**2020年上海市高等学校信息技术水平考试试卷**

**四级 人工智能（自然语言处理与识别方向 模拟卷）**

（本试卷考试时间 150 分钟）

**一、单选题 ( 本大题 15 道小题 ，每小题 1 分，共 15 分），从下面题目给出的A、B、C、D四个可供选择的答案中选择一个正确答案。**

1.在回归模型中，下列\_\_\_\_在权衡欠拟合和过拟合中影响最大。

A.多项式阶数

B.更新权重 w 时，使用的是矩阵求逆还是梯度下降

C.使用常数项

D.增加数据量

2.A和B分别代表两个事件，如果P(A, B)降低，同时P(A)上升，\_\_\_\_是正确的。

A.P(B|A)降低

B.P(A|B)降低

C.P(B)降低

D.P(B)上升

3.癌症检查数据样本有10000个，其中10个数据祥本是有癌症，其它是无癌症。假设分类模型在无癌症数据9990中预测正确了9980个，在10个癌症数据中预测正确了9个，此时真阳=9，真阴=9980，假阳=10，假阴=1 。则该分类模型的F1-score为\_\_\_\_。

A.62.07%

B.99.89%

C.47.36%

D.76.27%

4.在测试一假设h时，发现在一包含n=1000个随机抽取样例的样本s上，它出现r=300个错误, 计算errors(h)的标准差为\_\_\_\_。

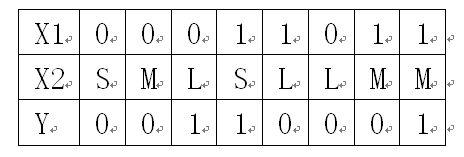
A.0.0145

B.0.145

C.1.45

D.14.5

5. 下表为某训练集数据，其中X1，X2为特征，Y为分类标记，则使用该训练集学习到的朴素贝叶斯分类器对x = (1，M) 的分类结果为\_\_\_\_。



A.0

B.1

C.不确定

D.0和1都有可能

6.关于主成分分析算法，以下步骤\_\_\_\_是错误的。

A.对所有样本进行去中心化

B.计算样本的协方差矩阵

C.对协方差矩阵做特征值分解

D.取最大的低维空间维数特征值所对应的特征向量输出投影矩阵

7.四个点坐标为(1,1)，(1,0)，(-1,-1)，(-1,0)，用 SVM 分类的决策边界是\_\_\_\_。

A.x = 0

B.y = x

C.y = -x

D.y = 0

8.在大数据集上训练决策树，为减少训练时间，可使用以下\_\_\_\_\_\_\_\_方法。

A.减少树的深度

B.增加树的深度

C.增加学习率

D.减少树的数量

9.关于偏差和方差，以下说法不正确的是\_\_\_\_。

A.如果能保证或验证一批训练集来自同一个分布，算法在这批训练集上的学习结果会是一致的

B.偏差-方差分解试图对学习算法的期望泛化错误率进行拆解

C.泛化误差可以分解为偏差、方差和噪声之和

D.方差与偏差通常是有冲突的，其中方差刻画数据扰动造成的影响，偏差刻画的是学习算法本身的拟合能力

10.下列哪个神经网络结构会发生权重共享\_\_\_\_\_\_\_\_。

A.卷积神经网络和循环神经网络

B.卷积神经网络

C.循环神经网络

D.全连接神经网络

11.Shanghai和Beijing的编辑距离是\_\_\_\_。

A.7

B.8

C.11

D.15

12.在自然语言处理中，实体关系学习是指\_\_\_\_。

A.发现文本中的有效实体

B.实体识别与链接

C.研究如何从文本中抽取事件信息并以结构化的形式呈现出来

D.检测文本中的实体是否具有某种预定义的关系

13.在文本挖掘中，可以使用\_\_\_\_命令完成将文本转换为tokens，然后将其转换为整数或浮点向量的操作。

A.CountVectorizer

B.TF-IDF

C.词袋模型（Bag of Words）

D.NERs

14.主题模型是自然语言处理中的常见的一类统计模型。对于常用的两类主题模型BTM、LDA，下列\_\_\_\_不是BTM相较于LDA模型的优点。

A.计算速度快

B.适用于短文本

C.占用空间少

D.解决稀疏性问题

15.以下关于神经网络自然语言模型的评述，错误的是\_\_\_\_\_。

A.BERT模型可用作特征提取

B.神经网络语言模型可解决维数灾难的问题

C.GPT和ELMo使用整个模型的全部参数来表示词向量

D.BERT是一个双向注意力模型

**二、多选题 ( 本大题 6 道小题 ，每小题 2 分，共 12 分），从下面题目给出的A、B、C、D四个可供选择的答案中选择所有正确答案。**

1.关于朴素贝叶斯分类方法描述正确的有\_\_\_\_。

A.需要计算先验概率

B.对缺失数据敏感

C.对小规模的数据表现很好

D.算法成立的前提是假设各属性之间互相独立

2.正则化能处理过拟合的原因是\_\_\_\_\_\_\_\_。

A.惩罚了模型的复杂度，避免模型过度学习训练集，提高泛化能力

B.正则项降低了每一次系数w更新的步伐，使参数更小，模型更简单

C.正则化使得训练集和测试集数据分布更为接近，因而避免了在训练集上过拟合

D.贝叶斯学派的观点，认为加入了先验分布（l1拉普拉斯分布，l2高斯分布），减少参数的选择空间

3.通过监督学习进行二分类模型训练过程中，可能会遇到正负样本分布不均的情况(比如正样本有50万但是负样本有100万)，下列\_\_\_\_方法可以进行恰当处理。

A.将所有的数据加入训练集，充分利用所有数据

B.从100万负样本中随机抽取50万

C.正样本权重设置为2，负样本权重设置为1

D.复制两份正样本参与到训练中去

4.可以实现关键词提取任务的常见算法包括( )。

A.TF/IDF

B.SVM

C.TextRank

D.最短编辑距离

5.对语义关系进行分类可以帮助我们更好理解语义关系的含义及特性，语义关系包括（ ）。

A.因果关系

B.上下位关系

C.部分整体

D.实体来源

6.评价一个推荐系统输出的Top-N推荐列表的好坏，常见的指标有:（ ）。

A.准确率Precision

B.召回率Recall

C.F1值

D.转化率

**三、是非题 ( 本大题 15 道小题 ，每小题 1 分，共 15 分），从下面题目给出的两个可供选择的答案中选择一个正确答案。**

1.如果决策树对训练集拟合不足，通过缩放输入特征可以有效改善。

A.正确

B.错误

2.Boosting和Bagging都是组合多个分类器投票的方法，二者都是根据单个分类器的正确率决定其权重。

A.正确

B.错误

3.如果两个变量的 Pearson 相关性系数为零，则它们不相关。

A.正确

B.错误

4.当不知道数据所带标签时，可以使用分类技术促使带同类标签的数据与带其他标签的数据相分离。

A.正确

B.错误

5.回归问题和分类问题都有可能发生过拟合。

A.正确

B.错误

6.向量x=[1,2,3,4,-9,0]的L1范数是1。

A.正确

B.错误

7.神经网络中激活函数引入了非线性。

A.正确

B.错误

8.参数化的方法可以使得类条件概率估计简化，但是估计结果的准确性严重依赖于所假设的概率分布形式是否符合真实数据分布。

A.正确

B.正确

9.假设我们有三个簇中心μ1=[1;2]，μ2=[-3 ;0];，μ3=[4 ;2]。此外，我们还有一个训练示例x(i)=[-2;1]。则在一个集群分配步骤之后，c(i)将会是2。

A.正确

B.错误

10.评价规则优劣的标准应该优先考虑规则准确率，同时考虑覆盖样例数和属性次序。

A.正确

B.错误

11.逆文档频率（IDF）的作用是减少常用词的权重，增加文档中不常用词的权重。

A.正确

B.错误

12.将词表示成向量被称为神经词嵌入（Neural Word Embeddings）。

A.正确

B.错误

13.信息抽取是指对源语言的长文本进行压缩，提取出关键句子的短文本的技术。

A.正确

B.错误

14.在给到任何神经网络之前，Tokens都会被转换成数字。

A.正确

B.错误

15.Word2Vec包含CBOW和Skip-gram这两个模型，其中CBOW模型是输入当前词的词向量，输出周围词的词向量，而Skip-gram模型正好相反。

A.正确

B.错误

**四、操作题**

**素材、样张、KS目录均在zip文件中，可双击此图标打开**

**以下第（一）题（案例应用题）题目请在文件“C:\KS\人工智能-自然语言处理与理解-答题纸.docx”中作答。**

**（一）案例应用题(共20分)**

虽然近年来智能对话系统取得了长足的进展，但是针对专业性较强的问答系统，如何准确的判别用户的输入是否为给定问题的语义等价问法仍然是智能问答系统的关键。举例而言，“请问什么是公益诉讼？”和“什么是公益诉讼？”可以认为是语义上等价的问题，而“请问什么是公益诉讼？”和“检察机关发起公益诉讼是什么意思？”则为不等价的问题。针对问题等价性判别而言，除去系统的准确性外，系统的鲁棒性也是很重要。

问题：现有某人工智能公司开发的自动问答系统用于某市法院的微信服务号，自动回答市民关于法律方面的常见问题。假设收集的提问数据集格式如下，数据以问题组的形式提供，每组问句又分为等价部分和不等价部分，等价问句之间互相组合可以生成正样本，等价问句和不等价问句之间互相组合可以生成负样本。请设计一个方案，用于预测一组文本是否为等价问法。

提示：这道题可能会有多种思路，建议使用基于BERT预训练的文本相似度预测思路。

|  |
| --- |
| 请在答题纸作答，此处答题一律无效！ |

|  |
| --- |
| <?xml version="1.0" encoding="utf-8"?> |
| <TrainCorpus> |
| <Questions number="0"> |
| <EquivalenceQuestions> |
| <question>哪些情形下，不予受理民事诉讼申请？</question> |
| <question>民事诉讼中对哪些情形的起诉法院不予受理</question> |
| <question>人民法院不予受理的民事案件有哪些情形？</question> |
| </EquivalenceQuestions> |
| <NotEquivalenceQuestions> |
| <question>民事诉讼什么情况下不能立案</question> |
| <question>哪些案件会给开具民事诉讼不予立案通知书</question> |
| <question>法院对于哪些案件再审申请不予受理</question> |
| </NotEquivalenceQuestions> |
| </Questions> |
| <Questions number="1"> |
| <EquivalenceQuestions> |
| <question>我被车撞了，肇事车是借的，我能找车主索赔吗？</question> |
| <question>借车撞人了，被撞者能找车主去索赔吗?</question> |
| <question>我被人家车撞了，可以去找原车主索要赔偿呢</question> |
| </EquivalenceQuestions> |
| <NotEquivalenceQuestions> |
| <question>交通事故未划分责任允许放车吗</question> |
| <question>人被车撞了,对方车是借的,对方全责,我怎么起诉</question> |
| <question>我的亲人被车撞如何索赔？</question> |
| <question>对方借车给我车撞了 对方全责</question> |
| </NotEquivalenceQuestions> |
| </Questions> |
| ... |
| </TrainCorpus> |