jasonfreak
一个懒惰的人，总是想设计更智能的程序来避免做重复性工作

关于线性模型你可能还不知道的二三事（一、样本）

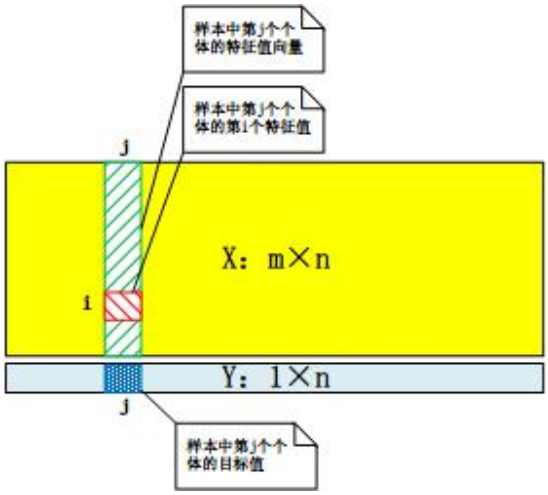
系列

- 关于线性模型你可能还不知道的二三事（一、样本）
- 关于线性模型你可能还不知道的二三事（二、也谈民主）
- 关于线性模型你可能还不知道的二三事（三、特征值与奇异值的魔力）

目录

1 样本的表示形式

在数据挖掘过程中，样本以特征值矩阵X和目标值向量Y的形式表示。容量为n，有m个特征的样本，其特征值矩阵X由n个维度为m的列向量组成，第j个列向量为样本中第j个个体的特征值向量；目标值向量Y的第j个分量为样本中第j个个体的目标值：



2 由线性模型产生的样本

机器学习(2)
数据分析(2)
数据挖掘(8)
特征工程(2)

随笔档案

2016年11月 (1)
2016年7月 (3)
2016年6月 (4)
2016年5月 (2)
2016年4月 (2)

最新评论

1. Re:使用sklearn做单机特征工程
标准化是依照特征矩阵的列处理数据，归一化是依照特征矩阵的行处理数据，这个不太理解，博主可以解释下吗？

--Stone1111

2. Re:使用sklearn进行集成学习——理论
你好博主！我想问一下“在bagging和boosting框架中，通过计算基模型的期望和方差，我们可以得到模型整体的期望和方差。为了简化模型，我们假设基模型的权重、方差及两两间的相关系数相等。”这里接.....

--implus

3. Re:使用Python进行描述性统计
很好，很清晰，赞！

--iuwai

4. Re:使用sklearn优雅地进行数据挖掘
@会飞的蜗牛引用@魔灵幽亭在你的数据集DataFrame上加一句df = df.fillna(0)...

--liuer2009

5. Re:谁动了我的特征？——sklearn特征转换行为全记录
@hnxsm这个调的哪里的...

--hustenn

阅读排行榜

1. 使用sklearn做单机特征工程(47866)
2. 使用sklearn优雅地进行数据挖掘

已知样本的特征值矩阵X，由线性模型生成样本的目标值向量的方式由以下公式定义：

$$Y = W * X + e$$

权值向量W是维度为m的行向量，误差向量e为维度为n的行向量，其分量独立同分布，服从均值为0的正态分布。之所以说这样的样本是由线性模型生成，是因为满足：

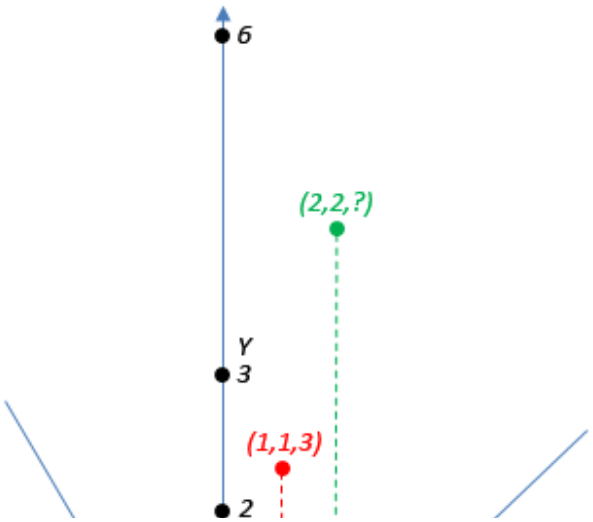
$$\begin{aligned} E(Y) &= E(W * X + e) \\ &= E(W * X) + E(e) \\ &= E(W) * E(X) + E(e) \\ &= W * E(X) + 0 \\ &= W * E(X) \end{aligned}$$

也就是说，从期望的角度来说，目标值和特征值存在线性关系！在假设样本是由线性模型产生的前提下，我们通常使用基于线性模型的机器学习算法来解决回归问题，例如：最小均方法（LMS），最小二乘法，回归支持向量机法等。但是，假设让一个完全没有机器学习背景的人来解决回归问题，他该如何入手呢？

解决回归问题，归根结底是要预测新个体的目标值。一个最直观的方式就是，让新个体（测试样本中的个体）与已知个体（训练样本中的个体）比较相似性（特征向量相似），相似度越高意味着新个体的目标值与该已知个体的目标值更接近。这样一来，计算新个体与已知个体的相似性成为预测工作的关键之处。

余弦相似性与欧式距离是衡量向量相似的最基本的两个方法。暂且让我们简化一下模型：假设样本只有2个特征，权值向量为[1, 2]，在期望情况下，特征值和目标值构成三维空间中的平面，权值向量为该平面的法平面。通过以下两例，我们可以得知余弦相似性和欧式距离在线性模型中无法使用。

例一、余弦相似性



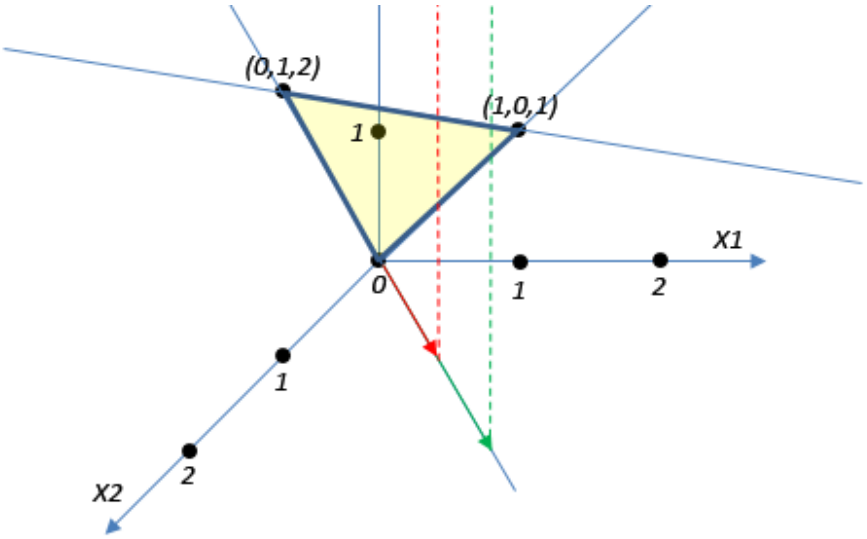
(42921)
3. 使用Python进行描述性统计(34743)
4. 使用sklearn进行集成学习——理论(22104)
5. 使用sklearn进行集成学习——实践(22084)

评论排行榜

1. 使用sklearn优雅地进行数据挖掘(21)
2. 使用sklearn做单机特征工程(18)
3. 使用sklearn进行集成学习——理论(9)
4. 虎扑论坛装备区到底有没有李宁水军？——论坛水军发现实践(6)
5. 使用sklearn进行集成学习——实践(3)

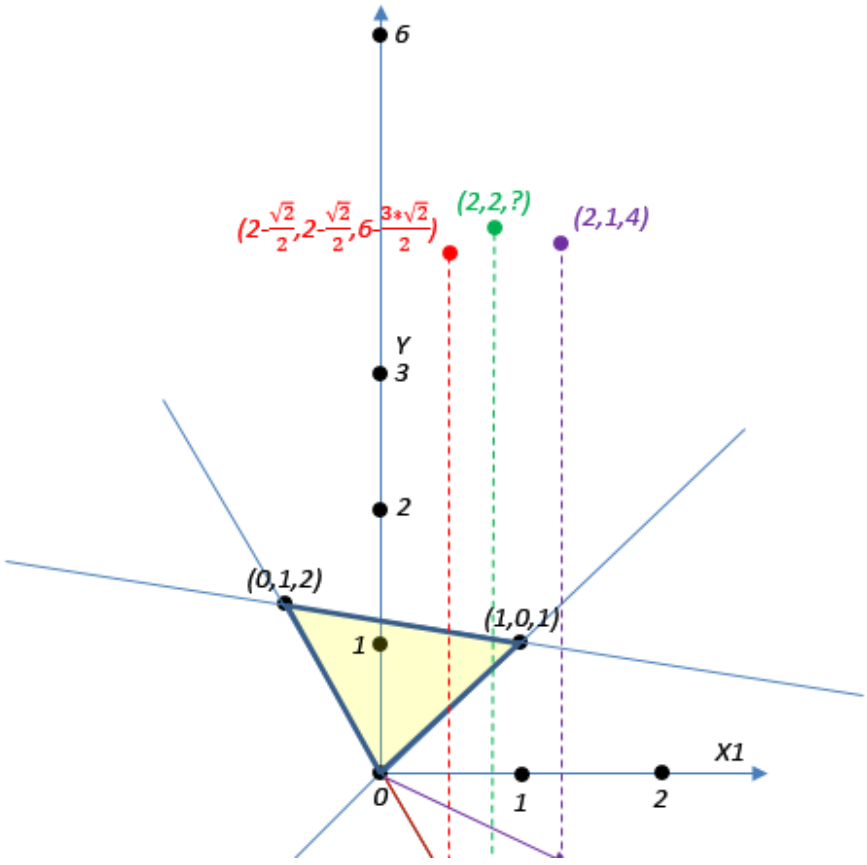
推荐排行榜

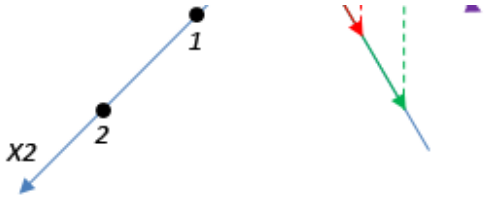
1. 使用sklearn做单机特征工程(20)
2. 使用sklearn优雅地进行数据挖掘(17)
3. 使用sklearn进行集成学习——理论(8)
4. 使用Python进行描述性统计(6)
5. easyconf——基于AugularJS的配置管理系统开发框架(5)



在本例中，已知个体（红色）的特征值向量为[1, 1]，未知个体（绿色）的特征向量为[2, 2]，通过计算余弦相似度，可得未知个体与该已知个体一致相似，其目标值也应当为$1 + 2 * 1 = 3$。但实际上，若样本是通过线性模型生成的话，其目标值应当约为$2 + 2 * 2 = 6$。由该例我们可以看到，余弦相似度只考虑了特征值向量的方向性，过于片面。

例二、欧式距离





在本例中，有两个已知个体（红色与紫色），其特征值向量与未知个体的特征值向量的欧式距离都等于1。在这种情况下，该未知个体的目标值应当与哪个已知个体更接近呢？如果样本是由线性模型产生的，该未知个体的目标值应当约为$2 + 2 * 2 = 6$。所以，以紫色的已知个体的目标值作为未知个体的目标值相对来说合适一点。通过该例可知，欧式距离也不适合在线性模型中使用。

---

3 逆矩阵的意义

那到底怎么才能准确地描述未知个体与已知个体的相似性呢？在此，我们不妨再次假设样本容量n=m，且特征值矩阵X是可逆的，也就是说样本中的个体是线性无关的。我们知道逆矩阵有这样的性质：

$$X^{-1} * X = E$$

这对我们有什么启发呢？假设未知个体的特征值向量为x，x可以用X的m个线性无关列向量（已知个体的特征值向量）表示：

$$\mathrm{x} = a_1 * X_1 + a_2 * X_2 + \cdots + a_m * X_m$$

此时将X的逆矩阵乘以未知个体x，可得：

$$X^{-1} * x = a_1 * X^{-1} * X_1 + a_2 * X^{-1} * X_2 + \cdots + a_m * X^{-1} * X_m$$

$$= \begin{bmatrix} a_1 \\ 0 \\ \cdots \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ a_2 \\ \cdots \\ 0 \end{bmatrix} + \cdots + \begin{bmatrix} 0 \\ 0 \\ \cdots \\ a_m \end{bmatrix}$$

$$= \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_m \end{bmatrix}$$

根据上式我们可以看到，在已知个体是线性无关的前提下，若未知个体能包含ai份第i个已知个体的特征，则其与第i个已知个体的近似度就为ai。显然。这样的近似表示方法，在线性模型中才是准确的。

如果样本的容量n大于m，我们该如何处理呢？假设X的秩仍然等于m，但由于X不是方阵，无法求解逆矩阵。此时我们可以将原线性模型改写成：

$$\mathrm{Y} * X^T = \mathrm{W} * \mathrm{X} * X^T + \mathrm{e} * X^T$$

此时，X乘以X的转置则变成了m维的方阵，由于X的秩为m，X与X转置的乘积的秩也为m，即可逆。此时我们需要将Y与X的转置的乘以看成新的目标值向量，X与X转置的乘积看成新的已知个体的特征值矩阵，e与X转置的乘积看成新的误差向量。不难看到，原始问题与新问题的解（回归问题的解通常是求权值向量）是“等价”的。在新问题中，特征值矩阵是方阵且可逆，这样便可通过求解新问题来解决原始问题了。

分类: 数据挖掘

标签: 数据挖掘, 线性模型, 线性代数

« 上一篇：使用sklearn优雅地进行数据挖掘
» 下一篇：关于线性模型你可能还不知道的二三事（二、也谈民主）

posted on 2016-06-02 09:35 jasonfreak 阅读(5236) 评论(0) 编辑 收藏

· 淘宝卧榻之侧，岂容拼多多安睡？
· Docker日志的10大陷阱
· OpenSSL改变开发策略：转用GitHub issue讨论补丁
» 更多新闻...

最新知识库文章:
· 领域驱动设计在互联网业务开发中的实践
· 步入云计算
· 以操作系统的角度述说线程与进程
· 软件测试转型之路
· 门内门外看招聘
» 更多知识库文章...