udesk机器学习岗

一、机器学习技术面

（1）讲讲做过的项目

（3）CNN的原理是什么？

卷积神经网络（Convolutional Neural Network，简称CNN），是一种前馈神经网络，人工神经元可以响应周围单元，可以进行大型图像处理。卷积神经网络包括卷积层和池化层。 卷积神经网络是受到生物思考方式启发的MLPs（多层感知器），它有着不同的类别层次，并且各层的工作方式和作用也不同

（4）调过什么参数？有改进吗？

机器学习模型中超级参数(hyperparameter)的调优问题（下文简称为调参问题），主要的方法有手动调优、网格搜索、随机搜索以及基于贝叶斯的参数调优方法。因为模型通常由它的超级参数确定，所以从更高的角度看调参问题就转化为模型选择问题。

如果训练误差和验证误差都停滞在一个很大的值上，那么可能的原因和可以尝试的解决方案：

```
欠拟合，采取增加模型容量的方法，如将weight decay 设为0
模型有bug，将训练数据集减小，再次训练看训练误差是否能减小
```

（5）过拟合问题怎么处理？

在对模型进行训练时，有可能遇到训练数据不够，即训练数据无法对整个数据的分布进行估计的时候，或者在对模型进行过度训练（overtraining）时，常常会导致模型的过拟合（overfitting）

为了防止过拟合，我们需要用到一些方法，如：early stopping、数据集扩增（Data augmentation）、正则化（Regularization）、Dropout等。

Early stopping

便是一种迭代次数截断的方法来防止过拟合的方法，即在模型对训练数据集迭代收敛之前停止迭代来防止过拟合。

数据机扩增

即需要得到更多的符合要求的数据，即和已有的数据是独立同分布的，或者近似独立同分布的。一般有以下方法：

```
从数据源头采集更多数据
复制原有数据并加上随机噪声
```

重采样

根据当前数据集估计数据分布参数，使用该分布产生更多数据等

Dropout

在神经网络中，有一种方法是通过修改神经网络本身结构来实现的，其名为Dropout。该方法是在对网络进行训练时用一种技巧（trick），对于如下所示的三层人工神经网络：

在训练开始时，随机删除一些（可以设定为一半，也可以为1/3，1/4等）隐藏层神经元，即认为这些神经元不存在，同时保持输入层与输出层神经元的个数不变，这样便得到ANN

（6）熟悉的机器学习算法，详细讲讲（讲了岭回归并推导）

岭回归是一种专用于共线性数据分析的有偏估计回归方法，实质上是一种改良的最小二乘估计法，通过放弃最小二乘法的无偏性，以损失部分信息、降低精度为代价获得回归系数更为符合实际、更可靠的回归方法，对病态数据的拟合要强于最小二乘法。

岭回归与最小二乘的区别在于这一项，称之为正则项，这一项可以看成是对A的各个元素，即各个特征的权的总体的平衡程度，也就是权之间的方差。

（7）聚类算法会哪些？（讲了k-means原理）

1. K-Means(K均值)聚类

算法步骤： (1) 首先我们选择一些类/组，并随机初始化它们各自的中心点。中心点是与每个数据点向量长度相同的位置。这需要我们提前预知类的数量(即中心点的数量)。 (2) 计算每个数据点到中心点的距离，数据点距离哪个中心点最近就划分到哪一类中。 (3) 计算每一类中中心点作为新的中心点。 (4) 重复以上步骤，直到每一类中心在每次迭代后变化不大为止。也可以多次随机初始化中心点，然后选择运行结果最好的一个。

2. 均值漂移聚类

均值漂移聚类是基于滑动窗口的算法，来找到数据点的密集区域。这是一个基于质心的算法，通过将中心点的候选点更新为滑动窗口内点的均值来完成，来定位每个组/类的中心点。然后对这些候选窗口进行相似窗口进行去除，最终形成中心点集及相应的分组。 具体步骤：

1. 确定滑动窗口半径r，以随机选取的中心点C半径为r的圆形滑动窗口开始滑动。均值漂移类似一种爬山算法，在每一次迭代中向密度更高的区域移动，直到收敛。
2. 每一次滑动到新的区域，计算滑动窗口内的均值来作为中心点，滑动窗口内的点的数量为窗口内的密度。在每一次移动中，窗口会想密度更高的区域移动。
3. 移动窗口，计算窗口内的中心点以及窗口内的密度，知道没有方向在窗口内可以容纳更多的点，即一直移动到圆内密度不再增加为止。
4. 步骤一到三会产生很多个滑动窗口，当多个滑动窗口重叠时，保留包含最多点的窗口，然后根据数据点所在的滑动窗口进行聚类。

目前就举出这两种聚类方法，详细请了解https://blog.csdn.net/Katherine_hsr/article/details/79382249

（8）分类算法会哪些？（讲了朴素贝叶斯）

在数据挖掘任务中通常分为两大类： 预测任务，根据其他属性的值，预测特定属性的值。 描述任务，概括数据中潜在联系的模式（相关性，趋势，聚类，轨迹和异常）

分类属于预测任务，就是通过已有数据集（训练集）的学习，得到一个目标函数f（模型），把每个属性集x映射到目标属性y（类），且y必须是离散的（若y为连续的，则属于回归算法） 分类过程首先需要将生活的数据处理成计算机可以理解的数据（通常为表）。阿里天池竞赛题目为例，已知客户行为信息，以及商品内容，预测推荐哪件商品给客户会被购买。 人的每一个行为都可以抽象成属性，是否购买过同类产品，买东西的频率是多少，从点进去网页到放进购物车的平均时间多少，从放入购物车到下单的时间多少，是否曾经把购物车的东西拿出来过，有无评论买过东西的习惯，有无退货习惯，买过最贵的东西是什么价位，最便宜是什么价位
https://blog.csdn.net/csdn595075652/article/details/51470415

（9）自然语言处理了解过哪些？（讲了词频逆文档频率矩阵）

1 Python 的几个自然语言处理工具

NLTK:NLTK 在用 Python 处理自然语言的工具中处于领先的地位。它提供了 WordNet 这种方便处理词汇资源的借口，还有分类、分词、除茎、标注、语法分析、语义推理等类库。 Pattern:Pattern 的自然语言处理工具有词性标注工具(Part-Of-Speech Tagger)，N元搜索(n-gram search)，情感分析(sentiment analysis)，WordNet。支持机器学习的向量空间模型，聚类，向量机。 TextBlob:TextBlob 是一个处理文本数据的 Python 库。提供了一些简单的api解决一些自然语言处理的任务，例如词性标注、名词短语抽取、情感分析、分类、翻译等等。 Gensim:Gensim 提供了对大型语料库的主题建模、文件索引、相似度检索的功能。它可以处理大于RAM内存的数据。作者说它是“实现无干预从纯文本语义建模的最强大、最高效、最无障碍的软件。 PyNLPI:它的全称是：Python自然语言处理库（Python Natural Language Processing Library，音发作: pineapple） 这是一个各种自然语言处理任务的集合，PyNLPI可以用来处理N元搜索，计算频率表和分布，建立语言模型。他还可以处理向优先队列这种更加复杂的数据结构，或者像 Beam 搜索这种更加复杂的算法。 spaCy:这是一个商业的开源软件。结合Python和Cython，它的自然语言处理能力达到了工业强度。是速度最快，领域内最先进的自然语言处理工具。 Polyglot:Polyglot 支持对海量文本和多语言的处理。它支持对165种语言的分词，对196中语言的辨识，40种语言的专有名词识别，16种语言的词性标注，136种语言的情感分析，137种语言的嵌入，135种语言的形态分析，以及69中语言的翻译。

可以针对NLTK做一个具体的讲解，讲一下NLTK做分词处理的逻辑

（10）在纸上当场写代码：二分查找（递归，while循环两种办法）；如果查找的列表里没有这个数怎么办？

二分查找又叫折半查找，二分查找应该属于减治技术的成功应用。所谓减治法，就是将原问题分解成若干个子问题后，利用了规模为n的原问题的解与较小规模（通常是n/2）的子问题的解之间的关系。 二分查找利用了记录按关键码有序的特点，其基本思想为：在有序表中，取中间记录作为比较对象，若给定值与中间记录的关键码相等，则查找成功；若给定值小于中间记录的关键码，则在中间记录的左半边继续查找；若给定值大于中间记录的关键码，则在中间记录右半边区继续查找。不断重复上述过程，直到查找成功，或所查找的区域无记录，查找失败。 二分查找的时间复杂度是O(log(n))，最坏情况下的时间复杂度是O(n)。

```
#!/usr/bin/python
#coding=utf-8

#自定义函数，实现二分查找，并返回查找结果
def binary_search(find, list1) :
  low = 0
  high = len(list1)
  while low <= high :
    mid = (low + high) / 2
    if list1[mid] == find :
      return mid
    #左半边
    elif list1[mid] > find :
      high = mid -1
    #右半边
    else :
      low = mid + 1
  #未找到返回-1
  return -1

list1 = [1,2,3,7,8,9,10,5]
#进行二分查找算法前必须保证要查找的序列时有序的，这里假设是升序列表
list1.sort()

print "原有序列表为:",list1
try :
  find = int(raw_input("请输入要查找的数："))
except :
  print "请输入正整数！"
  exit()

result = binary_search(find, list1)
if result != -1 :
  print "要找的元素%d的序号为：%d" %(find,result)
else :
  print "未找到！"
```

（11）对自然语言处理感兴趣吗？ （12）上一份工作薪资多少？期望薪资多少？ （13）有什么问题要问我的？

二、CTO （1）讲讲上一份工作 （2）公司一共几个人？ （3）你在项目组里做什么？ （4）你们技术总监什么背景？ （5）你在公司里参与的项目是怎么运作的？ （6）每两周周六加班一天，能接受吗？ （7）Python用得怎么样？ （8）在学校年级排名怎么样？ （9）有什么爱好？ （10）期望薪资？ （11）有什么要问我的？

三、Web组长 （1）Python用了多久了？ （2）用过什么其他语言？ （3）对比一下Java和python的区别？（java是强类型语言，类里方法的重载不同，python的封装是假的封装，python可以多继承。。。），python是一种弱类型语言，怎么来解释 （4）闭包是怎么回事？Python里的闭包是怎么实现的？ （5）一个函数有多少行代码合适？ （6）有什么问题要问我的？

四、人力 （1）上一份工作的公司有多少人？做技术的多少人？和你同岗位的多少人？薪资多少？ （2）为什么离职？为什么来北京？ （3）来上班路上要多久？九点半上班没问题吗？ （4）最近面试几

家了？感觉怎么样？（5）最早什么时间能入职？（6）对跳槽怎么看？觉得几年一跳合适？（7）自然语言处理是你的兴趣吗？（8）有什么问题要问我的？