蒋豆芽

特征工程 已购

来自 【机器学习面试题汇总与解析（蒋豆芽面试题总结）】 | 95 浏览 | 0 回复 | 2021-06-27

蒋豆芽

+关注

机器学习面试题汇总与解析——特征工程

1. **特征工程有哪些**☆☆☆☆☆
2. **遇到缺值的情况，有哪些处理方式**☆☆☆☆☆
3. **样本不均衡的处理办法**☆☆☆☆☆
4. **训练时样本不平衡问题如何解决；小样本问题如何解决**☆☆☆☆☆
5. **常见的筛选特征的方法有哪些？**☆☆☆☆☆
6. **数据怎么清洗，缺失值怎么填充**☆☆☆☆☆
7. **出现Nan的原因**☆☆☆☆☆
8. **特征筛选，怎么找出相似性高的特征并去掉**☆☆☆☆☆
9. **对于不同场景机器学习和深度学习你怎么选择，你更习惯机器学习还是深度学习？**☆☆☆☆☆
10. **包含百万、上亿特征的数据在深度学习中怎么处理**☆☆☆☆☆
11. **类别型数据你是如何处理的？比如游戏品类，地域，设备**☆☆☆☆
12. **计算特征之间的相关性方法有哪些？**☆☆☆☆

==========

- 本专栏适合于**Python已经入门**的学生或人士，有一定的编程基础。
- 本专栏适合于**算法工程师、机器学习、图像处理求职**的学生或人士。
- 本专栏针对面试题答案进行了**优化，尽量做到好记、言简意赅。这才是一份面试题总结的正确打开方式。这样才方便背诵**
- 如专栏内容有错漏，欢迎在评论区指出或私聊我更改，一起学习，共同进步。
- **相信大家都有着高尚的灵魂，请尊重我的知识产权，未经允许严禁各类机构和个人转载、传阅本专栏的内容。**

==========

关于**机器学习算法**书籍，我强烈推荐一本**《百面机器学习算法工程师带你面试》**，这个就很类似面经，还有讲解，写得比较好。私聊我进群。

蒋豆芽

参考资料

特征工程系列教程： https://blog.csdn.net/lc013/article/details/87898873 （写得很好，强烈推荐）

读者可以把参考文章看看

1. **特征工程有哪些**☆☆☆☆☆

   参考回答

   1. **数据预处理**
      1. **处理缺失值**
      2. **图片数据扩充**
      3. **处理异常值**
      4. **处理类别不平衡问题**
   2. **特征缩放**
      1. **归一化**
      2. **正则化**
   3. **特征编码**
      1. **序号编码(Ordinal Encoding)**
      2. **独热编码(One-hot Encoding)**
      3. **二进制编码(Binary Encoding)**
      4. **离散化**
   4. **特征选择**
      1. **过滤式(filter)**
      2. **包裹式(wrapper)**
      3. **嵌入式(embedding)**
   5. **特征提取**
      1. **降维**
      2. **图像特征提取**
      3. **文本特征提取**

蒋豆芽

答案解析

无。

类似的问题还有：

2. **遇到缺值的情况，有哪些处理方式** ☆☆☆☆☆

   参考回答

   1. **直接使用含有缺失值的特征**：当仅有少量样本缺失该特征的时候可以尝试使用；
   2. **删除含有缺失值的特征**：这个方法一般适用于大多数样本都缺少该特征，且仅包含少量有效值是有效的；
   3. **插值补全缺失值**
      1. 均值、众数、中位数、固定值、手动、最邻近补全
      2. 建模预测：回归、决策树
      3. 高维映射、压缩感知
      4. 多种方法插补

   答案解析

   无。

3. **样本不均衡的处理办法** ☆☆☆☆☆

   参考回答

   1. **扩充数据集**
   2. **尝试其他评价指标**
   3. **对数据集进行重采样**
      1. 对小类的数据样本进行采样来增加小类的数据样本个数，即**过采样**（over-sampling ，采样的个数大于该类样本的个数）
      2. 对大类的数据样本进行采样来减少该类数据样本的个数，即**欠采样**（under-sampling，采样的次数少于该类样本的个素）
   4. **尝试不同分类算法**：如**决策树**往往在类别不均衡数据上表现不错。
   5. **尝试对模型进行惩罚**：比如你的分类任务是识别那些小类，那么可以对分类器的小类样本数据增加**权值**，降低大类样本的**权值**（这种方法其实是产生了新的数据分布，即产生了新的数据集），从而使得分类器将重点集中在小类样本身上。如**focal loss**

无。

4. **训练时样本不平衡问题如何解决；小样本问题如何解决** ☆☆☆☆☆

   参考回答

   1. 扩充数据集，增加小类样本的数量。
   2. 针对小类样本进行**过采样**

   答案解析

   无。

5. **常见的筛选特征的方法有哪些?** ☆☆☆☆☆

   参考回答

   1. **过滤式(filter)**
   2. **包裹式(wrapper)**
   3. **嵌入式(embedding)**

   答案解析

   过滤式(filter)：先对数据集进行特征选择，其过程与后续学习器无关，即设计一些统计量来过滤特征，并不考虑后续学习器问题。如**方差选择**、**卡方检验**、**互信息**

   包裹式(wrapper)：实际上就是一个分类器，它是将后续的学习器的性能作为特征子集的评价标准。如**Las Vagas 算法**

   嵌入式(embedding)：实际上是学习器自主选择特征。如**基于惩罚项的选择**、**基于树的选择 GBDT**

6. **数据怎么清洗，缺失值怎么填充** ☆☆☆☆☆

   参考回答

   略

   答案解析

   无。

蒋豆芽

1. NaN的含义是没有意义的数，**not a number**，一般有这几种情况：0/0，Inf/Inf，Inf-Inf，Inf*0等，都会导致结果不确定，所以会得到NaN
2. 数据处理时，在实际工程中经常数据的**缺失**或者**不完整**，此时我们可以将那些缺失设置为nan
3. 读取数据时，某个**字符**不是数据，那么我们将它认为nan处理。

答案解析

无。

8. **特征筛选，怎么找出相似性高的特征并去掉**☆☆☆☆☆

参考回答

特征选择---过滤法（特征相关性分析）：可以采用**方差选择法**或**相关系数法。**

答案解析

无。

9. **对于不同场景机器学习和深度学习你怎么选择，你更习惯机器学习还是深度学习？**☆☆☆☆☆

参考回答

1. 当**数据量小**时，深度学习算法表现不佳。可采用**机器学习算法。**
2. 深度学习依赖于高端设备，而传统学习依赖于低端设备。如果成本限制严格，可采用**机器学习算法。**
3. 如果对时间要求苛刻，可采用**机器学习算法。**
4. 针对特定领域，如图像、视频流，深度学习在精度上更优异，可采用**深度学习**的方法。

答案解析

无。

10. **包含百万、上亿特征的数据在深度学习中怎么处理**☆☆☆☆☆

参考回答

这么多的特征，肯定不能直接拿去训练，**特征多，数据少**，很容易导致模型**过拟合。**

1. **降维**：PCA或LDA

蒋豆芽

4. **特征选择**：去掉不重要的特征

答案解析

无。

11. **类别型数据你是如何处理的？比如游戏品类，地域，设备** ☆☆☆☆

参考回答

序号编码、one-hot编码、二进制编码

答案解析

无。

12. **计算特征之间的相关性方法有哪些？** ☆☆☆☆

参考回答

1. **pearson系数**，对定距连续变量的数据进行计算。是介于-1和1之间的值
2. **Spearman秩相关系数**：是度量两个变量之间的统计相关性的指标，用来评估当前单调函数来描述俩个变量之间的关系有多好。
3. **Kendall（肯德尔等级）相关系数**：肯德尔相关系数是一个用来测量两个随机变量相关性的统计值。

答案解析

无。

资源分享 python 机器学习 算法工程师 春秋招 面试题 软件开发 面经

蒋豆芽

没有回复

请留下你的观点吧~

发布

牛客博客，记录你的成长

关于博客 | 意见反馈 | 免责声明 | 牛客网首页