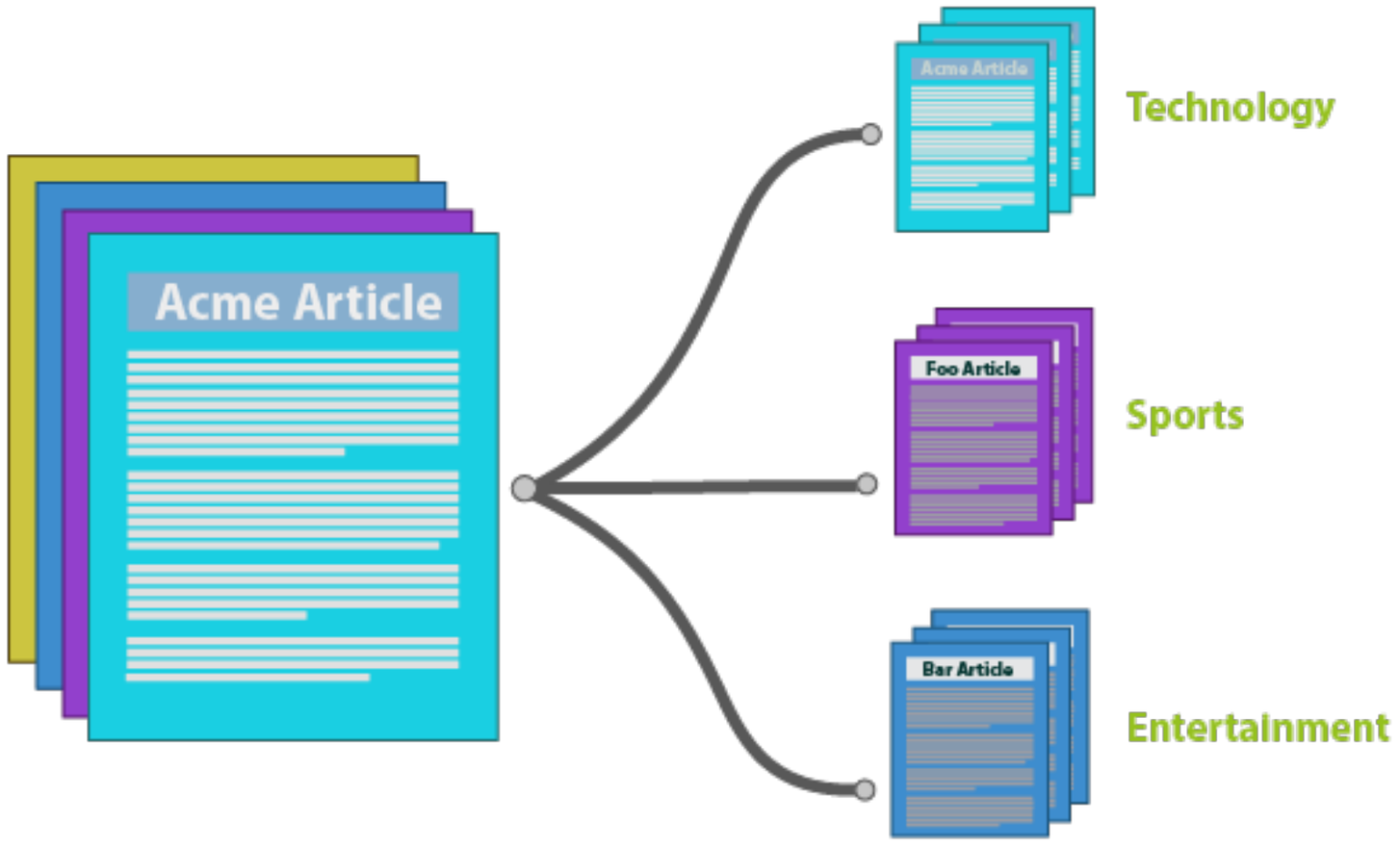
人工智能实验 4 ——新闻短文本分类

UESTC · 2021 fall

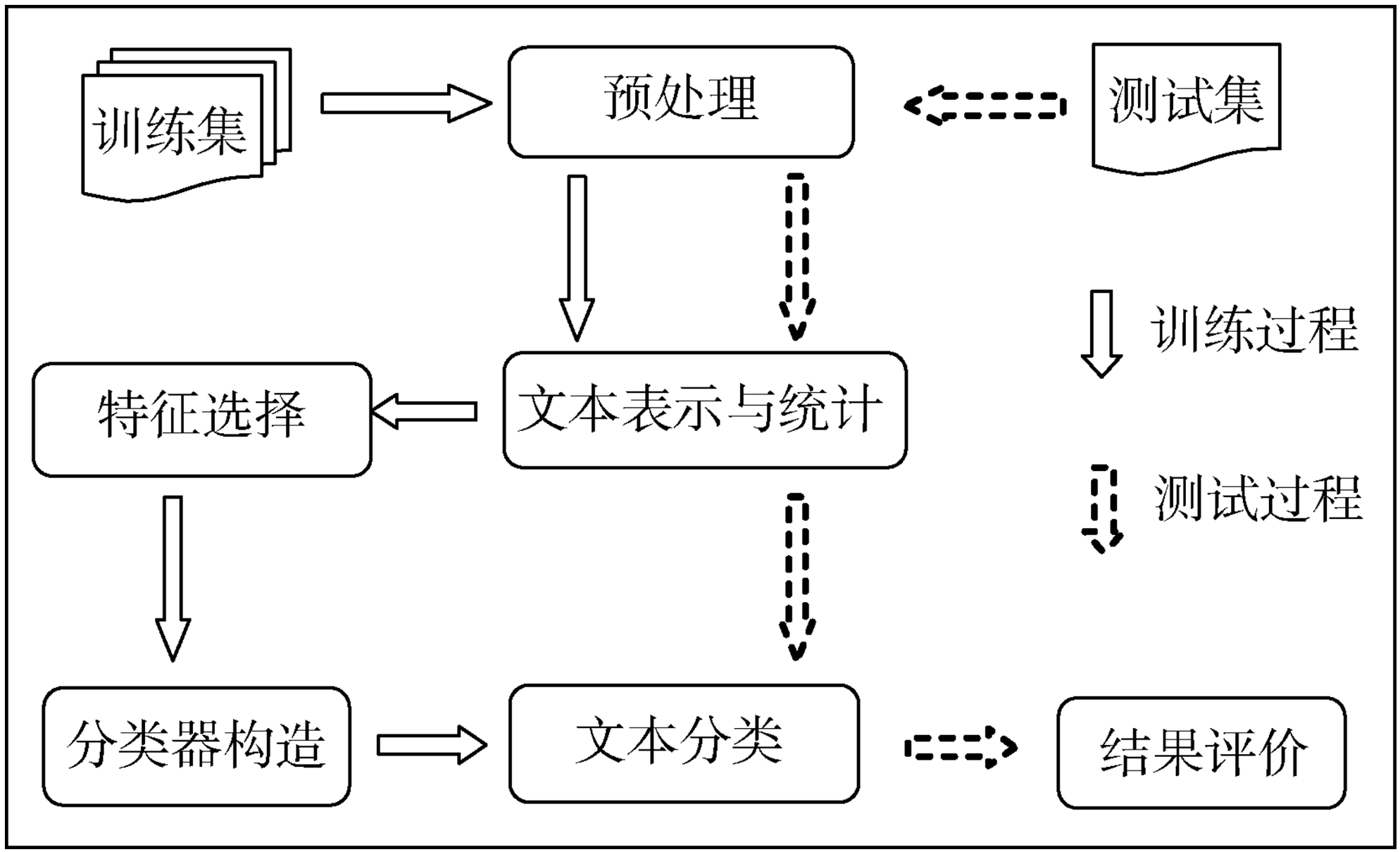
新闻类别与标题

财经 一个经营亏损的澡堂，2至3个月时间内如何收回现金300万
国际 7个绞死的甲级战犯，只有1个家属拒绝加入靖国神社，连骨灰都不要
汽车 东风风神AX7大战启辰T90
房产 惊！常德房价一涨再涨，直逼长沙一线楼盘！
文化 林则徐晚年为何提倡种植鸦片?
教育 30所中小学拟9月前投用 新建项目11月前全部开工
体育 如果火箭止步于西部决赛，下赛季火箭该如何补强?
文化 家里的装修太过时？只需一幅山水画，客厅立马高大尚
旅游 火了！沈阳这个区拥有国字号旅游桂冠，还开启了公交旅游新时尚！
财经 中资地产保险两大板块强势拉升 恒指收涨1.17%报30344点
财经 紫金财险分公司副总虚构交易套取一千多万 被监管撤职
农村 安徽省加大就业脱贫力度 确保零就业贫困户至少1人实现就业
国际 普京要求俄在2024年前成为全球五大经济体之一
国际 日外相：对美退出表示遗憾 日方将继续支持伊核协议
农村 玉米在土壤墒情不同时，播种深度也有大学问！
科技 京东成立时尚科技研究院 科技赋能时尚
旅游 杭州必去十大景区，去过七个的算合格，你去过几个?
娱乐 贾乃亮首次更博，从嘻哈帅哥变成郁郁寡欢判若两人，心疼亮哥！

文本分类


Acme Article
Acme Article
Technology
Foo Article
Sports
Bar Article
Entertainment

文本分类的基本流程



文本表示——词袋模型

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1

特征选择

- 很简单地选择高频词作为特征
- 去除掉了数字和停用词
- 思考：有没有更好的特征表示方式?

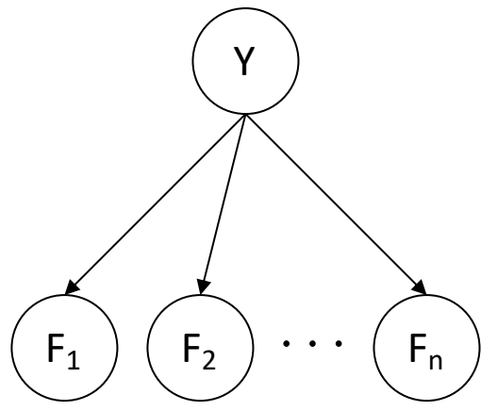
```
def get_feature_words(all_words_list, stopwords_set):
    feature_words = []
    for t in range(0, len(all_words_list), 1):
        if len(feature_words) > feature_size:  # feature_size是feature_words的维度
            break
        if not all_words_list[t].isdigit() and all_words_list[t] not in stopwords_set
and 1 < len(
                all_words_list[t]) < 5:
            feature_words.append(all_words_list[t])
    return feature_words
```

朴素贝叶斯 Naïve Bayes

- 朴素贝叶斯：假设所有特征都是标签的独立效果
- 文本分类版本
  - 每个词作为一个特征（变量） $F_{i,j}$
  - 特征取值为 1 / 0, 基于该词是否出现
  - 每个输入映射到一个特征向量, 例如

    现货价格继续松动 玻璃期货上行难度较大 -> [0,0,0,0,1,0,0,0,1,1,0,….,0]]

  - 有很多特征，每个特征都是二值的



- 朴素贝叶斯模型： $P(Y|F_{0,0}\dots F_{15,15}) \propto P(Y)\prod_{i,j} P(F_{i,j}|Y)$
- 需要学习的是什么?

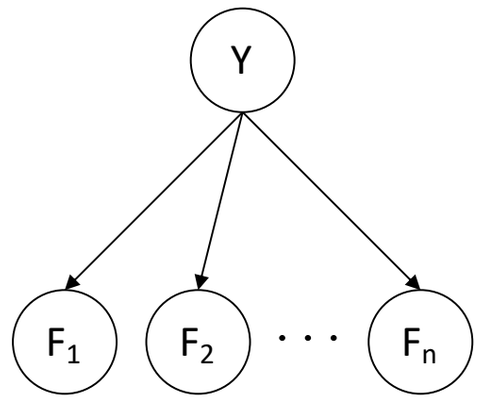
朴素贝叶斯

- 一般的朴素贝叶斯模型：

|Y| parameters

$$P(\mathsf{Y}, \mathsf{F}_1 \ldots \mathsf{F}_n) \propto P(\mathsf{Y}) \prod_i P(\mathsf{F}_i|\mathsf{Y})$$

|Y| x |F|^n values



n x |F| x |Y|
parameters

- 我们只需要指定每个特征如何依赖于类别
- 参数总数以n为单位呈线性
- 模型非常简单，但通常都是有效的

朴素贝叶斯的模型推断

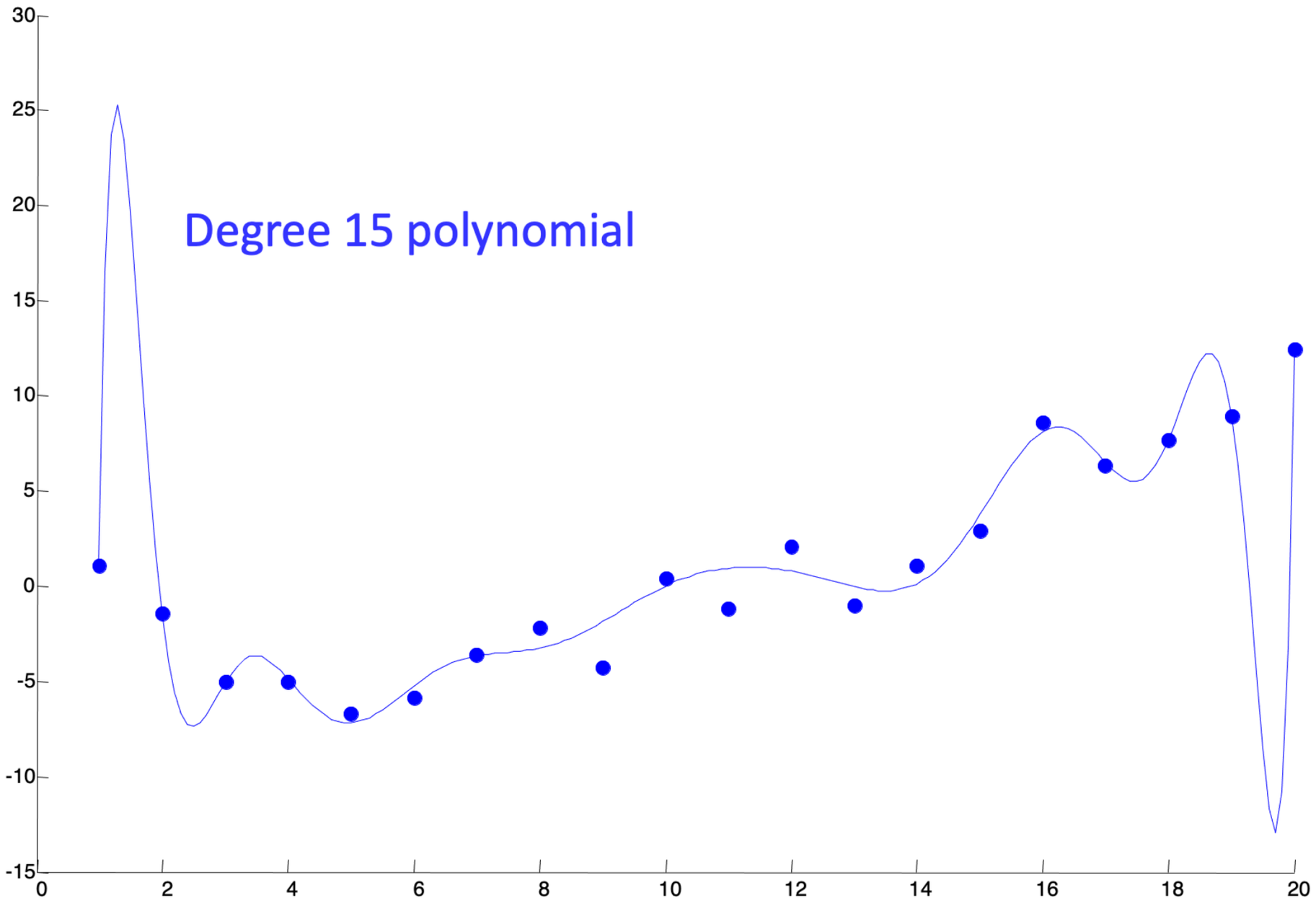
- 目标：计算标签变量Y的后验分布
  - 第1步：为每个标签获取标签和证据(特征)的联合概率

$$P(Y, f_1 \dots f_n) = \begin{bmatrix} P(y_1, f_1 \dots f_n) \\ P(y_2, f_1 \dots f_n) \\ \vdots \\ P(y_k, f_1 \dots f_n) \end{bmatrix} \Rightarrow \begin{bmatrix} P(y_1) \prod_i P(f_i|y_1) \\ P(y_2) \prod_i P(f_i|y_2) \\ \vdots \\ P(y_k) \prod_i P(f_i|y_k) \end{bmatrix}$$

$$\frac{\begin{bmatrix} P(y_1) \prod_i P(f_i|y_1) \\ P(y_2) \prod_i P(f_i|y_2) \\ \vdots \\ P(y_k) \prod_i P(f_i|y_k) \end{bmatrix}}{P(f_1 \dots f_n)} \Rightarrow P(Y|f_1 \dots f_n)$$

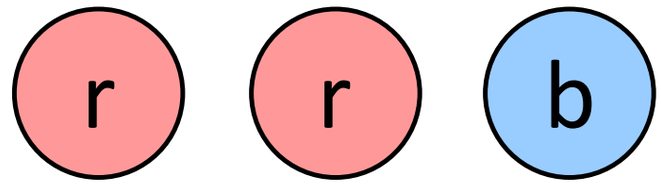
  - 第2步：求和得到证据的概率
  - 第3步：将步骤1除以步骤2进行规范化

过拟合


Degree 15 polynomial
30
25
20
15
10
5
0
-5
-10
-15
0
2
4
6
8
10
12
14
16
18
20

拉普拉斯平滑

- 拉普拉斯估计：
  - 假装你比实际看到的每一个结果都多看一次



$$P_{LAP}(x) = \frac{c(x)+1}{\sum_x [c(x)+1]} \qquad P_{ML}(X) = \left\langle \frac{2}{3}, \frac{1}{3} \right\rangle$$

$$= \frac{c(x)+1}{N+|X|} \qquad P_{LAP}(X) = \left\langle \frac{3}{5}, \frac{2}{5} \right\rangle$$

Sklearn中的朴素贝叶斯

```
from sklearn.naive_bayes import MultinomialNB

# 调用sklearn的MultinomialNB，训练朴素贝叶斯模型
classifier = MultinomialNB().fit(train_feature_list,
                                 train_class_list)

# 用训练好的模型对测试数据进行分类
test_class_pred = classifier.predict(test_feature_list)
```

结果评价

	precision	recall	f1-score	support
文化	0.91	0.87	0.89	7512
娱乐	0.78	0.78	0.78	5027
体育	0.78	0.75	0.76	3830
财经	0.76	0.72	0.74	5409
房产	0.81	0.84	0.83	7925
汽车	0.84	0.84	0.84	3569
教育	0.81	0.84	0.82	5380
科技	0.85	0.72	0.78	5556
军事	0.71	0.77	0.74	4300
旅游	0.87	0.86	0.86	7106
国际	0.86	0.84	0.85	5841
证券	0.64	0.76	0.69	1273
农村	0.74	0.82	0.77	8314
游戏	0.00	0.00	0.00	68
社会	0.74	0.73	0.73	5428
accuracy			0.80	76538
macro avg	0.74	0.74	0.74	76538
weighted avg	0.80	0.80	0.80	76538

实验要求

- 自己实现朴素贝叶斯算法
    - 实现拉普拉斯平滑
- 基于自己实现的算法进行文本分类
- 探索更好的特征提取方法，提高分类的准确率、召回率、F值