|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 为什么要对特征做归一化   |  | | --- | | **1.特征间的单位(尺度)可能不同。**比如身高和体重，比如摄氏度和华氏度，比如房屋面积和房间数，一个特征的变化范围可能是[1000, 10000]，另一个特征的变化范围可能是[−0.1,0.2]，在进行距离有关的计算时，单位的不同会导致计算结果的不同，尺度大的特征会起决定性作用，而尺度小的特征其作用可能会被忽略，**为了消除特征间单位和尺度差异的影响，以对每维特征同等看待，需要对特征进行归一化**。  2.原始特征下，**因尺度差异，其损失函数的等高线图可能是椭圆形**，梯度方向垂直于等高线，下降会走zigzag路线，而不是指向local minimum。通过对特征进行zero-mean and unit-variance变换后，其损失函数的等高线图更接近圆形，梯度下降的方向震荡更小，**收敛更快**。  https://mmbiz.qpic.cn/sz_mmbiz_png/gYUsOT36vfrVibKMK0MzlEQcPX1tF1EeG5OibQyOm1a75mD6vrIDuehSjbARtXcVb0ENVC8GA3ORJ406a27GghcQ/640?wx_fmt=png&tp=webp&wxfrom=5&wx_lazy=1&wx_co=1  a.涉及或隐含**距离计算**的算法，比如K-means、KNN、PCA、SVM等，一般需要feature scaling。  b.损失函数中含有**正则项**时，一般需要feature scaling。  c. **梯度下降算法，需要feature scaling**  d. 与距离计算无关的概率模型，不需要feature scaling，比如Naive Bayes | | 理解清楚特征归一化所适用的模型场景 |
| 2 | 什么是组合特征？如何处理高维组合特征？   |  | | --- | | 为了提高复杂关系的拟合能力，在**特征工程中经常会把一阶离散特征两两组合，构成高阶组合特征**。以广告点击预估问题为例，原始数据有语言和类型两种离散特征，表1.2是语言和类型对点击的影响。为了提高拟合能力，语言和类型可 以组成二阶特征，表1.3是语言和类型的组合特征对点击的影响  在这里插入图片描述  以逻辑回归为例，假设数据的特征向量为X=(x1,x2,…,xk)，则有  在这里插入图片描述  其中<xi, xj>表示xi和xj的组合特征，wij的维度等于|xi|·|xj|，|xi|和|xj|分别代表第i个特征 和第j个特征不同取值的个数。在表1.3的广告点击预测问题中，w的维度是 2×2=4（语言取值为中文或英文两种、类型的取值为电影或电视剧两种）。这种特 征组合看起来是没有任何问题的，但当引入ID类型的特征时，问题就出现了。以 推荐问题为例，表1.4是用户ID和物品ID对点击的影响，表1.5是用户ID和物品ID的 组合特征对点击的影响。  在这里插入图片描述  在这种情况下，一种有效办法就是将用户和物品分别用k维的低维向量表示(k<<m,k<<m)。这其实**等价于矩阵分解**，所以，这里也提供了另一个理解推荐系统中矩阵分解的。 | | 这里的特征组合主要指的是类别特征(Categorical Feature)之间的组合 |
| 3 | 请比较欧式距离与曼哈顿距离？   |  | | --- | | **欧氏距离就是我们最常用的两点之间的直线距离。**以二维空间为例，两点(x1,y1),(x2,y2)之间的欧式距离为：  在这里插入图片描述  **曼哈顿距离则表示两个点在标准坐标系上的绝对轴距之和。**还是以二维空间为例，两点(x1,y1),(x2,y2)之间的曼哈顿距离为：  在这里插入图片描述  用一张图来区分一下两者：  数据分析基础：距离度量方式（欧式距离、马氏距离、曼哈顿距离）  图中绿线是欧氏距离，红线是曼哈顿距离，蓝线和黄线等价于曼哈顿距离。 | | 比较曼哈顿距离和欧式距离的数值特点，并结合一两个具体例子做分析 |
| 4 | 为什么一些场景中使用余弦相似度而不是欧式距离   |  | | --- | | **余弦相似度**：取值范围[-1,1] **余弦距离**=1-余弦相似度：取值范围[0,2]   * 余弦相似度在高维的情况下依然保持“相同时为1，正交时为0，相反时为-1”的性质。 * **欧式距离的数值受维度的影响**，范围不固定，并且含义也比较模糊。 * 欧式距离体现数值上的绝对差异，而**余弦距离体现方向上的相对差异**。 | | 比较余弦相似度和欧式距离的数值特点，并结合一两个具体例子做分析 |
| 5 | One-hot的作用是什么？为什么不直接使用数字作为表示   |  | | --- | | One-Hot编码是分类变量作为二进制向量的表示。这首先要求将分类值映射到整数值。然后，每个整数值被表示为二进制向量，除了整数的索引之外，它都是零值，它被标记为1。举个例子，假设我们有四个样本（行），每个样本有三个特征（列），如图：  https://images2017.cnblogs.com/blog/1251096/201710/1251096-20171030163200996-742440926.png  one-hot编码就是保证每个样本中的单个特征只有1位处于状态1，其他的都是0。  https://images2017.cnblogs.com/blog/1251096/201710/1251096-20171030164625793-1703194224.png  相对于数字编码，One-hot编码后，能更方便于计算机处理，且表达容量更大。 | | 理解清楚并比较One-hot编码和数字编码的特点 |