：

1. SMOTE算法：对数据进行采用的过程中通过相似性同时生成并插样少数类别数据

2. 对多数类别进行随机降采样

3. 对数据先进行聚类，再将大的簇进行随机欠采样或者小的簇进行数据生成等

1. 什么是特征选择？为什么要做特征选择？怎样进行特征选择？

特征选择是寻找最优的特征子集，删除与标签集不相关的或者冗余的特征，以减少特征数量，简化模型，提高模型的请确定及减少运行时间。

特征选择的方法包括：

1. 方差选择法

2. 相关系数法

3. 卡方验证

4. 互信息法等

1. 什么是降维？常用的数据降维方法有哪些？它们的原理是什么？

特征降维是降低特征矩阵的维度，以解决特征过多时会导致计算量比较大，训练时间过长的问题。

常用的数据降维方法有：

1. 线性判别分析法（LDA):是有监督学习算法，通过一个旋转向量，将数据投影到一维空间。投影后标签值的类别间的均值越远越好，类别内的散列成度越小越好。

2. 主成分分析法(PCA)：是无监督学习算法，在原始空间中顺序找一组互相正交的坐标轴，第一个轴是所有样本点映射的直线中方差最大的线轴，第二个轴是与第一个轴正交的平面中样本映射的直线中方差最大的映射线轴，以此类推，找到前r个特征向量作为主成分。

1. 什么是One-Hot编码？

One-Hot编码是将离散特征的取值扩展为N维欧氏空间的向量，该离散特征出现的状态对应的向量元素设置为1，其他元素都为0。

1. 分类问题的评价指标主要有哪些？回归问题的评价指标主要有哪些？

分类问题的评价指标包括：

1.准确率、召回率、精确率、F1，FScore，

2.PR-曲线和AP值

3.ROC曲线和AUC值

回归问题的评价指标包括：

1.均方误差

2.绝对误差

3.R平方值

1. 请用简短的文字描述sklearn中以下模块的主要功能(每个模块一句话即可)：
2. sklearn.ensemble

sklearn.ensemble模块包括了集成学习的方法，可用于分类、回归和异常检测。

1. sklearn.impute

sklearn.impute模块包括了填补缺失值的方法。

1. sklearn.linear\_model

sklearn.linear\_model模块包括了多种线性模型分类器的实现方法。

1. sklearn.metrics

sklearn.metrics模块包括了计算评估模型性能的指标的方法。

1. sklearn.pipeline

sklearn.pipeline模块用于将多个数据预处理和建模的步骤封装起来，以实现更高效的建模流程。

1. sklearn.preprocessing

sklearn.preprocessing用于预处理数据，包括将数据标准化，归一化，正则化等。

1. sklearn.svm

sklearn.svm模块包括了支持向量机的数据分类和预测方法。

1. sklearn.tree

sklearn.tree模块包括了基于决策树的回归和预测方法。

1. sklearn.cluster

sklearn.cluster模块包括了无监督学习的聚类算法。

1. sklearn.datasets

sklearn.datasets模块包括了加载一些经典数据集，样本数据生成器的方法。

**第二部分（开卷）**

1. 了解Kaggle竞赛平台([www.kaggle.com](http://www.kaggle.com/))。
2. 了解Kaggle竞赛的竞赛规则和晋升体系。

参考：[Competitions Documentation | Kaggle](https://www.kaggle.com/docs/competitions)

1. 学习平台上关于Machine Learning的入门课程。

参考：[Learn Intro to Machine Learning Tutorials (kaggle.com)](https://www.kaggle.com/learn/intro-to-machine-learning)

1. 在平台上选择Spaceship Titanic项目参与竞赛。

链接：[Spaceship Titanic | Kaggle](https://www.kaggle.com/competitions/spaceship-titanic)