美丽联合2018校招算法工程师笔试试卷

一. 单项选择题

1. 使用以下模型解决分类问题时，一般情况下对输入数据是否进行归一化最不敏感的是?

A 软间隔SVM
B Lasso
C 深度神经网络
D 决策树

2. 使用以下模型解决分类问题时，一般情况下对输入数据内的噪音、离群点(outlier)最敏感的是?

A Adaboost
B 软间隔SVM
C Lasso
D 随机森林

3. 对于神经元q = x + y，f = q * z，当输入x=-2, y=5, z=-4时进行后向传播，则f对(x, y, z)的梯度分别为

A (-3, 4, 4)
B (-4, 4, 3)
C (-4, -4, 3)
D (3, -4, -4)

4. 请问以下和神经网络中的dropout作用机制类似的是?

A Bagging
B Boosting
C Stacking
D 都不是

5. 请问以下能为神经网络引入了非线性能力的是?

A 随机梯度下降
B ReLU
C 卷积
D 以上都不是

6.
假设以交叉熵为损失函数，则以下为其梯度的是?

$$\frac{1}{N}\sum_{i=1}^{N} ln(1+exp(-y_i w^T x_i))$$

A <img alt=""
src="https://uploadfiles.nowcoder.net/images/20170920/300823_1505895468374_701CCAF6EC164

B <img alt=""
src="https://uploadfiles.nowcoder.net/images/20170920/300823_1505895483831_5753F64FD8EED

C <img alt=""
src="https://uploadfiles.nowcoder.net/images/20170920/300823_1505895489730_EEE71F518F7405

D <img alt=""
src="https://uploadfiles.nowcoder.net/images/20170920/300823_1505895495841_54720E1A096141

7. 已知二叉树的先序遍历序列为ABCDEFI，中序遍历序列为CBDAEFI，则该二叉树的高度为?

A 2

B 3

C 4

D 5

二. 多选选择题

8.
在K近邻(KNN)模型中，超参数K的选择对模型的表现有较大的影响。一般而言，对比1近邻模型和3近邻模型，以下说法正确的是?

A 1近邻模型更容易过拟合

B 1近邻模型的Bias和Variance都更大

C 1近邻模型的Bias更大，Variance更小

D 1近邻模型的Bias更小，Variance更大

9. 假设我们有所有用户在一个月内的商品点击和购买行为的数据，我们想利用该数据基于有监督学习方法预测特定用户点击商品后产生购买行为的概率，即将购买率预估问题看做是一个二分类问题。假设用户的购买行为十分稀疏，即购买量远小于点击量。则以下说法正确的是?

A 这是一个正负类极度不平衡的数据集，因此AUC比较适合用于评估模型的效果

B 如果我们将模型在所有样本上的预估值都乘以2，则该模型模型的AUC有可能会下降(假设模型在所有样本上的预估值都小于0.5)

C 如果我们将模型在所有样本上的预估值都减去0.01，则该模型模型的AUC有可能会下降(假设模型在所有样本上的预估值都大于0.01)

D 我们可以将数据随机切分成10份，选择其中9份作为训练集，1份作为测试集，如此实现10重交叉检验可以较好地对模型的效果进行评估

10.
关于集成学习以下说法正确的是?

A Adaboost相对于单个弱分类器而言通过Boosting增大了模型的Bias

B 随机森林相对于单个决策树而言通过Bagging增大了模型的Variance

C 我们可以借鉴类似Bagging的思想对GBDT模型进行一定的改进，例如每个分裂节点只考虑某个随机的

特征子集或者每棵树只考虑某个随机的样本子集这两个方案都是可行的

D GBDT模型无法在树维度通过并行提速，因为基于残差的训练方式导致第i棵树的训练依赖于前i-1棵树的结果，故树与树之间只能串行

11.
如下数据集中，适合使用隐马尔科夫模型(HMM)建模的有?

A 基因序列集合

B 电影影评数据集合

C 股票市场数据集合

D 北京气温数据集合

12. 下列几个范数项可以为模型带来稀疏性的有：

A 向量内非零元素之和

B 向量内非零元素之和的平方

C 向量内所有元素之和

D 向量内所有元素之和的平方

13. 在深度神经网络的训练过程中发现损失函数值的变化非常缓慢，以下方案中很可能有助于解决该问题的是?

A 调小\learning rate

B 调大batch
size

C 使用ReLU

D 使用batch normaliztion

14.
以下关于排序算法的说法中错误的是?

A 快速排序的最坏时间复杂度为O(nlog(n))

B 快速排序不是一种稳定的排序

C 堆排序的最坏时间复杂度为O(nlog(n))

D 堆排序不是一种稳定的排序

三. 问答题

15. 请实现str2int函数，不限语言，但请不要用使用语言自带的类型转换。例如对于输入"123"，输出应为123

16.
假设有A、B两枚硬币，A硬币正面朝上的概率为T，B硬币正面朝上的概率为2T，按一定顺序抛了这两枚硬币之后，收集到如下样本：1) A硬币，正面 2) B硬币，反面 3) B硬币，反面 4) B硬币，反面 5) B硬币，正面。
a) 请写出上述样本关于T的log似然函数
b) 请给出对T的极大似然估计

17.

假设有一份数据量较大的数据，假设总共1亿行，现在要根据它的某个字段a进行全局降序排序，如果使用Map-Reduce来实现，程序大致应该怎么设计？要求：只要写出伪代码即可，注意避免控制数据多次分发和并行效率

18.
假设使用结构为(parent varchar(6), child varchar(6))的表来存储一颗树内所有子节点-父节点的关系。请使用SQL统计出该树内每个节点的儿子节点和孙子节点（即儿子的儿子）的数量之和。没有儿子且没有孙子的节点不应该出现在结果中。

样例输入
parent, child
A B
A C
A D
B E
B F
E G

样例输出
node, cnt
A 5
B 3