练习题

一、单项选择题

1、 属于监督学习的机器学习算法是(A)

**A.贝叶斯分类器**

B.主成分分析

C.K-Means

D.高斯混合聚类

1. 属于无监督学习的机器学习算法是(C)

A.支持向量机

B.Logistic回归

**C.层次聚类**

D.决策树

1. 朴素贝叶斯分类器的特点是(C)
2. 假设样本服从正态分布
3. 假设样本服从多项式分布
4. **假设样本各维属性独立**
5. 假设样本各维属性存在依赖

1. 下列方法中没有考虑先验分布的是(D)

A.最大后验估计

B.贝叶斯分类器

C.贝叶斯学习

**D.最大似然估计**

1. 对于正态密度的贝叶斯分类器，各类协方差矩阵相同时，决策函数为(A)

**A.线性决策函数**

B.非线性决策函数

C.最小距离分类器

D.以上都有可能

1. 下列属于线性分类方法的是(B)

A.决策树

**B.感知机**

C.最近邻

D. 集成学习

1. 下列方法使用最大似然估计的是(C)

A.线性鉴别分析

B.感知机

**C.Logistic回归**

D.SVM

1. SVM的原理的简单描述，可概括为(C)

A.最小均方误差分类

B.最小距离分类

**C.最大间隔分类**

D.最近邻分类

1. SVM 的算法性能取决于(D)
2. 核函数的选择
3. 核函数的参数
4. 软间隔参数C
5. **以上所有**

1. 支持向量机的对偶问题是(C)

A.线性优化问题

B.二次优化

**C.凸二次优化**

D.有约束的线性优化

1. 以下对支持向量机中的支撑向量描述正确的是(C)

A.最大特征向量

B.最优投影向量

**C.最大间隔支撑面上的向量**

D.最速下降方向

1. 假定你使用阶数为2的线性核SVM, 将模型应用到实际数据集上后，其训练准确率和测试准确率均为100%。现在增加模型复杂度(增加核函数的阶), 会发生以下哪种情况(A)

**A.过拟合**

B.欠拟合

C.什么都不会发生，因为模型准确率已经到达极限

D.以上都不对

1. 避免直接的复杂非线性变换，采用线性手段实现非线性学习的方法是(A)

**A.核函数方法**

B.集成学习

C.决策树

D.Logistic回归

1. 关于决策树节点划分指标描述正确的是(B)

A.类别非纯度越大越好

**B.信息增益越大越好**

C.信息增益率越小越好

D.基尼指数越大越好

1. 以下描述中，属于决策树策略的是(D)

A.最优投影方向

B.梯度下降方法

C.最大特征值

**D.最大信息增益**

1. 集成学习中基分类器的选择如何，学习效率通常越好(D)

A.分类器相似

B.都为线性分类器

C.都为非线性分类器

**D.分类器多样，差异大**

1. 集成学习中，每个基分类器的正确率的最低要求(A)

**A.50% 以上**

B.60% 以上

C.70% 以上

D.80% 以上

1. 下面属于Bagging方法的特点是(A)

**A.构造训练集时采用Bootstraping 的方式**

B.每一轮训练时样本权重不同

C.分类器必须按顺序训练

D.预测结果时，分类器的比重不同

1. 下面属于Boosting 方法的特点是(D)

A.构造训练集时采用Bootstraping的方式

B.每一轮训练时样本权重相同

C.分类器可以并行训练

**D.预测结果时，分类器的比重不同**

1. 随机森林方法属于(B)

A.梯度下降优化

**B.Bagging方法**

C.Boosting方法

D.线性分类

1. 假定有一个数据集S, 但该数据集有很多误差，采用软间隔SVM训练，阈值为C, 如果C的值很小，以下那种说法正确(A)

**A.会发生误分类现象**

B.数据将被正确分类

C.不确定

D.以上都不对

1. 软间隔SVM的阈值趋于无穷，下面哪种说法正确(A)

**A.只要最佳分类超平面存在，它就能将所有数据全部正确分类**

B.软间隔SVM分类器将正确分类数据

C.会发生误分类现象

D.以上都不对

1. 一般，K-NN最近邻方法在什么情况下效果好(B)

A.样本较多但典型性不好

**B.样本较少但典型性较好**

C.样本呈团状分布

D.样本呈链状分布注：最近邻属于分类算法，样本多而且典型性不好容易造成分类错误(尤其是 在分类边界上的样本点)。样本分布对聚类算法的影响较大。

1. 混合高斯聚类中，运用了以下哪种过程(A)

**A.EM算法**

B.集合运算

C.密度可达

D.样本与集合运算

1. 主成分分析方法是一种什么方法(C)

A.分类方法

B.回归方法

**C.降维方法**

D.参数估计方法

1. 过拟合现象中(A)

**A.训练样本的测试误差最小，测试样本的正确识别率却很低**

B.训练样本的测试误差最小，测试样本的正确识别率也很高

C.模型的泛化能力很高

D.通常为线性模型

1. 已知均值和方差，下面哪种分布的熵最大(C)

A.几何分布

B.指数分布

**C.高斯分布**

D.均匀分布

1. 梯度下降算法的正确步骤是什么(B)

(1)计算预测值和真实值之间的误差

(2)迭代跟新，直到找到最佳权重

(3)把输入传入网络，得到输出值

(4)初始化随机权重和偏差

(5)对每一个产生误差的神经元，改变相应的(权重)值以减小误差

A.1,2,3,4,5

**B.4,3,1,5,2**

C.3,2,1,5,4

D.5,4,3,2,1

1. 以下哪种方法会增加模型的欠拟合风险 (D)

A.添加新特征

B.增加模型复杂度

C.减小正则化系数

**D.数据增强**

1. 关于k-means算法，正确的描述是 (B)

A.能找到任意形状的聚类

**B.初始值不同，最终结果可能不同**

C.每次迭代的时间复杂度是0(n^2), 其中n是样本数量

D.不能使用核函数

1. 下列关于过拟合现象的描述中，哪个是正确的 (A)

**A.训练误差小，测试误差大**

B.训练误差小，测试误差小

C.模型的泛化能力高

D.其余选项都不对

1. 下方法中属于无监督学习算法的是 (D)

A.线性回归

B.支持向量机

C.决策树

**D.K-Means聚类**

1. 下面关于贝叶斯分类器描述错误的是 (B)

A.以贝叶斯定理为基础

**B.是基于后验概率，推导出先验概率**

C.可以解决有监督学习的问题

D.可以用极大似然估计法解贝叶斯分类器

1. 下面关于Adaboost算法的描述中，错误的是 (D)

A.是弱分类器的线性组合

B.提升树是以分类树或者回归树为基本分类器的提升办法

C.该算法实际上是前向分步算法的一个实现，在这个方法里，模型是加法模 型，损失函数是指数损失，算法是前向分步算法。

**D.同时独立地学习多个弱分类器**

1. 二分类任务中，有三个分类器h1,h2,h3, 三个测试样本x1,x2,x3 。 假设 1 表 示分类结果正确，0 表示错误， h1 在 x1,x2,x3 的结果分别

(1,1,0),h2,h3 分别为 (0,1,1),(1,0,1),按投票法集成三个分类器，下列说法正确的是 (A)

**A.集成提高了性能**

B.集成没有效果

C.集成降低了性能

D.集成效果不能确定

1. 下列哪个不属于常用的文本分类的特征选择算法 (D)

A.卡方检验值

B.互信息

C.信息增益

**D.主成分分析**

1. 以下哪个模型不是分类模型 (B)

A.最近邻 **B.K均值** C.朴素贝叶斯 D.逻辑回归

1. 在混淆矩阵中，识别率可以表示为（）。

A.(TP)/(TP+TN+FP+FN)

B.(TP+FN)/(TP+TN+FP+FN)

**C.(TP+TN)/(TP+TN+FP+FN)**

D.(FP+TN)/(TP+TN+FP+FN)

正确答案：C

1. 在混淆矩阵中，系统召回率定义为（）。

**A.TP/（TP+FN）**

B.TN/（FP+TN）

C.TP/（FP+TP）

D.TN/（TP+TN）

正确答案：A

1. 给人脸打上标签再让模型进行学习训练的方法，属于（）

A.强化学习

B.半监督学习

**C.监督学习**

D.无监督学习正确答案：C

1. 机器学习进行的第一步是（）

A.数据收集

**B.特征提取**

C.交叉验证

D.模型训练

正确答案：B

1. 若我们用一类对另一类的方法来解决多分类问题，当有K类时，我们需要训练（）个支持向量机。

**A.K(K-1)/2**

B.K-1

C.K

D.K(K+1)/2

正确答案：A

1. 如果一个样本空间线性可分，那么，我们能找到（）个平面来划分样本。

A.1

**B.无数**

C.K

D.不确定正确答案：B

1. 向量x=[1,2,3,4,-9,0]的L1范数是多少
2. 1
3. **19**
4. 6

D.20

正确答案: B

45、向量 X=[1,2,3,4,-9,0] 的 L2 范数为（ ）

1. 1
2. 19
3. 6
4. **√111** 正确答案: D

46、一般，k-NN最近邻方法在（ B ）的情况下效果较好

1. 样本较多但典型性不好
2. **样本较少但典型性好**
3. 样本呈团状分布

D.样本呈链状分布

47、以下哪些方法不可以直接来对文本分类？

1. **K-Means**
2. 决策树
3. 支持向量机
4. kNN

正确答案: A

48、以下说法错误的一项是

1. 负梯度方向是使函数值下降最快的方向
2. 当目标函数是凸函数时，梯度下降法的解是全局最优解
3. **梯度下降法比牛顿法收敛速度快**
4. 拟牛顿法不需要计算Hesse矩阵正确答案: C

49、下列说法错误的是？

1. 当目标函数是凸函数时，梯度下降算法的解一般就是全局最优解
2. 进行 PCA 降维时，需要计算协方差矩阵
3. **沿负梯度的方向一定是最优的方向**
4. 利用拉格朗日函数能解带约束的优化问题正确答案: C

50、下面是交叉验证的几种方法：

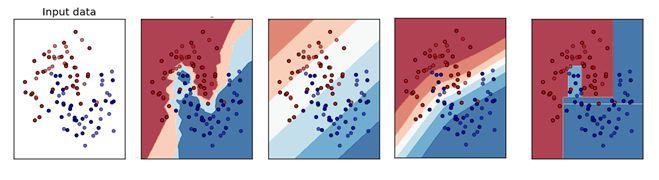
1. Bootstrap
2. 留一法交叉验证
3. 5 折交叉验证
4. 重复使用两次 5 折交叉验证

请对上面四种方法的执行时间进行排序，样本数量为 1000。

1. 1 > 2 > 3 > 4
2. 2 > 3 > 4 > 1
3. 4 > 1 > 2 >３
4. **2 > 4 > 3 > 1**

正确答案: D

51、下面哪个对应的是正确的 KNN 决策边界？



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

正确答案: A

52、下面哪句话是正确

1. 机器学习模型的精准度越高，则模型的性能越好
2. 增加模型的复杂度，总能减小测试样本误差
3. **增加模型的复杂度，总能减小训练样本误差**
4. 以上说法都不对

正确答案: C

53、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？

1. 基本模型之间相关性高
2. **基本模型之间相关性低**
3. 集成方法中，使用加权平均代替投票方法
4. 基本模型都来自于同一算法

正确答案: B

54、在训练完 SVM 之后，我们可以只保留支持向量，而舍去所有非支持向量。

仍然不会影响模型分类能力。这句话是否正确？

1. **正确**
2. 错误

正确答案: A

55、假设我们使用原始的非线性可分版本的 Soft-SVM 优化目标函数。我们需要做什么来保证得到的模型是线性可分离的？

1. C = 0
2. C = 1
3. **C 正无穷大**
4. C 负无穷大

正确答案: C

56、如果在大型数据集上训练决策树，为了花费更少的时间来训练这个模型，下列哪种做法是正确的？

1. 增加树的深度
2. 增加学习率
3. **减小树的深度**
4. 减少树的数量

正确答案: C

57、点击率预测是一个正负样本不平衡问题（例如 99% 的没有点击，只有 1% 点击），假如在这个非平衡的数据集上建立一个模型，得到训练样本的正确率是

99%，则下列说法正确的是？

1. 模型正确率很高，不需要优化模型了
2. **模型正确率并不高，应该建立更好的模型**
3. 无法对模型做出好坏评价
4. 以上说法都不对

正确答案: B

58、关于 k 折交叉验证，下列说法正确的是？

1. k 值并不是越大越好，k 值过大，会降低运算速度
2. 选择更大的 k 值，会让偏差更小，因为 k 值越大，训练集越接近整个训练样本
3. 选择合适的 k 值，能减小验方差
4. **以上说法都正确**

正确答案: D

1. 下列描述无监督学习错误的是（ ）。

A.无标签

B.核心是聚类

**C.不需要降维**

D.具有很好的解释性

正确答案：C

1. 下列描述有监督学习错误的是（ ）。

A.有标签

B.核心是分类

**C.所有数据都相互独立分布**

D.分类原因不透明

正确答案：C

1. 混淆矩阵的假正是指（ ）。

A.模型预测为正的正样本

**B.模型预测为正的负样本**

C.模型预测为负的正样本

D.模型预测为负的负样本

正确答案：B

1. 混淆矩阵的真负率公式是为（ ）。

A.TP/(TP+FN)

B.FP/(FP+TN)

C.FN/(TP+FN)

**D.TN/(TN+FP)**

正确答案：D

1. 混淆矩阵中的 TP=16，FP=12，FN=8，TN=4，准确率是（ ）。

A.1/4

**B.1/2**

C.4/7

D.4/6

正确答案：B

1. 混淆矩阵中的 TP=16，FP=12，FN=8，TN=4，精确率是（ ）。

A.1/4

B.1/2

**C.4/7**

D.2/3

正确答案：C

1. 混淆矩阵中的 TP=16，FP=12，FN=8，TN=4，召回率是（ ）。

A.1/4

B.1/2

C.4/7

**D.2/3**

正确答案：D

1. 混淆矩阵中的 TP=16，FP=12，FN=8，TN=4，F1-score 是（ ）。

A.4/13

**B.8/13**

C.4/7

D.2/3

0.00/2.00

正确答案：B

1. EM 算法的 E 和 M 指什么？（ ）

**A.Expectation-Maximum**

B.Expect-Maximum

C.Extra-Maximum

D.Extra-Max 正确答案：A

1. EM 算法的核心思想是？（ ）

**A.通过不断地求取目标函数的下界的最优值，从而实现最优化的目标。**

B.列出优化目标函数，通过方法计算出最优值。

C.列出优化目标函数，通过数值优化方法计算出最优值。

D.列出优化目标函数，通过坐标下降的优化方法计算出最优值。

正确答案：A

1. SVM 算法的性能取决于（ ）。

A.核函数的选择

B.核函数的参数

C.软间隔参数

**D.以上所有**正确答案：D

1. SVM 中的代价参数 C 表示什么？（ ）

**A.在分类准确性和模型复杂度之间的权衡**

B.交叉验证的次数

C.以上都不对

D.用到的核函数

正确答案：A

1. 下列有关支持向量机说法不正确的是（ ）。

**A.得到的是局部最优解**

B.具有很好的推广能力

C.是凸二次优化问题

D.采用结构风险最小化原理

正确答案：A

1. 下列有关核函数不正确的是（ ）。

A.可以采用 cross-va1idalion 方法选择最佳核函数

**B.满足 Mercer 条件的函数不一定能作为支持向量机的核函数**

C.极大地提高了学习机器的非线性处理能力

D.函数与非线性映射并不是一一对应的关系正确答案：B

1. 一对一法分类器，k 个类别需要多少个 SVM（ ）。
2. **k(k-1)/2**
3. k(k-1)

C.k

D.k!

正确答案：A

1. 有关聚类分析说法错误的是（ ）。

A.无须有标记的样本

B.可以用于提取一些基本特征

C.可以解释观察数据的一些内部结构和规律

**D.聚类分析一个簇中的数据之间具有高差异性**正确答案：D

1. 两个 n 维向量𝛼(𝑥11, 𝑥12, ⋯ , 𝑥1𝑛) 和𝛽(𝑥21, 𝑥22, ⋯ , 𝑥2𝑛)之间的欧式距离（euclidean distance)为（ ）。

**A.𝑑12 = √(𝛼 − 𝛽)(𝛼 − 𝛽)𝑇**

B.

C.𝑑12 = max (|𝑥1𝑖 − 𝑥2𝑖|) D.cos(𝜃) = (𝛼 ∙ 𝛽)/(|𝛼||𝛽|) 正确答案：A

1. 闵可夫斯基距离表示为曼哈顿距离时 p 为（ ）。

**A.1**

B.2

C.3

D.4

正确答案：A

1. 关于 K-means 说法不正确的是（ ）。

A.算法可能终止于局部最优解

B.簇的数目 k 必须事先给定

C.对噪声和离群点数据敏感

**D.适合发现非凸形状的簇正确答案**：D

1. 假设某事件发生的概率为 p，则此事件发生的几率为（ ）。

A.p

B.1-p

**C.p/(1-p)**

D.(1-p)/p

正确答案：C

1. 同质集成中的个体学习器亦称（）
2. 组件学习器
3. **基学习器**
4. 异质学习器

D、同质学习器正确答案：B

1. 假设我们使用原始的非线性可分版本的 Soft-SVM 优化目标函数。我们需要做什么来保证得到的模型是线性可分离的？

A、C = 0

B、C = 1

**C、C 正无穷大**

D、C 负无穷大正确答案：C

1. 关于logistic回归和SVM不正确的是（）
2. **Logistic 回归目标函数是最小化后验概率**
3. Logistic回归可以用于预测事件发生概率的大小
4. SVM可以有效避免模型过拟合
5. SVM目标是结构风险最小化正确答案：A

82、下面关于SVM算法叙述不正确的是（ ）

1. **SVM是一种基于经验风险最小化准则的算法**
2. SVM求得的解为全局唯一最优解
3. SVM在解决小样本、非线性及高维模式识别问题中具有优势
4. SVM最终分类结果只与少数支持向量有关正确答案：A

83、KNN算法属于一种典型的（）算法

A、无监督学习

B、半监督学习

C、弱监督学习

**D、监督学习**正确答案：D

84、按照求解方法进行分类算法的划分，下列中为生成模型的是（ ）

1. 决策树
2. 支持向量机SVM
3. K近邻
4. **贝叶斯分类器正确答案**：D

85、EM算法的停止条件（）

1. 数据样本异常
2. **已达到最大迭代轮数**
3. 似然函数减小
4. 训练器异常正确答案：B

86、SVM中的代价参数C表示什么？

1. 交叉验证的次数
2. 以上都不对
3. **用到的核函数**
4. 在分类准确性和模型复杂度之间的权衡正确答案：D

87、对函数dist（.，.）若它是一个距离度量则需要满足的基本特性中以下哪个不是正确答案

A、非负性 B、同一性 C、对称性 D、递增性

正确答案：D

88、关于决策树结点划分指标描述正确的是

1. **信息增益越大越好**
2. 类别非纯度越大越好
3. 基尼指数越大越好
4. 信息增益率越小越好正确答案：A

89、以下关于训练集、验证集和测试集说法不正确的是( )。

1. 验证集用于调整模型参数
2. 测试集是纯粹是用于测试模型泛化能力
3. 以上说法都不对
4. **训练集是用来训练以及评估模型性能**

正确答案：D

90、对于在原空间中线性不可分问题，支持向量机（）。

1. 无法处理
2. **将数据映射到核空间中**
3. 在原空间中寻找线性函数划分数据
4. 在原空间中寻找非线性函数的划分数据正确答案：B

91、一对一法分类器，k个类别需要多少个SVM：

A、k(k-1)

B、k

C、**k(k-1)/2**

D、k!

正确答案：C

92、以下关于感知器说法错误的是: ( )

1. 感知器中的偏置只改变决策边界的位置
2. 可为感知器的输出值设置阈值使其用于处理分类问题
3. **单层感知器可以用于处理非线性学习问题**
4. 感知器是最简单的前馈式人工神经网络正确答案：C

93、当数据分布不平衡时，我们可采取的措施不包括( )。

1. 对数据分布较少的类别过采样
2. 对数据分布较少的类别赋予更大的权重
3. 对数据分布较多的类别欠采样
4. **对数据分布较多的类别赋予更大的权重**正确答案：D

94、朴素贝叶斯是一种典型的基于概率的机器学习方法，它利用了

A、先验概率

B、后验概率

**C、以上都是**

D、以上都不是

正确答案：C

95、对于非概率模型而言，可按照判别函数线性与否分成线性模型与非线性模型。下面哪个模型不属于线性模型？

1. **感知机**
2. AdaBoost

C、K-means

D、k近邻

正确答案：A

96、模型评估的常用方法有哪些

1. 留出法
2. 交叉验证法
3. 自助法
4. **以上都是**正确答案：D

97、KNN算法是基于（）

A、颜色空间 B、概率空间 C、距离空间 D、线性空间正确答案：C

98、下列哪种方法可以用来缓解过拟合的产生：( )。

1. **正则化**
2. 增加更多的特征
3. 以上都是
4. 增加模型的复杂度正确答案：A

99、以下哪个不是常见的决策树算法

1. ID3
2. C4.5
3. CART

**D、DBSCAN**

正确答案：D

100、关于EM算法正确的是

**A、EM算法包括两步：E算法和M算法**

B、EM算法一定能收敛到全局最大值点

1. 英文全称是Expectation-Minimization
2. 以上都不正确正确答案：A

101、不属于KNN算法要素的是:

1. 距离度量
2. 分类决策的规则
3. k值的选择
4. **训练样本的个数**正确答案：D

102、“没有免费的午餐定理”告诉我们

1. 对于一个特定的问题，任何算法都是一样好的
2. **没有可以适应一切问题的算法**
3. 设计好的算法是徒劳的
4. 我们不能对问题有先验假设正确答案：B

二、多项选择题

1、 对单层感知机判别分类，描述正确的是(ABC)

1. 线性分类
2. 监督学习
3. 错误误差最小
4. **错误误差最大**

2、 对聚类问题描述不正确的(ACD)

1. 监督学习
2. **无监督学习**
3. 线性决策
4. 增量学习

3、 以下可行的最近邻分类的加速方案(AB)

1. **分层搜索**
2. **训练样本缩减**
3. 样本增加
4. 非线性投影

4、Adaboost方法中，需要迭代调整的两个重要参数是(AB)

1. **样本权重**
2. **分类器权重**
3. 梯度变化率
4. 梯度

5、以下对层次聚类描述正确的(BD)

1. 监督学习
2. **自顶向下寻找最优划分**
3. 集成学习
4. **自底向上寻找最优合并**

6、支持向量机可能解决的问题(ABC)

1. 线性分类
2. 非线性分类
3. 回归分析
4. **BP 算法**

7、影响K-Means聚类算法结果的主要因素有(BC)

1. 样本顺序
2. **相似性度量**
3. **初始聚类中心**
4. 样本类别

1. 下面关于集成学习的描述，正确的是(AD)

**A.Bagging方法可以并行训练**

B.Bagging方法基学习器的比重不同

**C.Boosting方法可以并行训练**

D.Boosting方法基学习器的比重不同

1. 下面关于最大熵模型的描述，正确的是(ABC)
2. 思想是在满足一定约束条件下，概率分布的熵最大
3. 是一种信息论模型
4. 在已知均值和方差的条件下，最大熵分布是高斯分布
5. **在已知均值和方差的条件下，最大熵分布是指数分布**

10、给定两个特征向量，以下哪些方法可以计算这两个向量相似度 (ABD)

1. 欧式距离
2. 夹角余弦
3. **信息熵**
4. 曼哈顿距离

11、类别不平衡就是指分类问题中不同类别的训练样本相差悬殊的情况，例如 正例有 900 个，而反例只有 100 个，这个时候我们就需要进行相应的处理来平衡这个问题，下列方法正确的是 (ACD)

1. 在训练样本较多的类别中进行欠采样
2. **在训练样本较多的类别中进行过采样**
3. 直接基于原数据集进行学习，对预测值进行再缩放处理
4. 通过对反例中的数据进行插值，来产生额外的反例

12、以下可以有效解决过拟合的方法是 (ABD)

1. 增加样本数量
2. 通过特征选择减少特征数量
3. **训练更多的迭代次数**
4. 采用正则化方法

13、 在分类问题中，我们经常会遇到正负样本数据量不等的情况，比如正样本为 10w条数据，负样本只有1w条数据，以下最合适的处理方法是 (ACD)

1. 将负样本重复10次，生成10w样本量，乱顺序参与分类
2. **直接进行分类，可以最大限度利用数据**
3. 从10w正样本中随机抽取1w参与分类
4. 将负样本每个权重设置为10,正样本权重为1,参与训练过程

1. 下面哪几种机器学习的分类，完全不需要人工标注数据？

A.半监督学习

**B.强化学习**

**C.无监督学习**

D.监督学习

正确答案：B、C

1. 以下哪些算法是无监督学习算法?

**A.空间聚类**

**B.主成分分析**

C.支持向量机

D.Q-LEARNING

正确答案：A、B

1. 以下哪些算法是监督学习算法?

**A.人工神经网络**

B.高斯混合模型概率密度估计

C.ACTOR-CRITIC 算法

**D.支持向量机**

正确答案：A、D

1. 当我们利用二分类支持向量机来解决多分类问题是，我们有哪两种策略？（）

**A.一类对另一类**

**B.一类对K-1类**

C.一类对K类

D.2类对K-2类正确答案：A、B

1. 在利用二分类支持向量机来解决多分类的问题中，为了减少支持向量机的个数，我们可以用（）来构建树状结构的多分类模型。

A.强化学习

**B.聚类**

C.人工神经网络

**D.决策树**

正确答案：B、D

1. 核函数满足的两个条件（）。

**A.交换性**

B.正交性

C.鲁棒性

**D.半正定性**

正确答案：A、D

1. 下列对核函数的说法中，正确的有（）。

A.核函数是为了解决映射函数的无限维问题

B.核函数是为了解决在低维空间中线性不可分的情况

C.不需要映射函数的具体形式

**D.核函数与映射函数是一一对应的关系**正确答案：A、B、C、

1. 下列属于生成模型的有（）。

**A.支持向量机**

B.混合高斯模型

C.隐马尔可夫模型

D.朴素贝叶斯

正确答案：B、C、D

1. **下列属于判别模型的有（）。**

**A.SVM**

**B.多层感知机**

**C.决策树**

**D.逻辑递归模型**

**正确答案：A、B、C、D**

1. 小明参加某公司的大数据竞赛，他的成绩在大赛排行榜上原本居于前二十，后来他保持特征不变，对原来的模型做了1天的调参，将自己的模型在自己本地测试集上的准确率提升了 5%，然后他信心满满地将新模型的预测结果更新到了大赛官网上，结果懊恼地发现自己的新模型在大赛官方的测试集上准确率反而下降了。对此，他的朋友们展开了讨论，请将说法正确的选项打勾（ BD ）

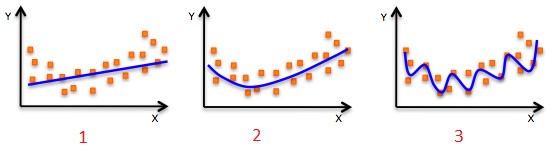
A.小芳：从机器学习理论的角度，这样的情况不应该发生，快去找大赛组委会反应

**B.小刚：你这个有可能是由于过拟合导致的**

C.小月：早就和你说过了，乖乖使用默认的参数就好了，调参是不可能有收益的

**D.小平：你可以考虑一下，使用交叉验证来验证一下是否发生了过拟合**

1. 下面三张图展示了对同一训练样本，使用不同的模型拟合的效果（蓝色曲线）。那么，我们可以得出哪些结论（ ACD ）？

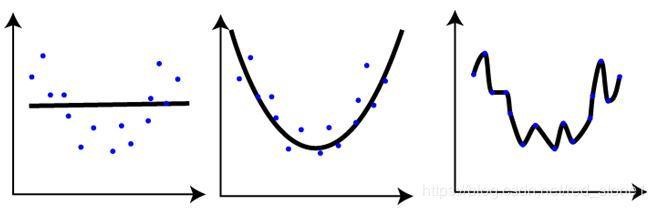


1. 第 1 个模型的训练误差大于第 2 个、第 3 个模型
2. **最好的模型是第 3 个，因为它的训练误差最小**
3. 第 2 个模型最为“健壮”，因为它对未知样本的拟合效果最好
4. 第 3 个模型发生了过拟合
5. 所有模型的表现都一样，因为我们并没有看到测试数据

**25、下列哪种方法可以用来减小过拟合？（ABCD ）**

1. **更多的训练数据**
2. **L1 正则化**
3. **L2 正则化**
4. **减小模型的复杂度**

26、如下图所示，对同一数据集进行训练，得到 3 个模型。对于这 3 个模型的评估，下列说法正确的是？（ACD）



1. 第一个模型的训练误差最大
2. **第三个模型性能最好，因为其训练误差最小**
3. 第二个模型最稳健，其在测试集上表现应该最好
4. 第三个模型过拟合

**27、我们想要训练一个 ML 模型，样本数量有 100 万个，特征维度是 5000，面对如此大数据，如何有效地训练模型（多选）？**

1. **对训练集随机采样，在随机采样的数据上建立模型**
2. **尝试使用在线机器学习算法**
3. **使用 PCA 算法减少特征维度**

正确答案: ABC

28、我们知道二元分类的输出是概率值，一般设定输出概率大于或等于 0.5，则预测为正类；若输出概率小于 0.5，则预测为负类。

那么，如果将阈值 0.5 提高，例如 0.6，大于或等于 0.6 的才预测为正类，则准确率（Precision）和召回率（Recall）会发生什么变化（多选）？

1. **准确率（Precision）增加或者不变**
2. 准确率（Precision）减小
3. **召回率（Recall）减小或者不变**
4. 召回率（Recall）增大

正确答案: A C

29、观察样本次数如何影响过拟合（多选）？注意：所有情况的参数都保持一致。

1. **观察次数少，容易发生过拟合**
2. 观察次数少，不容易发生过拟合
3. 观察次数多，容易发生过拟合
4. **观察次数多，不容易发生过拟合**

正确答案: A D

30、下列关于极大似然估计（Maximum Likelihood Estimate，MLE），说法正确的是（多选）？

1. **MLE 可能并不存在**
2. MLE 总是存在
3. **如果 MLE 存在，那么它的解可能不是唯一的**
4. 如果 MLE 存在，那么它的解一定是唯一的

正确答案: A C

三、判断题

1. 信息熵是度量样本集合纯度最常⽤的⼀种指标。信息熵的值越⼤，说明样本集合的纯度

越⾼。（×）

1. 数据集⼀般划分为训练集、验证集和测试集三部分，训练集⽤于建模，验证集（开发集）

⽤于模型验证与矫正，测试集⽤于模型的最终评估。（√）

1. F1值是精确率与查全率的加权平均值，综合平衡了精确率与查全率两个指标的特点，F1 值突出对分类错误的评估。（√）
2. 训练/测试集的划分要尽可能保持数据分布的⼀致性，避免因数据划分过程引⼊额外的偏

差⽽对最终结果产⽣影响，例如在分类任务中⾄少要保持样本的类别⽐例相似。（√）

1. 逻辑斯蒂（对数⼏率）回归是⼀种回归学习⽅法。（×）
2. 正则化是为了防⽌模型过拟合⽽引⼊额外信息，对模型原有逻辑进⾏外部⼲预和修正，

从⽽提⾼模型的泛化能⼒。（√）

1. 在各类机器学习算法中，过拟合和⽋拟合都是可以彻底避免的。（×）
2. 寻找最优超参数费时费⼒，应该在模型训练之前就指定最优参数。（×）
3. 训练集与验证集的样本是不同的。（√）
4. k均值算法，是⼀种原型聚类算法。（√）
5. 学习率越⼤，训练速度越快，最优解越精确。（×）
6. 决策树学习是⼀种逼近离散值⽬标函数的⽅法，学习到的函数被表⽰为⼀棵决策树。（√）
7. 聚类和分类的区别在于⽤于聚类的训练样本的类标记是未知的。（√）
8. 在凸优化中只有一个全局最优解。（√） 15、支持向量是指分布在样本簇边缘的样本点。（×）
9. 通过计算模型的对样本识别率，就可以完全确定模型性能的好坏。（×）
10. 利用ROC曲线来度量分类模型性能是，曲线越靠近左上角，则模型分类性能越好。（√）
11. ⼀般的，⼀棵决策树包含⼀个根结点、若⼲个内部结点和若⼲个叶结点；叶结点对应于

决策结果，其他每个结点则对应于⼀个属性测试；根结点包含样本全集。（√）

1. 监督学习的学习数据既有特征（feature），也有标签（label）。（√）
2. 直观上看，我们希望“物以类聚”，即聚类的结果“簇内相似度”⾼，且“簇间”相似

度低。（√）

1. 关于EM算法的收敛性，EM算法理论上不能够保证收敛。（×）
2. 关于EM算法的用途，EM算法只适用不完全数据的情形。（×）
3. Jessen不等式等号成立的条件是：变量为常数。

正确答案：√

1. Jessen不等式E(f(x)) >= f(E(x)), 左边部分大于等于右边部分的条件是函数f是凸函数，如果f是凹函数，左边部分应该是小于等于右边部分。

正确答案：√

1. EM算法因为是理论可以保证收敛的，所以肯定能够取得最优解。（×）
2. EM 算法首先猜测每个数据来自哪个高斯分布，然后求取每个高斯的参数，之后再去重新猜测每个数据来自哪个高斯分布，类推进一步迭代，直到收敛，从而得到最后的参数估计值。

正确答案：√

1. EM算法，具有通用的求解形式，因此对任何问题，其求解过程都是一样，都能很容易求得结果。（×）
2. EM算法通常不需要设置步长，而且收敛速度一般很快。

正确答案：√

1. 吉布斯采样是一种通用的采样方法，对于任何概率分布都可以采样出对应的样本。（×）
2. LDA模型的核心假设是：假设每个文档首先由主题分布表示，然后主题由词概率分布表示，形成文档-主题-词的三级层次。

正确答案：√

1. Gibbs采样是一类通用的采样方法，和M-H采样方法没有任何关系。（×）
2. 关于LDA模型中的K，K的指定，必须考虑数据集合的特点，选择一个较为优化的数值。

（√）

1. LDA模型是一种生成式模型。

正确答案：√

1. 主题建模的关键是确定数据集合的主题个数。（×）
2. 主题建模本质上是：一种新的文档表示方法，主要是通过主题的分布来表示一个文档。

一种数据压缩方法，将文档压缩在更小的空间中。

正确答案：√

1. SVM中的泛化误差代表SVM对新数据的预测准确度。

正确答案：√

1. 若参数C（cost parameter）被设为无穷，只要最佳分类超平面存在，它就能将所有数据全部正确分类。

正确答案：√

1. “硬间隔”是指SVM允许分类时出现一定范围的误差。（×）
2. 支持向量是最靠近决策表面的数据点。

正确答案：√

1. 数据有噪声，有重复值，不会导致SVM算法性能下降。（×）
2. 分裂层次聚类采用的策略是自底向上。（×）
3. DBSCAN对参数不敏感。（×）
4. EM聚类属于软分聚类方法。

正确答案：√

44、k-means算法、EM算法是建立在凸球形的样本空间上的聚类方法。

正确答案：√

1. 逻辑斯蒂回归模型是一种回归算法。（×）
2. 熵最大时，表示随机变量最不确定，也就是随机变量最随机，对其行为做准确预测最困难。

正确答案：√

1. 从最大熵思想出发得出的最大熵模型，采用最大化求解就是在求 P(y|x)的对数似然最大化。

正确答案：√

1. GIS算法的收敛速度由计算更新值的步长确定。C值越大，步长越大，收敛速度就越快。

（×）

1. 贝叶斯网络具有的条件独立性是结点与其后代结点条件独立。（×）
2. 最大似然估计方法是实例数据不完备情况下的学习方法。（×）
3. 隐马尔可夫模型是一种统计模型，经常用来描述一个含有隐含未知参数的马尔可夫过程。

正确答案：√

1. LSTM和GRU网络因为引入了门控单元，可以缓解梯度消失问题。

正确答案：√

1. 循环神经网络按时间展开后就可以通过反向传播算法训练了。

正确答案：√

1. 贝叶斯分类器中只有一个判别函数 （×）
2. 回归问题和分类问题都有可能发生过拟合

A、正确 B、错误

正确答案：A

59、过拟合比欠拟合更容易克服。

A、正确 B、错误

正确答案：B

60、概率模型的训练过程就是参数估计的过程

A、正确 B、错误

正确答案：A

61、随机森林的两个随机指的是随机选取样本和随机选取属性

A、正确 B、错误

正确答案：A

1. 逻辑回归是一个回归模型（×）
2. 支持向量是最靠近决策表面的数据点

A、正确 B、错误

正确答案：A

64、随机森林的训练效率通常低于Bagging

A、正确 B、错误

正确答案：B

65、PCA是一种有效的降维去噪方法

A、正确 B、错误

正确答案：A

66、决策树的生成是一个递归过程。在决策树基本算法中，有三种情形会导致递归返回。

A、正确 B、错误

正确答案：A

67、K均值算法的E值越小则簇内样本相似度越低

A、正确 B、错误

正确答案：B

68、在初始数据量足够时，自助法比交叉验证法更为常用。

A、正确 B、错误

正确答案：B

69、逻辑回归分类的精度不够高，因此在业界很少用到这个算法

A、正确 B、错误

正确答案：B

1. 在基于SGD随机梯度下降算法的神经网络中，每次打乱数据是非常重要和必不可少 A、正确 B、错误

正确答案：A

1. AGNES算法分为单链接、全链接、均链接算法

A、正确 B、错误

正确答案：A

72、PCA是有监督学习，是有参数限制的

A、正确 B、错误

正确答案：B

73、Bagging是并行式集成学习方法最著名的代表

A、正确 B、错误

正确答案：A

75、模型泛化能力与训练样本数量无关

A、正确 B、错误

正确答案：B

77、剪枝是决策树学习算法对付“过拟合”的主要手段，决策树剪枝的基本策略有“预剪枝” 和“后剪枝”。

A、正确 B、错误

正确答案：A

78、一般情形下，后剪枝决策树的欠拟合风险很小，泛化性能往往优于预剪枝决策树。

A、正确 B、错误

正确答案：A

79、过拟合是有监督学习的挑战，而不是无监督学习（×）

81、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

A、正确 B、错误

正确答案：A

82、硬投票计算出每个类别的平均估算概率，然后选出概率最高的类别。

A、正确 B、错误

正确答案：B

83、"过拟合是有监督学习的挑战，而不是无监督学习"

A、正确 B、错误

正确答案：B

84、线性回归模型只能处理具有线性关系的数据。

A、正确 B、错误

正确答案：B

85、一个贝叶斯网由结构和参数两个部分构成

A、正确 B、错误

正确答案：A

86、Bagging只适用于二分类任务

A、正确 B、错误

正确答案：B

87、任何一个有效的机器学习算法必须有其归纳偏好

A、正确 B、错误

正确答案：A

88、最近邻分离器的泛化错误率不会超过 贝叶斯最优分类器错误率的两倍

A、正确 B、错误

正确答案：A

89、一般来说，查准率高时，查全率也高。

A、正确 B、错误

正确答案：B

90、BP算法的反向传播是为了对权值进行调整。

1. 正确
2. 错误

正确答案：A

91、预剪枝决策树通常比后剪枝决策树保留了更多的分支。

A、正确 B、错误

正确答案：B

92、预剪枝决策树的训练时间开销比后剪枝决策树要大得多。

A、正确 B、错误

正确答案：B

93、簇内相似度高且簇间相似度低的聚类结果比较好

A、正确 B、错误

正确答案：A

94、决策树是基于树结构来进行决策的，决策树学习的目的是为了产生一棵泛化能力强的决策树。

A、正确 B、错误

正确答案：A

95、梯度下降法中梯度方向是函数值下降最快方向。

A、正确 B、错误

正确答案：B

96、“独依赖估计”是半朴素贝叶斯分类器最常用的一种策略

A、正确 B、错误

正确答案：A

97、流形学习是一种非线性的维数约简方法

A、正确 B、错误

正确答案：A

98、反应事件或对象在某方面的表现或性质的事项称为属性

A、正确 B、错误正确答案：A

99、SVM中的泛化误差代表SVM对新数据的预测准确度

A、正确 B、错误

正确答案：A

100、神经网络算法不能用于数据降维

A、正确 B、错误

正确答案：B

101、决策树的适用面较广，对于分类应用和回归应用，决策树都可以被用来构建模型。

A、正确 B、错误

正确答案：A

102、集成学习方法只能用于分类任务

A、正确 B、错误

正确答案：B

103、密度直达和密度可达满足对称性

A、正确 B、错误

正确答案：B

104、通常，我们认为对于一个系统来说，误差越小越好，因此无论是泛化误差还是经验误差，都是越小越好。

A、正确 B、错误

正确答案：B

105、KNN算法的基本思路是近朱者赤，近墨者黑

A、正确 B、错误

正确答案：A

106、无监督学习任务中研究最多、应用最广的是聚类

A、正确 B、错误正确答案：A