数据预处理

一.数据处理对象的相关介绍

1.数据处理对象是什么?

现实生活中所有的事物.

例如 网站的一个用户,购物车中的一项清单,新闻网站上的一则新闻等

一个数据处理对象一般可以理解为一项记录.

2.数据对象(记录)如何表示?

数据对象(记录)一般可以用若干个属性(特征)进行表示,而属性是描述某对象的一个侧面或多个侧面的综合值.

例 1 一个网站的用户,可以用以下方式进行描述

	昵称	年龄	性别	城市	等级	注册时间	已充值金额	
U1(	Mr.K	22	男	GZ	8	2016-01-01	100	)
U2(	Ms.Q	20	女	GZ	8	2016-01-01	80	)

其中 U1 和 U2 为数据对象;昵称,年龄,性别,城市,等级,注册时间和已充值金额是对同一研究数据对象的属性,各属性有对应的值进行表示.

例 2 某个句子的表示(文本)

D1: 数据挖掘的基础步骤是什么 D2:数据预处理是数据挖掘的基础步骤

	数据	预处理	是	数据挖掘	的	基础	步骤	什么	数挖类	
D1(	0	0	1	1	1	1	1	1	1	)
D2(	1	1	1	1	1	1	1	0	1	)

D1,D2 为数据对象,而各个词为表示描述数据对象的属性,属性值 1 表示含有该词,0 表示不含有该词. 而最后一个属性”数挖类”的值为 1 表示是数据挖掘相关的文本,0 表示不是数据挖据相关的文本.

3.属性(特征)有那些常用的数据类型?

描述数据对象的属性有着对应的值,那么不同的属性值,可以进行数据类型来划分.

常用的两种类型为 连续数值型属性 和 离散数值型属性

连续数值型属性:属性值的表示为连续的值,例如上面 “已充值金额”的值

离散数值型属性:属性值的表示为离散的值,例如上面文本例子中所有的属性值

二.Python 分词包(jieba)的安装

1.打开命令行 安装 python-pip

sudo apt-get install python-pip

2.安装 jieba 分词包

sudo pip install jieba

3.Jieba 分词包的使用

```
import jieba            #导入 jieba 分词包
ws = jieba.cut(“数据预处理是数据挖掘的基础步骤”)
for w in ws:
    print w
```

4.练习时间

已有类别 C1,C2 文本各 500 条,找出所有出现过的词 ws.然后对每个文本 d,计算它每个 w 出现的次数.作为一行结果写到文件中.

三.数据统计特征

1.算术平均数 arithmetic mean

算术平均数是指在一组数据中所有数据之和再除以数据的个数。它是反映数据集中趋势的一项指标。

把 n 个数的总和除以 n，所得的商叫做这 n 个数的算术平均数。

公式：$$A_n = \frac{a_1 + a_2 + a_3 + \cdots + a_n}{n}$$

2.加权平均数 weighted average

加权平均数是不同比重数据的平均数，加权平均数就是把原始数据按照合理的比例来计算，若 n 个数中，x_1 出现 f_1 次，x_2 出现 f_2 次，…，x_k 出现 f_k 次，那么

$\frac{x_1 f_1 + x_2 f_2 + \cdots + x_k f_k}{f_1 + f_2 + \cdots + f_k}$ 叫做 x_1、x_2、…、x_k 的加权平均数。f_1、f_2、…、f_k 是 x_1、x_2、…、x_k 的**权**(weight)。平均数是加权平均数的一种特殊情况，即各项的权相等时，加权平均数就是算术平均数。

3.几何平均数 geometric mean

n 个观察值连乘积的 n 次方根就是几何平均数。根据资料的条件不同，几何平均数分为加权和不加权之分。

公式：$G_n = \sqrt[n]{a_1 \bullet a_2 \bullet a_3 \bullet \cdots \bullet a_n}$

几何意义:作一正方形，使其面积等于以 a,b 为长宽的矩形，则该正方形的边长即为 a、b 的几何平均数

4.截断均值 trimmed mean

对一组数据排序后,丢弃高端和低端的(p/2)%数据,再进行均值计算.

通常 p 取 20,但也会根据实际问题进行特定的定义.

5.极差 Range

R = Max-Min

6.四分位极差 inter-quartile range

对 N 个数据进行排序,并给各个数数据标上序号,则 各四分位序号为

Q1 = (n+1)/4　Q2 = 2(n+1)/4　Q3 = 3(n+1)/4
IQR = Q3 - Q1

7.方差 variance

方差是各变量值与其均值离差平方的平均数，它是测算数值型数据离散程度的最重要的方法。标准差为方差的算术平方根，用 S 表示。方差相应的计算公式为

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

8.练习时间

对二(4)计算得到的矩阵的每一列进行以上 7 种统计特征的计算并其表示为一行数据写进一个文件中.

四.数据转换

1.分箱平滑

目的:去除数据中的噪声点

措施:对一组数据进行升序排序后,每 k 个数据一组,每组内所有数据用组内的均值表示

如: 13,15,16,16,19,20,20,21,22　　k = 3

第一组:13,15,16　　均值化后　　14.666,14.666,14.666

第二组:16,19,20　　均值化后　　18.333,18.333,18.333

第三组:20,21,22　　均值化后　　21,21,21

结果 14.666,14.666,14.666,18.333,18.333,18.333,21,21,21

2.数据泛化(概念分层)

对数据进行层次性的划分,现实业务中不需追求精确的区别.

例如广州每天的平均温度对人影响

温度:8,10,13,15,16,18,19,20,23,24,27,29,30,32,35

由于人对温度的感受内心没有精确到 1 度为单位.大多是某种层次性的程度的求别

(t<=8)　(8<t<=15) (15<t<=20) (20<t<=24) (24<t<=28)

极冷-3　　很冷-2　　稍冷-1　　暖和 0　　稍热 1

(28<t<=31)　(31<t)

很热 2　　极热 3

结果:-3,-2,-2,-2,-1,-1,-1,-1,0,0,1,2,2,3,3

3.规范化

目的:平衡具有较大初始值域属性与较小初始值域属性的可比性

最小-最大规范化: y=[(x-Min)/(Max-Min)](b-a) + a

表示将 x 值变到换区间[a,b]中去,其结果值为 y

例 3,5,7,9　这些值变换到[0,1]区间中

0=[(3-3)/(9-3)](1-0) + 0

0.333=[(5-3)/(9-3)](1-0) + 0

0.666=[(7-3)/(9-3)](1-0) + 0

1=[(9-3)/(9-3)](1-0) + 0

结果为 0,0.333,0.666,1

4.离散化

背景:有些算法的输入要求只能为分类属性(离散型属性),需要将连续型属性进行离散化

措施:对连续型属性值的区间进行分割,分割为 k 个区间,落在每个区间的值投影为一个特定的分类值.

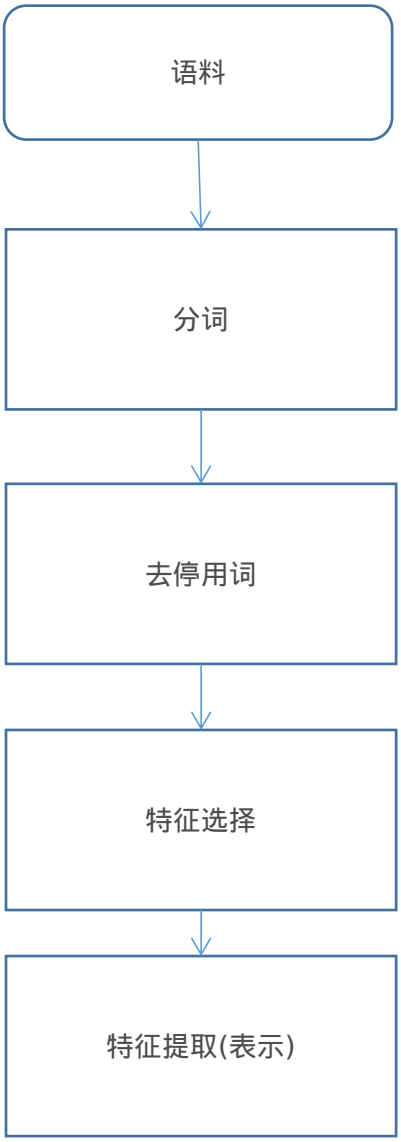
例如:概念分层中提到的温度问题.

5.练习时间

对二(4)中得到的矩阵按列进行 [0,1]最小-最大规范化

五.文本挖掘数据预处理

1.文本挖掘数据预处理流程



2.停用词

停用词:没有具体含义和分类意义的词

一般在文本数据预处理的步骤对停用词进行剔除处理.这样不仅减低了特征空间的纬度,而且提高了后期建模和计算的效率.

3.特征选择

由于中华文化博大精深,中文词汇呈现多样化,在进行数据预处理的时候,即使使用了去停用词操作后,剩下的特征仍然有可能是上 10 万级别的.这样的高纬度灾难,是当代计算机难以承受的,而且还有可能因为算法或建模数据的局限性,出现过拟合的现象.

所以需要有一种评价标准进行特征选择.选择一定量的有较强分类能力的特征进行文本表示.

具体常用方法见附录论文

例如:

(1)特征选择之前,总特征有 10w 个,选用其中任意一种特征选择方法,计算出这 10w 个特征的的数值信息,进行降序排列.

特征 T	数值 V
T1	0.98
T2	0.51
T3	0.30
…	…
T100000	0.0003

(2)选择手段

a.选 top N 个特征

b.选 前 p% 的特征

c.选 数值 V > 阈值 P 的特征

(3) 得到一个低数量级的特征空间,接下来就是使用这些特征对每一个文本进行特征提取

4.特征提取 TF*IDF

Term frequency		Document frequency		Normalization	
n (natural)	$\mathrm{tf}_{t,d}$	n (no)	1	n (none)	1
l (logarithm)	$1+\log(\mathrm{tf}_{t,d})$	t (idf)	$\log \frac{N}{\mathrm{df}_t}$	c (cosine)	$\frac{1}{\sqrt{w_1^2+w_2^2+\ldots+w_M^2}}$
a (augmented)	$0.5+\frac{0.5\times \mathrm{tf}_{t,d}}{\max_t(\mathrm{tf}_{t,d})}$	p (prob idf)	$\max\{0, \log \frac{N-\mathrm{df}_t}{\mathrm{df}_t}\}$	u (pivoted unique)	$1/u$ (Section 6.4.4)
b (boolean)	$\begin{cases}1 & \text{if } \mathrm{tf}_{t,d}>0\\ 0 & \text{otherwise}\end{cases}$			b (byte size)	$1/CharLength^{\alpha}$, $\alpha<1$
L (log ave)	$\frac{1+\log(\mathrm{tf}_{t,d})}{1+\log(\mathrm{ave}_{t\in d}(\mathrm{tf}_{t,d}))}$				

选用 第二行 ltc 模式

tf 表示该特征(词)在该文本的出现次数,

N 表示所有文本的个数,

df 表示含有该特征(词)的文本总个数.

结果表示为(假设特征选择选 top 1000)

	T1	T2	T3	T4	T5	...	T1000
D1(	0.153	0.52	0.53	0.12	0.68	...	0.71)
	.....						
D1000(	0.153	0.52	0.53	0.12	0.68	...	0.71)

六.空间向量的相似性计算

1.距离计算

距离度量（Distance）:用于衡量个体在空间上存在的距离，距离越远说明个体间的差异越大。

(1)欧几里得距离（Euclidean Distance）

欧氏距离是最常见的距离度量，衡量的是多维空间中各个点之间的绝对距离。公式如下：

$$dist(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

注:欧氏度量需要保证各维度指标在相同的刻度级别，比如对身高（cm）和体重（kg）两个单位不同的指标使用欧式距离可能使结果失效。

(2)明可夫斯基距离（Minkowski Distance）

明氏距离是欧氏距离的推广，是对多个距离度量公式的概括性的表述。公式如下：

$$dist(X,Y) = (\sum_{i=1}^{n}|x_i - y_i|^p)^{1/p}$$

这里的 p 值是一个变量，当 p=2 的时候就得到了上面的欧氏距离。

(3)曼哈顿距离（Manhattan Distance）

曼哈顿距离来源于城市区块距离，是将多个维度上的距离进行求和后的结果，即当上面的明氏距离中 p=1 时得到的距离度量公式，如下：

$$dist(X,Y) = \sum_{i=1}^{n}|x_i - y_i|$$

(4)切比雪夫距离（Chebyshev Distance）

切比雪夫距离起源于国际象棋中国王的走法，我们知道国际象棋国王每次只能往周围的 8 格中走一步，那么如果要从棋盘中 A 格（x1， y1）走到 B 格（x2， y2）最少需要走几步？扩展到多维空间，其实切比雪夫距离就是当 p 趋向于无穷大时的明氏距离：

$$dist(X,Y) = \lim_{p\to\infty}(\sum_{i=1}^{n}|x_i - y_i|^p)^{1/p} = max\,|x_i - y_i|$$

其实上面的曼哈顿距离、欧氏距离和切比雪夫距离都是明可夫斯基距离在特殊条件下的应用。

2.相似度度量

相似度度量（Similarity），即计算个体间的相似程度，与距离度量相反，相似度度量的值越小，说明个体间相似度越小，差异越大。

(1)向量空间余弦相似度（Cosine Similarity）

余弦相似度用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小。相比距离度量，余弦相似度更加注重两个向量在方向上的差异，而非距离或长度上。公式如下：

$$sim(X,Y) = cos\theta = \frac{\vec{x}\cdot\vec{y}}{\|x\|\cdot\|y\|}$$

(2)皮尔森相关系数（Pearson Correlation Coefficient）

即相关分析中的相关系数 r，分别对 X 和 Y 基于自身总体标准化后计算空间向量的余弦夹角。公式如下：

$$r(X,Y) = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2} \cdot \sqrt{n\sum y^2 - (\sum y)^2}}$$

(3)Jaccard 相似系数（Jaccard Coefficient）

$$f(A,B) = \frac{A \cdot B}{|A|^2 + |B|^2 - A \cdot B}$$