朴素贝叶斯分类器
Naive Bayes Classifier
Movie Reviews

课程版本：v1.0

扫描二维码关注微信/微博
获取最新面试题及权威解答

微信: ninechapter
微博: http://www.weibo.com/ninechapter
知乎: http://zhuanlan.zhihu.com/jiuzhang
官网: http://www.jiuzhang.com

版权声明

九章的所有课程均受法律保护，不允许录像与传播录像

一经发现，将被追究法律责任和赔偿经济损失

Naïve Bayes

三种模型

伯努利型贝叶斯

Bernoulli Naïve Bayes

伯努利型贝叶斯

- Bernoulli Naïve Bayes
- 随机变量x满足伯努利分布（Bernoulli Distribution），也叫二项分布
- 丢硬币就是一个经典的Bernoulli Distribution，意思就是一个事件要么成功要么失败。
- 事件成功，则随机变量取值为1。若事件失败，则伯努利随机变量取值为0。
- 若成功概率p，失败概率为 q=1-p

伯努利型贝叶斯

- 对于这个垃圾邮件问题，如果一句话里面出现重复词语。比如：
- “A Great Great Problem”
- 伯努利型就视作重复的词语都只出现1次考虑
- $P(\text{“A } Great \text{ Greate } Problem | Spam\text{”}) = P(\text{“A} | Spam\text{”}) * P(\text{“}Great | Spam\text{”}) * P(\text{“}Problem | Spam\text{”})$
- 并且
- $P(\text{“}Great | Spam\text{”}) = \frac{\text{出现“}Great\text{”的}Spam\text{的封数}}{\text{所有}spam\text{邮件中所有词出现次数（出现了次只计算一次）的总和}} = \frac{1}{5}$

邮件	是否是垃圾邮件
A Great Problem	Ham
A Great Great Game	Spam
I Love You	Ham
A Complex Problem	Spam
A Great Great Problem	Ham

多项式型贝叶斯

Multinomial Naïve Bayes

多项式型贝叶斯

- Multinomial Naïve Bayes
- 随机变量x满足多项式分布（ Multinomial Distribution ）
- 丢骰子就是经典的Multinomial Distribution，二项式的拓展
- 骰子有6个面对应6个不同的点数，这样单次每个点数朝上的概率分别是{p1…,p6}
- 如果做下面实验重复扔n次，如果问有m次都是点数3朝上的概率就是 $C(n,m)p_3^m(1-p_3)^{(n-m)}$

多项式型贝叶斯

- 对于这个垃圾邮件问题，如果一句话里面出现重复词语。比如：
- “A Great Great Problem”
- 多项式型贝叶斯就对于重复的词语都也会视作多次
- $P(\text{“A } Great \text{ Greate } Problem | Spam\text{”}) = P(\text{“A } | Spam\text{”}) * P(\text{“}Great | Spam\text{”})^2 * P(\text{“}Problem | Spam\text{”})$
- $P(\text{“}Great | Spam\text{”}) = \frac{\text{每封出现“}Great\text{”的}Spam\text{的次数总和}}{\text{所有}spam\text{邮件中所有词出现次数（计算重复次数）的总和}} = \frac{2}{6}$

邮件	是否是垃圾邮件
A Great Problem	Ham
A Great Great Game	Spam
I Love You	Ham
A Complex Problem	Spam
A Great Great Problem	Ham

高斯型贝叶斯

Gaussian Naïve Bayes

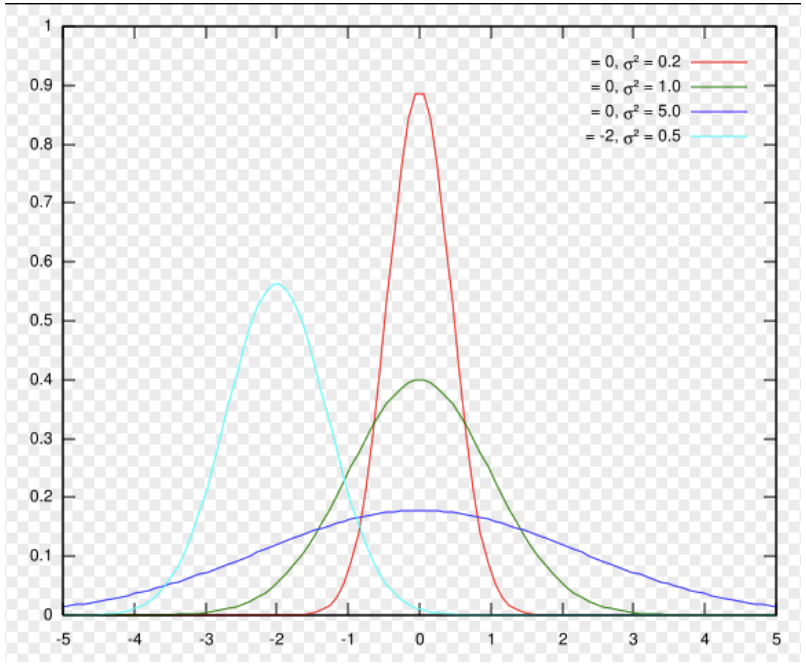
高斯型贝叶斯

- 刚才的随机变量都是离散型分布。
- 我们来看下面一个例子， 已知某人身高6、体重130，脚掌8，请预测该人是男是女？
- 已知某人身高6英尺、体重130磅，脚掌8英寸，请问该人是男是女？
- 问题：
  - 由于身高、体重、脚掌都是连续变量，不能采用离散变量的方法计算概率。
  - 而且由于样本太少，所以也无法分成离散区间计算。怎么办？

性别	身高	体重	脚掌
男	6	180	12
男	5.92	190	11
男	5.58	170	12
男	5.92	165	10
女	5	100	6
女	5.5	150	8
女	5.42	130	7
女	5.75	150	9

高斯型贝叶斯

- 假设男性和女性的身高、体重、脚掌都是正态分布（高斯分布）
- 通过样本计算出均值和方差，也就是得到正态分布的密度函数
- $P(x \mid y_k) = \frac{1}{\sqrt{2\pi\delta_k^2}} e^{-\frac{(x-\mu_k)^2}{2\delta_k^2}}$



- δ_k是x 在第k个分类的标准差，δ_k^2 是方差
- μ_k是x 在的k个分类的期望值
- 比如男性的身高是均值5.855、方差0.035的正态分布
- $P(\text{身高} \mid \text{男}) = \frac{1}{\sqrt{2\pi 0.035}} e^{-\frac{(\text{身高}-5.855)^2}{2*0.035}}$
- 男性的身高为6的概率的相对值等于1.5789
- 大于1并没有关系，因为这里是密度函数的值，只用来反映各个值的相对可能性

高斯型贝叶斯

- 已知某人身高6、体重130，脚掌8，请问该人是男是女？
- $P(男|身高=6，体重=130，脚掌=8) \propto$
  - $P(身高=6|男) x\ P(体重=130|男) x\ P(脚掌=8|男) x\ P(男) =$
  - 6.1984 x 10^{-9}
- $P(女|身高=6，体重=130，脚掌=8) \propto$
  - $P(身高=6|女)\ x\ P(体重=130|女)\ x\ P(脚掌=8|女)\ x\ P(女)$
  - 5.3778 x 10^{-4}
- 所以这个人只为男的可能性大

采用哪种模型

关键看具体的场景

TF-IDF

Term frequency–Inverse document frequency

TF-IDF

$$P(\text{“}Spam|A\ Great\ \text{Great}\ Problem\text{”}) \propto P(\text{“A}\mid Spam\text{”})\ * P(\text{“}Great|\ Spam\text{”})^2\ * P(\text{“}Problem|\ Spam\text{”})\ * P(\text{“}Spam\text{”})$$

$$\begin{aligned}&\text{Log}(P(\text{“}Spam|A\ Great\ \text{Great}\ Problem\text{”}))\\ &= Log(P(\text{“A}\mid Spam\text{”})) + 2 * Log(P(\text{“}Great|\ Spam\text{”})) + Log(P(\text{“}Problem|\ Spam\text{”})) + Log(P(\text{“}Spam\text{”}))\\ &= sum(词频率*Log(词概率)\ ) + Log(P(\text{“}Spam\text{”}))\end{aligned}$$

回归数学

- 如果 $(\{x_1, x_2, \dots, x_n\}, y)$ 表示一个数据样例, x_i是第i个向量，y是标签
  - （比如x_i是第i个单词，y 是垃圾或者正常邮件这种分类）
- 贝叶斯定理定义
  - $P(y \mid x_1, x_2, \dots, x_n) = \frac{P(x_1, x_2, \dots, x_n \mid y) * P(y)}{P(x_1, x_2, \dots, x_n)} \propto P(x_1, x_2, \dots, x_n \mid y) * P(y)$
- 因为贝叶斯假设 $\{x_1, x_2, \dots, x_n\}$ 之间相互独立
  - $P(x_1, x_2, \dots, x_n \mid y) = \prod_{i=1}^{n} P(x_i \mid y)$
  - $P(y \mid x_1, x_2, \dots, x_n) \propto P(x_1, x_2, \dots, x_n \mid y) * P(y) = \prod_{i=1}^{n} P(x_i \mid y) * P(y)$
- 因为有y是我们的类别， 我们求解的是使得 $P(y \mid x_1, x_2, \dots, x_n)$ 最大的时候， y的类别
  - 如果有m种类别 $\{y_1, y_2 \dots y_m\}$
  - 问题是求 $P(y_1 \mid x_1, x_2, \dots, x_n)$, $P(y_2 \mid x_1, x_2, \dots, x_n) \cdots P(y_m \mid x_1, x_2, \dots, x_n)$ 谁概率最大
- 最终公式
  - $\hat{y} = \arg max_{j=1}^{m} \{P(y_j \mid x_1, x_2, \dots, x_n)\} = \arg max_{j=1}^{m} \{\prod_{i=1}^{n} P(x_i \mid y_j) * P(y_j)\}$
  - Maximum a posteriori estimation（最大后验估计）

更通俗

$$\hat{y} = \arg max_{j=1}^{m}\{P(y_j \mid x_1, x_2, \dots, x_n)\} = \arg max_{j=1}^{m}\{\prod_{i=1}^{n} P(x_i|y_j)^{f_i} * P(y_j)\}, tc_i \text{是第i个变量的次数}$$

$$= \arg max_{j=1}^{m}\{\sum_{i=1}^{n} tc_i \log P(x_i|y_j) + \log P(y_j)\}$$

Term frequency–Inverse document frequency

- **词频**（term frequency，tf）指的是某一个给定的词语在该文件中出现的频率
  - **词数**（term count）的归一化，以防止它偏向长的文件
  - 表示词语在一个句子的重要性
  - 如果词w在文档d中出现次数count(w, d)和文档d中总词数size(d)
  - tf = count(w, d) / size(d)
  - tf(Great, id =2) = 2/ 4
- **逆向文件频率**（inverse document frequency，idf）总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到
  - 表示词语在所有句子的重要性,词语出现的越多，说明词语越通用，但是越没有意义，比如the
  - 文档总数n与词w所出现文件数docs(w, D)
  - idf = log(n / docs(w, D))
  - Idf(word) = log(5/3)
- TF-IDF = tf*idf

标号	邮件	是否是垃圾邮件
1	A Great Problem	Ham
2	A Great Great Game	Spam
3	I Love You	Ham
4	A Complex Problem	Spam
5	A Great Great Problem	Ham

TF-IDF

$$\hat{y} = \arg max_{j=1}^{m}\{P(y_j \mid x_1, x_2, \dots, x_n)\} = \arg max_{j=1}^{m}\{\sum_{i=1}^{n} tfidf_i \log P(x_i|y_j) + \log P(y_j)\}$$

$tfidf_i$是第i个变量的$tfidf$值

调查问卷

http://www.jiuzhang.com/course/11/questionnaire/

QA

谢谢大家

扫描二维码关注微信/微博
获取最新面试题及权威解答

微信: ninechapter
微博: http://www.weibo.com/ninechapter
知乎: http://zhuanlan.zhihu.com/jiuzhang
官网: http://www.jiuzhang.com