

计算题：

- 基本概念计算（各种距离、各种相似度、Precision、Recall、ACC、F1）

104

检测冗余样本

思想：数据样本之间的相关性，数据融合、去除冗余

方法：距离度量

- 欧几里得距离
- 曼哈顿距离
- 汉明距离
- 明氏距离
-

方法：相似度计算

- 余弦相似度
- Jaccard相似度
-

分类——模型评估方法：混淆矩阵

	PREDICTED CLASS	
	Class=Yes	Class=No
ACTUAL CLASS	Class=Yes a (TP)	b (FN)
	c (FP)	d (TN)

Most widely-used metric

$$Accuracy = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

正确率 Precision (p) = $\frac{a}{a+c}$

召回率 Recall (r) = $\frac{a}{a+b}$

F值 F-measure (F) = $\frac{2rp}{r+p} = \frac{2a}{2a+b+c}$

Count	PREDICTED CLASS	
	Class=Yes	Class=No
ACTUAL CLASS	a	b
	c	d

- 计算Spearman相关系数、CG、DCG、NDCG
- PCA和TF-IDF忘了有没有了？，但也挺重要的，需要看一看
- 参数估计——抛硬币
 - 似然估计
 - 最大后验估计（Beta分布）



参数估计——最大后验估计

61

- 最大后验概率估计(MAP)—理解先验 $p(\theta)$
 - 扔硬币的例子：我们期望先验概率（待估计的参数 θ ）分布在0.5处取得最大值，也可以选用Beta分布（ θ 服从Beta分布）即：

$$p(\theta|\alpha, \beta) \triangleq Beta(\alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

- 其中，Beta函数是 $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
- Gamma函数 $\Gamma(n) = (n-1)!$
- Beta分布的随机变量范围是[0,1]，不同参数情况下的Beta分布的概率密度函数形式如图



参数估计——最大后验估计

63

- 最大后验概率估计(MAP)—理解先验 $p(\theta)$
 - 扔硬币的例子：我们期望待估计的参数 θ 的先验分布在0.5处取得最大值，可以选用Beta分布（ θ 服从Beta分布）即：

$$p(\theta|\alpha, \beta) \triangleq Beta(\alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

- 取 $\alpha = \beta = 5$ ，使得先验分布Beta分布在0.5处取得最大值
- 使用MAP方法求解参数

· 贝叶斯估计

参数估计—贝叶斯估计

贝叶斯估计

下面仍然以抛硬币为例，此时选择Beta分布作为先验，类似MAP：

$$P(\theta) = \gamma \theta^{\alpha-1} (1-\theta)^{\beta-1} \quad \text{其中 } \gamma \text{ 为归一化常数}$$

Beta分布在这里作为先验来做参数估计尤为有用

假设我们现在只有先验，没有数据，此时来考虑一次单独的硬币投掷 X_1 ，那么贝叶斯方法预测该硬币朝上的概率为：

$$P(x_1 = 1) = \int_0^1 P(x_1 = 1 | \theta) P(\theta) d\theta = \int_0^1 \theta P(\theta) d\theta = \int_0^1 \theta \gamma \theta^{\alpha-1} (1-\theta)^{\beta-1} d\theta$$

积分后可得： $P(x_1 = 1) = \frac{\alpha}{\alpha + \beta}$ (积分过程复杂，此处省略)

结论：Beta分布作为先验表明（假设）我们已经看到 α 次正面朝上和 β 次反面朝上

参数估计—贝叶斯估计

贝叶斯估计

现在，让我们在先验的基础上加入更多观测，抛硬币实验 X 中有正面 M_1 ，反面 M_2 ，则后验估计为：

$$P(\theta | X) = \frac{P(X | \theta) P(\theta)}{\int P(X | \theta) P(\theta) d\theta} = \frac{\theta^{M_1} (1-\theta)^{M_2} \frac{\gamma}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}}{\int \theta^{M_1} (1-\theta)^{M_2} \frac{\gamma}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} d\theta} = \frac{\theta^{M_1 + \alpha - 1} (1-\theta)^{M_2 + \beta - 1}}{\int \theta^{M_1 + \alpha - 1} (1-\theta)^{M_2 + \beta - 1} d\theta} = \text{Beta}(\theta | M_1 + M_2, \alpha + M_1, \beta + M_2)$$

观察：在抛硬币的实验中(虽然 $X(i)$ 为二项分布)，当先验 $P(\theta)$ 为Beta分布时，后验 $P(\theta | X)$ 也为Beta分布，即更新后的参数服从一个新的Beta($\alpha+M_1, \beta+M_2$)分布

即，在把后验概率看作为和先验概率一样的分布形式的连续，贝叶斯公式分母 $\mu(X)$ 可以视为一个常数，往往起到了归一化的作用。

参数估计—贝叶斯估计

贝叶斯估计

现在，让我们在先验的基础上加入更多观测，抛硬币实验 X 中有正面 M_1 ，反面 M_2 ，则后验估计为：

$$P(\theta | X) = \frac{P(X | \theta) P(\theta)}{\int P(X | \theta) P(\theta) d\theta} = \frac{\theta^{M_1} (1-\theta)^{M_2} \frac{\gamma}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}}{\int \theta^{M_1} (1-\theta)^{M_2} \frac{\gamma}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} d\theta} = \frac{\theta^{M_1 + \alpha - 1} (1-\theta)^{M_2 + \beta - 1}}{\int \theta^{M_1 + \alpha - 1} (1-\theta)^{M_2 + \beta - 1} d\theta} = \text{Beta}(\theta | M_1 + M_2, \alpha + M_1, \beta + M_2)$$

观察：在抛硬币的实验中(虽然 $X(i)$ 为二项分布)，当先验 $P(\theta)$ 为Beta分布时，后验 $P(\theta | X)$ 也为Beta分布，即更新后的参数服从一个新的Beta($\alpha+M_1, \beta+M_2$)分布

这种情况我们称之为Beta分布是二项分布的自然共轭

关联规则中的若干概念计算：

关联规则挖掘

关联规则挖掘的基本概念

Itemset (项集)

- 一个或多个项目(Items)的集合
- k-Itemset: 大小为k的项集
- 例: {Milk, Bread, Diaper}是3项集

Support (支持度)

- 一个项集在数据中的出现频率
- 例: $\text{support}(\{\text{Milk, Bread, Diaper}\}) = \frac{2}{5}$

Frequent Itemset (频繁项集)

- 用户自行设定最小支持度阈值 min_sup ，支持度大于 min_sup 的项集称为频繁项集
- 例: 设 $\text{min_sup} = 0.3$ ，则{Milk, Bread, Diaper}为频繁项集

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

关联规则挖掘

关联规则挖掘的基本概念

Association Rule (关联规则)

- 形如 $X \rightarrow Y$ 的表达式， X, Y 均为项集
- 例: {Milk, Diaper} \rightarrow {Beer}

Confidence (置信度)

- 度量包含 X 的事务中同时出现 Y 的频率
- 例: 对于关联规则{Milk, Diaper} \rightarrow {Beer}

强关联规则

- 用户自行设定最小置信度阈值 min_conf ，置信度大于 min_conf 的规则称为强关联规则
- 例: 设 $\text{min_conf} = 0.5$ ，则{Milk, Diaper} \rightarrow {Beer}为强关联规则

$$\text{Confidence}(A \rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)}$$
$$\text{confidence}(\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\}) = \frac{2}{3}$$

练习

请依据下表计算出关于早餐的关联规则 (面包) \rightarrow (豆浆) 的置信度

	买豆浆	不买豆浆	
买面包	90	30	120
不买面包	390	90	480
	480	120	600

买面包的次数=120，
买面包的同时买豆浆的次数=90
置信度= $\frac{90}{120} = \frac{3}{4}$

Example:
{Milk, Diaper} \Rightarrow Beer
 $s = \frac{\sigma(\{\text{Milk, Diaper, Beer}\})}{|\mathcal{T}|} = \frac{2}{5} = 0.4$
 $c = \frac{\sigma(\{\text{Milk, Diaper, Beer}\})}{\sigma(\{\text{Milk, Diaper}\})} = \frac{2}{3} = 0.67$

关联规则挖掘

关联规则挖掘——Statistical-based Measures

Measures that take into account statistical dependence

$$\text{Lift} = \frac{P(Y | X)}{P(Y)}$$
$$\text{Interest} = \frac{P(X, Y)}{P(X)P(Y)}$$
$$\text{PS} = P(X, Y) - P(X)P(Y)$$
$$\phi\text{-coefficient} = \frac{P(X, Y) - P(X)P(Y)}{\sqrt{P(X)[1 - P(X)]P(Y)[1 - P(Y)]}}$$

A-Priori算法
略

决策树（信息熵、信息增益、基尼指数、Error、如何利用这些指标分裂）
略

贝叶斯分类器（类似下图的一道题）

贝叶斯分类器

For example:

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

A

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

A: attributes
M: mammals
N: non-mammals

$$P(A|M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A|N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A|M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A|N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

$P(A|M)P(M) > P(A|N)P(N)$
=> Mammals

感知机（PPT原题↓）

例 如图3所示的训练数据集，其正实例点是 $x_1 = (3,3)^T$ ， $x_2 = (4,3)^T$ ，负实例点是 $x_3 = (1,1)^T$ ，试用感知机学习算法的原始形式求感知机模型 $f(x) = \text{sign}(w \cdot x + b)$ ，即求出 w 和 b 。这里 $w = (w^{(1)}, w^{(2)})^T$ ， $x = (x^{(1)}, x^{(2)})^T$

假设错分类点定义为

$$y_i(w \cdot x_i + b) \leq 0$$

假设学习速率为**1**，则每步更新为：


$$w = w + y_i x_i$$

$$b = b + y_i$$

迭代次数	误分类点	w	b	$w \cdot x + b$
0		0	0	0
1	x_1	$(3,3)^T$	1	$3x^{(1)} + 3x^{(2)} + 1$
2	x_3	$(2,2)^T$	0	$2x^{(1)} + 2x^{(2)}$
3	x_3	$(1,1)^T$	-1	$x^{(1)} + x^{(2)} - 1$
4	x_3	$(0,0)^T$	-2	-2
5	x_1	$(3,3)^T$	-1	$3x^{(1)} + 3x^{(2)} - 1$
6	x_3	$(2,2)^T$	-2	$2x^{(1)} + 2x^{(2)} - 2$
7	x_3	$(1,1)^T$	-3	$x^{(1)} + x^{(2)} - 3$
8	0	$(1,1)^T$	-3	$x^{(1)} + x^{(2)} - 3$

注：ARIMA、ETS、Theta我没印象，疑似是今年新加的内容

简答题：

· 原题 

□ 特征工程（重复迭代）的流程

1. 对特征进行头脑风暴

深入分析问题，观察数据的基本统计信息，结合问题的相关领域知识和参考其他问题的相关特征工程的方法并应用到自身的问题中来。

2. 特征的设计—基础且重要的步骤

人工设计特征、自动提取特征，或两者结合，得到模型使用的特征。

3. 特征的选择

使用不同的特征重要性评分方法或者特征选择方法，对特征的有效性进行分析，选出有效的特征。

4. 评估模型

利用所选择的特征对测试数据进行预测，评估模型的性能。

5. 上线测试

通过在线测试的效果判断特征是否有效，若不能达到要求，则重复2-5步骤，直到模型的性能达到要求。

3/6/2025

· 结构化数据、半结构化数据、非结构化数据的定义与区别

· 其它忘了

其它：

除了第一章以及太偏机器学习的内容（如SVM）不会考，其它基本概念和方法都可能出现在选择、简答中，不过在我印象中都不难