2023年上海市高等学校信息技术水平考试试卷

四级 人工智能——自然语言处理与理解（模拟卷）

（本试卷考试时间 150 分钟）

一、单选题（ 本大题 18 道小题 ，每小题 1 分，共 18 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。

1.两位同事从上海出发前往深圳出差，他们在不同时间出发，搭乘的交通工具也不同，能准确描述两者“上海到深圳”距离差别的是_____。

A.欧式距离

B.余弦距离

C.曼哈顿距离

D.切比雪夫距离

2.SVM（支持向量机）与LR（逻辑回归）在数学本质上的区别是_____。

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

3._____不能防止过拟合。

A.交叉验证

B.低维嵌入

C.剪枝

D.集成学习

4.一个计算机程序从经验E中学习任务T，并用P来衡量表现。并且，T的表现P随着经验E的增加而提高。假设我们给一个学习算法输入了很多历史天气的数据，让它学会预测天气。_____是P的合理选择。

A.计算大量历史气象数据的过程

B.天气预报任务

C.正确预测未来日期天气的概率

D.其余选项皆不是

5.关于PCA（主成分分析）算法和KPCA（核主成分分析）算法的说法中，正确的是_____。

A.PCA所做的是对坐标轴线性变换，即变换后的新基是一条曲线

B.KPCA对坐标轴做了线性变换，数据所映射的新基是一条直线

C.PCA与KPCA都可以对数据进行降维

D.KPCA算法应用到了核函数，所以在计算量上有一定减小

6.以下关于VC维（Vapnik-Chervonenkis Dimension）的描述中，正确的是_____。

A.在一个假设空间H中，如果d_vc(H)是无穷大，则存在VC bound

B.在一个假设空间H中，如果d_vc(H)是有限的，则不存在VC bound

C.模型较为复杂时，d_vc 通常较小
D.VC 维反应了模型的学习能力，VC 维越大，则模型的容量越大

7.以下关于梯度下降法的描述，错误的是_____。
A.随机梯度下降算法每次基于一个样本来更新梯度值，这导致更新过程波动较大，收到噪声影响较大
B.批量梯度下降算法每次基于全部样本来更新梯度值，这使得其更新速度较慢，但是更新过程较为平稳
C.小批量梯度下降算法每次选取一个 Batch Size 的数据样本进行梯度值更新，它综合了批量梯度下降与随机梯度下降的优缺点，是常用的梯度下降方法
D.批量梯度下降方法在大数据集上的收敛速度最快

8.某工程师在构建机器学习模型时发现模型在训练集和测试集上表现都很差，可能的原因是_____。
A.模型过于复杂，但是训练集比较小
B.正则化参数中惩罚项系数取值太小
C.测试集样本数量太少
D.模型训练时间过长

9.在进行分类模型建模时，如果样本类别不均衡，处理方式错误的是_____。
A.过采样小样本，欠采样大样本
B.丢弃小样本类别
C.通过集成方法解决
D.采集新的数据

10.A 工厂生产配件的合格率是 97%，B 工厂生产配件的合格率是 96%，C 公司从 A 工厂采购了 60%的配件，从 B 工厂采购了 40%的配件，现抽检到一个不合格的配件，该配件是 A 工厂生产的概率是_____。
A.52.9%
B.43.9%
C.67%
D.47.3%

11.____作为 NLP 领域最经典的使用场景之一，是指将自然语言的文本(例如，网页，电子邮件，新闻，公司文档等)归入到预定义的语义类别中。
A.命名实体识别
B.信息抽取
C.自动摘要
D.文本分类

12.关于 N-gram 模型，描述正确的是____。
A.N 越大，区分能力越强；N 越小，参数估计的可靠性越低
B.N 越大，区分能力越弱；N 越小，参数估计的可靠性越高
C.N 越大，区分能力越强；N 越小，参数估计的可靠性越高

D.N 越大，区分能力越弱；N 越小，参数估计的可靠性越低

13.在大规模的语料中，挖掘词的相关性是一个重要的问题。____不能用于确定两个词的相关性。
A.互信息
B.最大熵
C.卡方检验
D.最大似然比

14.在处理自然结构的新闻性句子的时候，_____是基于语法的文本句法分析方法，可以用于名词短语检测、动词短语检测、主语检测和宾语检测。
A.部分语音标注
B.依存句法分析和选取句法分析
C.Skip Gram 和 N-Gram 提取
D.连续性词包

15.在 LSTM 中使用的 Tahn 激活函数的模块结构是____。
A.遗忘门
B.输入门
C.输出门
D.生成候选记忆

16.关于 CBOW、Skip-gram、gloVe 的描述，错误的是____。
A.Skip-gram, CBOW 可以良好的体现词与词之间的语义相似性
B.CBOW 的输入是特定的一个词的词向量，而输出是特定词对应的上下文词向量
C.gloVe 模型除了可以体现出语义相似性，还可以体现词的全局统计特征
D.使用 negative sampling 训练 Skip-gram 可以提高训练速度

17.语料库中共有 100 篇文档，其中有含有“蜜蜂”的文档有 33 篇，文档 A 中共有 10 个词，其中“蜜蜂”出现了 3 次，则文档 A 在中“蜜蜂”的 TF-IDF 值为____。
A.log(100/34)*3/10
B.log(100/33)*3/10
C.log(33/100)*3/10
D.log(34/100)*3/10

18.某课程推荐系统可以根据学生的兴趣推荐选修课。选修课的一个重要属性是其所属的学科类别，例如：哲学、理学、工学、法学等。将这些学科类别用一个整数编码，用于计算学生的兴趣与课程的相似度。如下编码方式，最为合理的是____。
A.0-哲学，1-理学，2-工学，3-法学…
B.1-哲学，2-理学，3-工学，4-法学…
C.0-哲学，2-理学，4-工学，8-法学…
D.1-哲学，2-理学，4-工学，8-法学…

二、多选题（ 本大题 10 道小题 ，每小题 2 分，共 20 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.关于神经网络，说法正确的是_____。
A.增加网络层数，可能会增加测试集分类错误率
B.增加网络层数，一定会增加训练集分类错误率
C.减少网络层数，可能会减少测试集分类错误率
D.减少网络层数，一定会减少训练集分类错误率

2.属于稀疏表示应用领域的有_____。
A.图像去噪
B.压缩感知
C.人脸识别
D.目标跟踪

3.关于集成学习的说法中，正确的是_____。
A.Bagging 是通过结合几个模型降低泛化误差的技术。主要思想是分别训练几个不同的模型，然后让所有模型表决测试样例的输出
B.Boosting 的代表算法是随机森林
C.Boosting 是一种框架算法,主要是通过对样本集的操作获得样本子集,然后用弱分类算法在样本子集上训练生成一系列的基分类器
D.Stacking 是将各个弱学习器的学习成果，并行结合起来，形成以预测值（标签）为数据的训练集，用来训练下一层学习器

4.关于决策树算法的描述，正确的是_____。
A.构建决策树时可以选择信息熵与 GINI 系数作为分割指标
B.构建决策树时，数据集分割前与分割后的"纯度"差异越小，决策树越好
C.决策树既可以处理分类问题，也可以处理回归问题
D.不同的决策树算法差异点主要在特征选择过程中

5.关于不同的模型评价指标，正确的是_____。
A.对于样本类别分布均衡的分类问题，准确率是有意义的
B.回归问题可以采用 MSE 作为评价指标
C.F1 值综合了准确率和召回率
D.召回率越高意味着模型性能越好

6.文本语料库的特征包括____。
A.文本中词计数
B.词的向量标注
C.词性标注
D.基本依存语法

7.协同过滤算法经常被用于推荐系统，包含基于内存的协同过滤、基于模型的协同过滤以及混合模型。关于协同过滤，说法错误的是____。
A.基于内存的协同过滤可以较好解决冷启动问题
B.基于内存的协同过滤实现比较简单，新数据可以较方便的加入

C.基于模型的协同过滤不需要 item 的内容信息
D.基于模型的协同过滤能比较好的处理数据稀疏的问题

8.关于 HMM（隐马尔科夫模型）和 CRF（条件随机场），正确的是____。
A.HMM 是生成模型，CRF 是判别模型
B.HMM 是概率有向图，CRF 是概率无向图
C.HMM 求解过程可能是局部最优，CRF 可以全局最优
D.HMM 与 CRF 都具有三个基本问题：概率计算问题、预测问题和学习问题

9.Transformer 和 LSTM 是 NLP 中的常用模型。Transformer 相比较于 LSTM 等循环神经网络模型的优点有____。
A.采用门机制
B.可以直接捕获序列中的长距离依赖关系
C.模型并行度高，使得训练时间大幅度降低
D.适合用于处理与时间序列高度相关的问题

10.关于 Seq2Seq 的说法正确的是____。
A.Seq2Seq 通过深度神经网络将一个输入的序列映射为一个作为输出的序列
B.编码器和解码器各由一个循环神经网络组成
C.在 Seq2Seq 中，两个模型是串行训练的
D.Seq2Seq 要求输出的长度定长

三、是非题（本大题 23 道小题，每小题 1 分，共 23 分）。

1.xgboost 是一种优秀的集成算法，其优点包括速度快，对异常值不敏感，支持自定义损失函数等。

2.线性回归的自变量和残差不一定保持相互独立。

3.L1 正则和 L2 正则的共同点是都会让数据集中的特征数量减少。

4.过拟合是有监督学习的挑战，而不是无监督学习。

5.对于一个 SVM（支持向量机），去除不支持的向量后仍然能分类。

6.给定 n 个数据点，如果其中一半用于训练，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 的增加而减少。

7.随机梯度下降（Stochastic Gradient Descent）算法是用小规模的样本近似估计梯度的方法，适合在大规模数据上训练深度神经网络，但在逻辑回归、SVM 等算法中的作用很有限。

8.某个神经网络，其激活函数是 ReLU。若使用线性激活函数代替 ReLU，该神经网络仍然能表征 XNOR 函数。

9.ID3 决策树学习算法以信息增益为准则来选择划分属性；C4.5 决策树算法使用增益率来选择最优划分属性；CART 决策树使用 Gini（基尼）指数来选择划分属性。

10.梯度下降法可能会陷于局部极小值，EM 算法则不会。

11.深度学习与机器学习算法之间的区别在于，后者无需进行特征提取。

12.每次使用 K-means 聚类算法得到的聚类结果可能会不一样。

13.在某神经网络的隐藏层输出中，包含-1.5，那么该神经网络采用的激活函数不可能是 sigmoid、tanh、relu。

14.在线商品 A 的用户点击率为 1%，用某模型进行点击预测，得到了 99%的预测准确率，则该模型的预测准确率很高，可以投入使用。

15.强化学习理论受到行为主义心理学启发，侧重在线学习并试图在探索-利用（exploration-exploitation）间保持平衡。不同于监督学习和非监督学习，强化学习不要求预先给定任何数据，而是通过接收环境对动作的奖励（反馈）获得学习信息并更新模型参数。

16.Self-attention 的特点是无视词（token）之间的距离，直接计算依赖关系，从而能够学习到词语序列的内部结构。

17.Word2vec 模型是一种用于给文本目标创建矢量标记的机器学习模型。Word2vec 包含多个深度神经网络。

18.通常情况下，用户点击广告意味着用户对广告感兴趣。CTR（Click-Through Rate）是衡量用户点击率的指标，通过机器学习等方法，智能广告系统可以通过用户行为预测用户对广告的偏好，从而提升 CTR。

19.在统计语言模型中，通常以概率的形式描述任意语句的可能性，利用最大相似度估计进行度量。然而，对于一些低频词，无论如何扩大训练数据量，低频词出现的频度仍然很低，采用数据平滑技术可以解决这一问题。

20.知识图谱本质上是语义及关系网络，是一种基于图的数据结构语义知识库。

21.用余弦相似度表示的词之间的差异将显著高于 0.5。

22.TF-IDF（Term Frequency-Inverse Document Frequency）是一种常见的文本检索统计方法，用于评估一个词条对一个文档的重要程度。一个词条的 TF-IDF 与它在文档中出现的次数成正比，与它在全部语料库中出现的频率成反比。

23.每一篇文章都有各自的主题分布，该主题分布应该服从多项式分布，每个主题都有各自的词分布，其词分布应该服从多项式分布。

四、操作题

以下试题（案例应用题）题目请在文件"C:\KS\人工智能-自然语言处理-答题纸.docx"中作答！

知识图谱技术已在众多领域发挥积极的作用。例如，在网上购物时，消费者往往会根据自己的购物体验与实际使用情况，从商品本身、竞品对比、服务态度、物流体验、品牌活动好感度等多个方面对本次购物进行评价。在充分保证用户个人信息与隐私安全，严格遵守法律法规的前提下，基于公开评论信息，快速有效地识别、统计、分析各方面评价反馈的好坏程度，从而使得厂家能够及时针对性改进产品、店铺能够有效优化服务环节、品牌方能够更合理得策划活动，最后实现客户满意度提升、购买意愿加强的良性循环。

从评论中关键信息的挖掘、信息间关联性的分析再到其评价好坏程度的判断，这是一个典型的知识图谱任务应用，需要实体识别、关系识别、情感识别的一系列流程进行配合。

问题：

1）请分析在该知识图谱任务中，实体识别、关系识别、情感识别在该任务中分别解决什么问题？

请在答题纸作答，此处答题一律无效！

2）在该类任务的训练中，数据样本不均衡是一个常见的问题，请简述两种解决数据不均衡的方法。

请在答题纸作答，此处答题一律无效！

3）Bert在上述三项任务中都发挥了重要的作用，请问Bert预训练时的两种下游任务是什么？Bert采用何种Normalization（归一化）技术？

请在答题纸作答，此处答题一律无效！

4）请列举知识图谱技术的其他任务场景，并简单描述其中一个任务场景的知识图谱任务流程。

请在答题纸作答，此处答题一律无效！