

第九周补讲内容

陈欣鸿 陈昱夫 毛润泽

目录

- 问题
- 补充知识点
- 决策树实现细节
- 决策树优化
- 任务布置

报告

- 页数
- 原理
- 伪代码
- 结果
- 思考题

原理

(1) 贝叶斯定理 (Bayes rule)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

其中 $P(A|B)$ 是在 B 发生的情况下 A 发生的可能性。

在贝叶斯定理中，每个名词都有约定俗成的名称：

- $P(A)$ 是 A 的先验概率或边缘概率。之所以称为“先验”是因为它不考虑任何 B 方面的因素。
- $P(A|B)$ 是已知 B 发生后 A 的条件概率，也由于得自 B 的取值而被称作 A 的后验概率。
- $P(B|A)$ 是已知 A 发生后 B 的条件概率，也由于得自 A 的取值而被称作 B 的后验概率。
- $P(B)$ 是 B 的先验概率或边缘概率，也作标准化常量 (*normalized constant*)。

按这些术语，Bayes 定理可表述为：

$$\text{后验概率} = (\text{相似度} * \text{先验概率}) / \text{标准化常量}$$

也就是说，后验概率与先验概率和相似度的乘积成正比。

另外，比例 $P(B|A)/P(B)$ 也有时被称作标准相似度 (*standardised likelihood*)，Bayes 定理可表述为：后验概率 = 标准相似度 * 先验概率

推导：

根据条件概率的定义。在事件 B 发生的条件下事件 A 发生的概率是

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

同样地，在事件 A 发生的条件下事件 B 发生的概率

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

整理与合并这两个方程式，我们可以找到

$$P(A|B)P(B) = P(A \cap B) = P(B|A)P(A)$$

这个引理有时称作概率乘法规则。上式两边同除以 $P(B)$ ，若 $P(B)$ 是非零的，我们可以得到贝叶斯定理：

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

贝叶斯定理是关于随机事件 A 和 B 的条件概率（或边缘概率）的一则定理。

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

其中 $P(A|B)$ 是在 B 发生的情况下 A 发生的可能性。

在贝叶斯定理中，每个名词都有约定俗成的名称：

- $P(A)$ 是 A 的先验概率或边缘概率。之所以称为“先验”是因为它不考虑任何 B 方面的因素。
- $P(A|B)$ 是已知 B 发生后 A 的条件概率，也由于得自 B 的取值而被称作 A 的后验概率。
- $P(B|A)$ 是已知 A 发生后 B 的条件概率，也由于得自 A 的取值而被称作 B 的后验概率。
- $P(B)$ 是 B 的先验概率或边缘概率，也作标准化常量 (*normalized constant*)。

按这些术语，Bayes 定理可表述为：

$$\text{后验概率} = (\text{相似度} * \text{先验概率}) / \text{标准化常量}$$

也就是说，后验概率与先验概率和相似度的乘积成正比。

另外，比例 $P(B|A)/P(B)$ 也有时被称作标准相似度 (*standardised likelihood*)，Bayes 定理可表述为：

$$\text{后验概率} = \text{标准相似度} * \text{先验概率}$$

从条件概率推导贝叶斯定理

根据条件概率的定义。在事件 B 发生的条件下事件 A 发生的概率是

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

同样地，在事件 A 发生的条件下事件 B 发生的概率

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

整理与合并这两个方程式，我们可以找到

$$P(A|B)P(B) = P(A \cap B) = P(B|A)P(A).$$

这个引理有时称作概率乘法规则。上式两边同除以 $P(B)$ ，若 $P(B)$ 是非零的，我们可以得到贝叶斯定理：

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

(2) 朴素贝叶斯处理分类问题 (Classification for Naïve Bayes)

Bernoulli Model (伯努利模型) : a document is represented by a feature vector with **binary** elements taking value 1 if the corresponding word is present in the document and 0 if the word is not present.

$$p(x_k|e_i) = \frac{n_{e_i}(x_k)}{N_{e_i}} \quad p(e_i) = \frac{N_{e_i}}{N}$$

where $n_{e_i}(x_k)$ is the number of documents of emotion e_i in which x_k is observed, and N_{e_i} and N is the number of documents with emotion e_i and total documents, respectively.

Multinomial Model (多项式模型) : a document is represented by a feature vector with integer elements whose value is the **frequency** of that word in the document.

$$p(x_k|e_i) = \frac{nw_{e_i}(x_k)}{nw_{e_i}} \quad p(e_i) = \frac{N_{e_i}}{N}$$

where $nw_{e_i}(x_k)$ is the number of times word x_k occurs in documents with emotion e_i , and nw_{e_i} is the total number of words occurs in documents with emotion e_i .

(3) 朴素贝叶斯处理回归问题 (Regression for Naïve Bayes)

To predict emotion e_i of test document $X_3 = (x_3, x_4)$, we need to estimate:

$$\arg \max_{e_i} p(e_i|X) = \arg \max_{e_i} \sum_{j=1}^M \prod_{k=1}^{K'} p(x_k|e_i, d_j) p(d_j, e_i)$$

(4) 拉普拉斯平滑 (Laplace Smoothing)

Notice that if the word x_k in test document does not occur in the training set, $p(x_k|d_j, e_i)$ will be zero and thus cause the resulting value becoming 0.

Solution: **Laplace Smoothing!** (拉普拉斯平滑)

- regression model:

$$p(x_k|d_j, e_i) = \frac{x_k + 1}{\sum_{k=1}^K x_k + K}$$

- classification model:

$$\text{Bernoulli: } p(x_k|e_i) = \frac{n_{e_i}(x_k) + 1}{N_{e_i} + 2}$$

$$\text{Multinomial: } p(x_k|e_i) = \frac{nw_{e_i}(x_k) + 1}{nw_{e_i} + V_{e_i}}$$

where nw_{e_i} is the total number of words in documents with emotion e_i , and V_{e_i} is the number of non-repetitive words with label e_i .

伪代码反例

```
void AplusB() {
    ifstream "A", "B"; //读取文件 A, B
    rows = max{A.rows, B.rows};
    cols = max{A.cols, B.cols};

    push A -> C;
    push B -> C; //去重，词向量的值加 1 (相同项的值)
    sort(C); //排序，不重复按行再列
    ofstream fout << "C"; //写入文件 C
```

伪代码反例

```
void Div() {
    ifstream fin>>;           //读取数据集文件
    while(getline){             //取一条完整的文本
        strtok(",\t");         //去掉序号、情感权重，得到分析文本

        while(strtok(", ")){    //取一个单词

void smatrix() {
    ofstream fout<< //输出到文件
    cs                  //文本总数（行数）
    rs                  //单词总数（列数）
    for cs//遍历每个文本
        for ws //遍历每个单词
            if(wf != 0)//词频不为 0，即存在
                cnt++;
            cnt;          //输出三元表第三行
    //遍历，找到单词存在的位置，或是在 onehot 矩阵中的坐标
    fout << x << y << 1;
```

伪代码反例

第三步：构建所有文本的不重复词向量 `total_words(vector<string>)` 和每个文本的词向量
`train_set(vector<each_row>)`

将上一步分出的每个单词，与 `total_words` 中已有的单词进行比较，若非已有单词
则将其加入到 `total_words` 中

将单词与该(第 `i` 个)文本的词向量比较，若不存在，则将其加入到
`train_set[i].row_words` 中；若存在则增加相应的 `num` 值

第四步：获取 one-hot 矩阵

打开文件 `onehot.txt`

将 `i` 从 0 到 `row_size(文本数量)-1`：

对第 `i` 个文本的词向量 `train_set[i]` 根据单词在 `total_words` 中位置 `position` 来从

结果

Knn 回归-测试集结果:

A	B	C	D	E	F	G
textid	anger	disgust	fear	joy	sad	surprise
1	0.088756	0	0.218283	0.183867	0.289888	0.219206
2	0.070767	0	0.195792	0.200479	0.385266	0.147696
3	0.103633	0	0.24463	0.293582	0.197396	0.160759
4	0.104061	0	0.166424	0.39876	0.113612	0.217144
5	0.067391	0	0.10605	0.489761	0.130146	0.206652
6	0.110435	0	0.273227	0.127544	0.319231	0.169562
7	0.007549	0	0.107002	0.424915	0.015912	0.444622
8	0.11562	0	0.283973	0.191539	0.270856	0.138012
9	0	0.060011	0.199845	0.35603	0.076909	0.307205
10	0.117811	0	0.230617	0.272414	0.165707	0.213451
11	0.020326	0	0.143423	0.475613	0.13281	0.227829
12	0.148325	0	0.291524	0.22124	0.217365	0.121546
13	0.090285	0	0.096877	0.479275	0.10736	0.226203
14	0.11766	0	0.208597	0.347275	0.241376	0.085091
15	0.116727	0	0.379226	0.203046	0.171326	0.129674
16	0.116101	0	0.263185	0.25227	0.211457	0.156986
17	0.097866	0	0.207287	0.409234	0.125429	0.160184
18	0.113285	0	0.159483	0.326618	0.1561	0.244613
19	0.087718	0	0.169184	0.36579	0.146317	0.230992
20	0.016726	0	0.1779	0.35932	0.230259	0.215795
21	0.116467	0	0.260992	0.309613	0.172158	0.140769
22	0.066658	0	0.211246	0.32387	0.250918	0.147385
23	0.08492	0	0.183316	0.35213	0.243527	0.136107
24	0.118993	0	0.074374	0.022605	0.124245	0.659783

Knn 回归-验证集结果:

A	B	C	D	E	F	G
textid	anger	disgust	fear	joy	sad	surprise
1	0.153261	0	0.304928	0.181966	0.214596	0.145248
2	0.038158	0	0.159768	0.39959	0.118547	0.283937
3	0.091285	0	0.229783	0.292669	0.220007	0.166257
4	0.130093	0	0.216042	0.355166	0.173065	0.125633
5	0.025574	0	0.112089	0.429346	0.153048	0.279944
6	0.083212	0	0.160249	0.347983	0.142676	0.265879
7	0.142601	0	0.340099	0.047838	0.372533	0.09693
8	0.089098	0	0.177352	0.463886	0.11746	0.152204
9	0.113072	0	0.236335	0.300509	0.137332	0.212752

textid, label	21, joy	278, joy
0, joy	22, joy	279, sad
1, joy	23, fear	280, joy
2, joy	24, fear	281, joy
3, joy	25, sad	283, joy
4, joy	26, fear	284, joy
5, joy	27, joy	285, joy
6, sad	28, sad	286, joy
7, joy	29, joy	287, joy
8, joy	30, joy	288, sad
9, joy	31, fear	289, joy
10, joy	32, joy	290, joy
11, joy	33, joy	292, joy
12, joy	34, sad	293, joy
13, sad	35, joy	294, joy
14, joy	36, joy	295, joy
15, joy	37, joy	296, joy
16, joy	38, joy	297, joy
17, joy	39, joy	298, joy
18, joy	40, joy	299, joy
19, joy	41, joy	301, joy
20, fear	42, joy	302, joy

上述为 k 取 18 时的部分结果

(2) Knn 回归

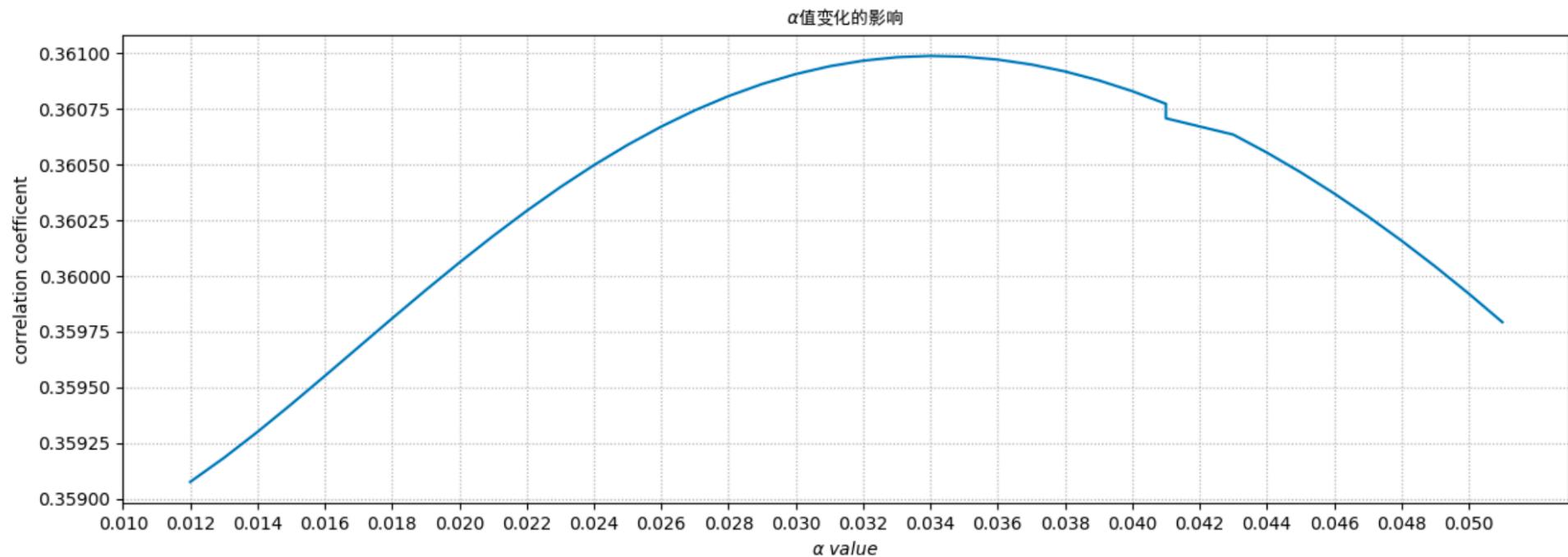
通过验证集得到的部分结果:

textid	anger	disgust	fear	joy	sad	surprise
0	0.0779	0.0759	0.1106	0.258	0.254	0.2236
1	0.0771	0.0432	0.1077	0.4079	0.0935	0.2706
2	0.0941	0.0781	0.1368	0.307	0.1562	0.2278
3	0.0784	0.0428	0.142	0.243	0.2326	0.2613
4	0.0866	0.0318	0.1655	0.3915	0.1192	0.2055
5	0.07	0.04	0.1142	0.4159	0.0274	0.3325
6	0.0834	0.055	0.2057	0.2205	0.285	0.1504
7	0.0791	0.0722	0.1099	0.3856	0.1017	0.2315
8	0.0903	0.0443	0.219	0.2396	0.1583	0.2484
9	0.0915	0.0659	0.2306	0.2616	0.1596	0.1908
10	0.0767	0.0531	0.145	0.3703	0.1692	0.1857
11	0.0795	0.0325	0.1757	0.3289	0.2105	0.1728
12	0.0865	0.0619	0.2087	0.1958	0.1789	0.2683

结果

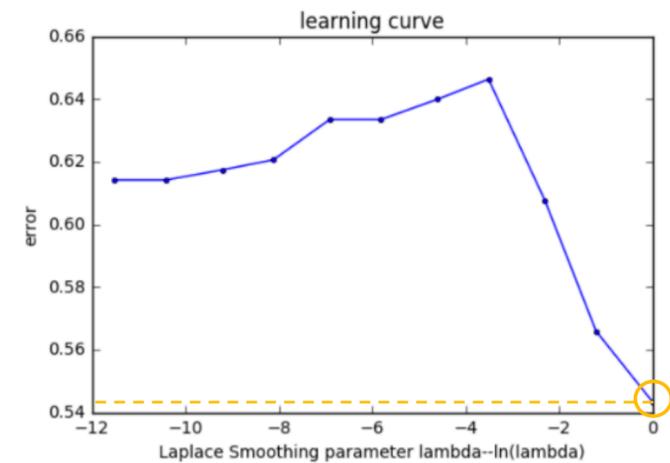
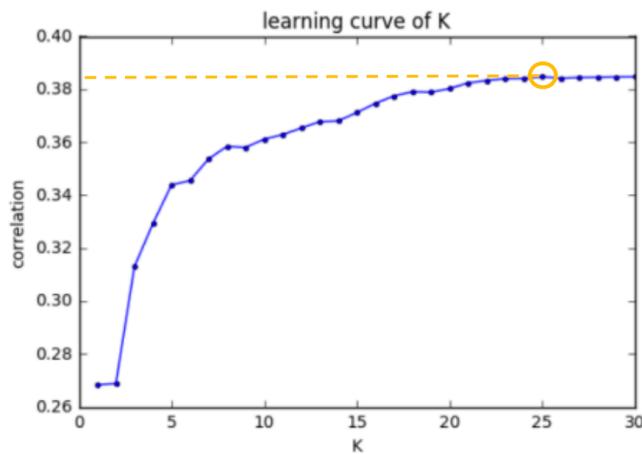
NB回归

在不同的 α 值下，相关系数变化为

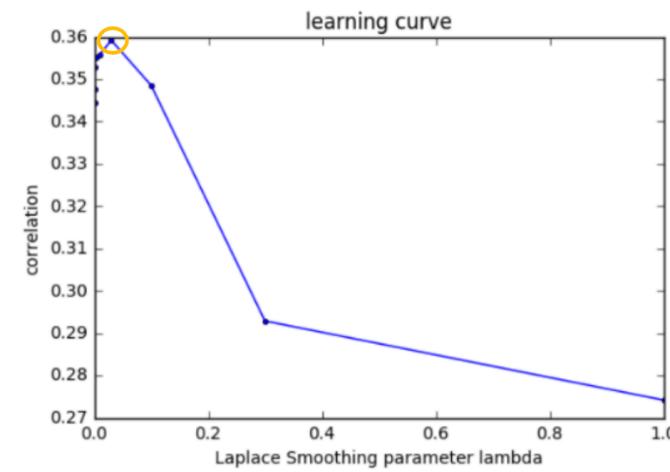
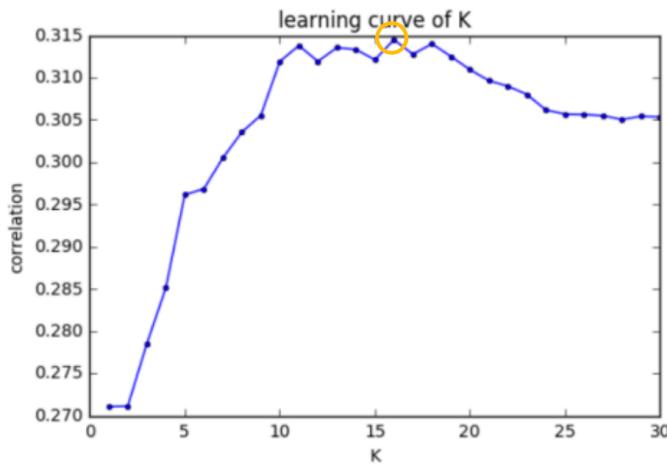


结果

曼哈顿



欧氏



思考题

3. IDF 数值有什么含义？TF-IDF 数值有什么含义

百度百科：“IDF 的主要思想是：如果包含词条 t 的文档越少，也就是 n 越小，IDF 越大，则说明词条 t 具有很好的类别区分能力。”

3. 在稀疏矩阵程度不同的时候，曼哈顿距离和欧氏距离表现有什么区别？为什么？

(好像) 无区别；

4、伯努利模型和多项式模型分别有什么优缺点?

思考题



对在情感 e 下，单词 x 出现的概率：

伯努利模型是这样算的：

$$P(x|e) = \frac{\text{情感为 } e \text{ 下包含单词 } x \text{ 的文本数}}{\text{情感为 } e \text{ 下的文本总数}}$$

而多项式模型是这样算的：

$$P(x|e) = \frac{\text{情感为 } e \text{ 下 } x \text{ 出现的总次数}}{\text{情感为 } e \text{ 下单词的总数}}$$

优缺点不明。

思考题

TF-IDF 算法是一种统计方法，用于统计一个字词在一个文件集中的一份文件的重要性。一般来说，字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。**TF-IDF** 算法正是结合这两者的特点，统计词频并用逆向文件频率进行加权来计算出该字词的重要程度。这样可以过滤掉常见的词语，保留重要的词语。（以上大部分内容来自网络）

TF-IDF 是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降

简介

 编辑

TF-IDF是一种**统计方法**，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在**语料库**中出现的频率成反比下降。**TF-IDF**加权的各种形式常被**搜索引擎**应

归一化

- 归一化 \neq 让和等于1
- 让和等于 1 是一种归一化
- 归一化方法可参考 lab2 课件
- 对什么归一化？
- 为什么归一化？

归一化

	售出数量	评价(1-5分)	是否值得购买	与4距离	
1	20000	2	否	4002	✓
2	10000	5	是	6001	
3	10000	1	否	6003	
4	16000	4	?		

KNN决定		
如果是用街区距离，会由谁决定？		不值得购买

- 对列归一化？对行归一化？全局一起归一化？

归一化

	售出数量	评价(1-5分)	是否值得购买	与4距离	
1	1	0.25	否	0.9	
2	0	1	是	0.85	√
3	0	0	否	1.35	
4	0.6	0.75	?		

- 归一化会抵消由原数据可能产生的影响模型的效果
- **可能会提高准确度**

概率

$$\sum_x p(x|a, b) = ?$$

$$\sum_a p(x|a, b) = ?$$

贝叶斯公式

- $P(A|B)$ 是 ?
- $P(B|A)$ 是 ?
- $P(B)$ 是 ?
- $P(A)$ 是 ?

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

概率

$$\sum_x p(x|a, b) = ?$$

$$\sum_a p(x|a, b) = ? \quad p(x|b)$$

贝叶斯公式

- $P(A|B)$ 是后验概率
- $P(B|A)$ 是似然度
- $P(B)$ 是标准化常量
- $P(A)$ 是先验概率

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

拉普拉斯平滑

$$\text{Multinomial: } p(x_k|e_i) = \frac{n w_{e_i}(x_k) + 1}{n w_{e_i} + V}$$

- 如果平滑是为了防止概率乘积为 0，为什么要用拉普拉斯平滑的方式？为什么不直接默认成一个很小的数值，比如 0.001，也不会怎么影响其他的概率？
- 平滑的时候分子的 1 怎么理解？分母的 V 怎么理解？
- 如果分子加的是一个实数值 α ，分母加什么？为什么？

拉普拉斯平滑

$$\text{Multinomial: } p(x_k | e_i) = \frac{n w_{e_i}(x_k) + 1}{n w_{e_i} + V}$$

- 分子既然是该单词在这个词袋里面出现的词数，如果加上1，说明每个单词都多出现了一次
- 既然每个单词都在这个词袋里面多出现了一次，说明这个词袋现在多了V个单词

拉普拉斯平滑

$$p(x_k | d_j, e_i) = \frac{x_k + 1}{\sum_{k=1}^K x_k + K}$$

- 这个数值算出来是什么值？分子的 x_k 是什么？
- 如果要用 TFIDF 做特征，还是这样平滑吗？

词袋

- 文本被看成是无序的词汇集合，忽略语法或者是单词的顺序
 - I read a book today
 - today I read a book
 - read I a book today
 - book today a read I
-
- 什么时候就导致了模型是词袋模型？
 - 忽略单词的顺序或者语法，可能会有怎样的问题？

总结

- 不认真对待 = 没有分数
- 认真做了实验不好好写报告 = 没有分数
- 报告好好写了但是补交 = 降分
- 只知道照着PPT说的来做， 没有自己的思考 = 学完这门课还是等于没学
- Project 抱大腿的可能性不大， 就算组队也是分开打分， 一个高分一个没分是完全可能的

决策树实现细节

剪枝

- 预剪枝
 - 限制决策树深度
 - 为每个叶节点的数据个数设定阈值
 - 为准确率的增益设定阈值
- 后剪枝
 - 基于错误率剪枝
 - 基于模型复杂度剪枝

剪枝

- 后剪枝
 - 基于错误率剪枝

$$E(L) = \sum_{\text{All leaf } l} e(l) \quad E(L') = \sum_{\text{All leaf } t \text{ except } k} e(l)$$

- 基于模型复杂度剪枝

$$e_c = \frac{\sum_{l=1}^L (r(n_l) + \zeta(n_l))}{\sum_{l=1}^L m(n_l)}$$

优化

- 剪枝
- 尝试使用随机森林
- 改变叶子节点的基础模型

课堂任务

- 上课当天完成三种决策树的选择属性的函数中任意一种的编写
- ID3 => 根据信息增益选择
- C4.5 => 根据信息增益率选择
- C&RT => 根据 GINI 指数选择
- 输入：完整数据集
- 输出：在当前数据集下应该选择 **哪种属性** 作为根节点的属性
- 只提交代码文件，命名为“**1535xxxx_xiaoming.xx**”，xx视编程语言而定，多版本的在后面加“_v1”，数字视提交版本号而定
- 该任务一定程度参与实验报告评分，请务必认真完成
- DDL: **上课当天晚上十二点**，交到 ftp 上的 **Week 9 Mission**