

1	<p>为什么要对特征做归一化</p> <p>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。</p> <p>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。</p> <div data-bbox="304 763 1118 1211"> <p>Feature Scaling Idea: Make sure features are on a similar scale.</p> <p>E.g. $x_1 = \text{size (0-2000 feet}^2\text{)}$ \leftarrow $\rightarrow x_1 = \frac{\text{size (feet}^2\text{)}}{2000}$</p> <p>$x_2 = \text{number of bedrooms (1-5)}$ \leftarrow $\rightarrow x_2 = \frac{\text{number of bedrooms}}{5}$</p> </div> <p>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。</p> <p>b. 损失函数中含有正则项时，一般需要 feature scaling。</p> <p>c. 梯度下降算法，需要 feature scaling</p> <p>d. 与距离计算无关的概率模型，不需要 feature scaling，比如 Naive Bayes</p>	<p>理解清楚特征归一化所适用的模型场景</p>
---	--	--------------------------

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

表1.2 语言和类型对点击的影响

是否点击	语言	类型
0	中文	电影
1	英文	电影
1	中文	电视剧
0	英文	电视剧

表1.3 语言和类型的组合特征对点击的影响

是否点击	语言=中文 类型=电影	语言=英文 类型=电影	语言=中文 类型=电视剧	语言=英文 类型=电视剧
0	1	0	0	0
1	0	1	0	0

以逻辑回归为例，假设数据的特征向量为 $X=(x_1, x_2, \dots, x_k)$ ，则有

$$Y = \text{sigmoid}(\sum_i \sum_j w_{ij} \langle x_i, x_j \rangle)$$

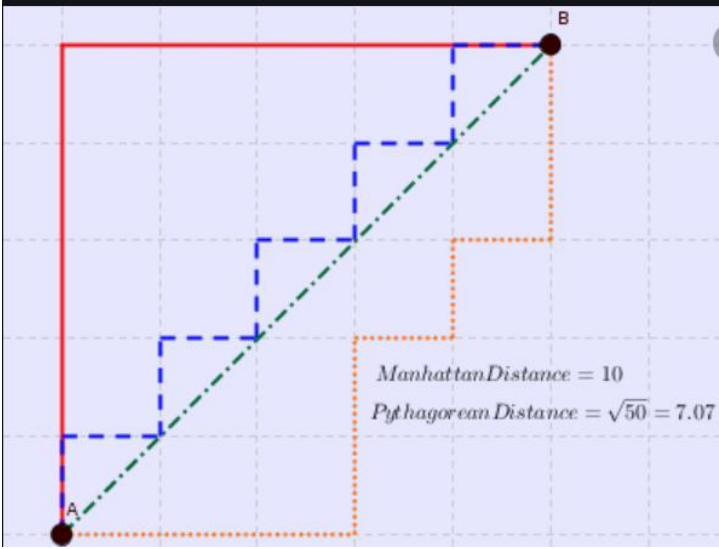
其中 $\langle x_i, x_j \rangle$ 表示 x_i 和 x_j 的组合特征， w_{ij} 的维度等于 $|x_i| \cdot |x_j|$ ， $|x_i|$ 和 $|x_j|$ 分别代表第 i 个特征 和第 j 个特征不同取值的个数。在表 1.3 的广告点击预测问题中， w 的维度是 $2 \times 2 = 4$ （语言取值为中文或英文两种、类型的取值为电影或电视剧两种）。这种特征组合看起来是没有任何问题的，但当引入 ID 类型的特征时，问题就出现了。以 推荐问题为例，表 1.4 是用户 ID 和物品 ID 对点击的影响，表 1.5 是用户 ID 和物品 ID 的 组合特征对点击的影响。

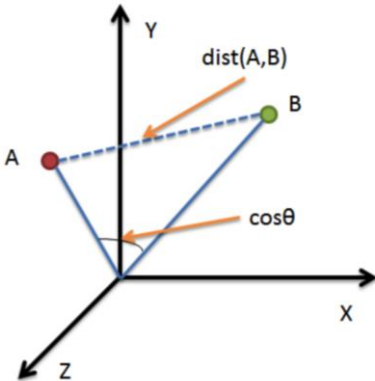
表1.5 用户ID和物品ID的组合特征对点击的影响

是否点击	用户ID=1 物品ID=1	用户ID=2 物品ID=1	...	用户ID=m 物品ID=1	用户ID=1 物品ID=2	用户ID=2 物品ID=2	...	用户ID=m 物品ID=n
0	1	0	...	0	0	0	...	0

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 ($k \ll m, k \ll n$)。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	<p>请比较欧式距离与曼哈顿距离？</p> <p>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 (x_1, y_1), (x_2, y_2) 之间的欧式距离为：</p> $\rho = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$ <p>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 (x_1, y_1), (x_2, y_2) 之间的曼哈顿距离为：</p> $c = x_1 - x_2 + y_1 - y_2 $ <p>用一张图来区分一下两者：</p>  <p>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。</p>	<p>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</p>
---	---	--

4	<div>为什么一些场景中使用余弦相似度而不是欧式距离</div> <div></div> <div>余弦相似度：取值范围$[-1, 1]$ 余弦距离$=1 - \text{余弦相似度}$：取值范围$[0, 2]$<ul style="list-style-type: none">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。</div>	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	<div>One-hot 的作用是什么？为什么不直接使用数字作为表示</div> <div>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：</div> <table><tr><th></th><th>Feature_1</th><th>Feature_2</th><th>Feature_3</th></tr><tr><td>Sample_1</td><td>1</td><td>4</td><td>3</td></tr><tr><td>Sample_2</td><td>2</td><td>3</td><td>2</td></tr><tr><td>Sample_3</td><td>1</td><td>2</td><td>2</td></tr><tr><td>Sample_4</td><td>2</td><td>1</td><td>1</td></tr></table> <div>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。 1 -> 0001 2 -> 0010 3 -> 0100 4 -> 1000</div> <div>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。</div>		Feature_1	Feature_2	Feature_3	Sample_1	1	4	3	Sample_2	2	3	2	Sample_3	1	2	2	Sample_4	2	1	1	理解清楚并比较 One-hot 编码和数字编码的特点
	Feature_1	Feature_2	Feature_3																			
Sample_1	1	4	3																			
Sample_2	2	3	2																			
Sample_3	1	2	2																			
Sample_4	2	1	1																			

