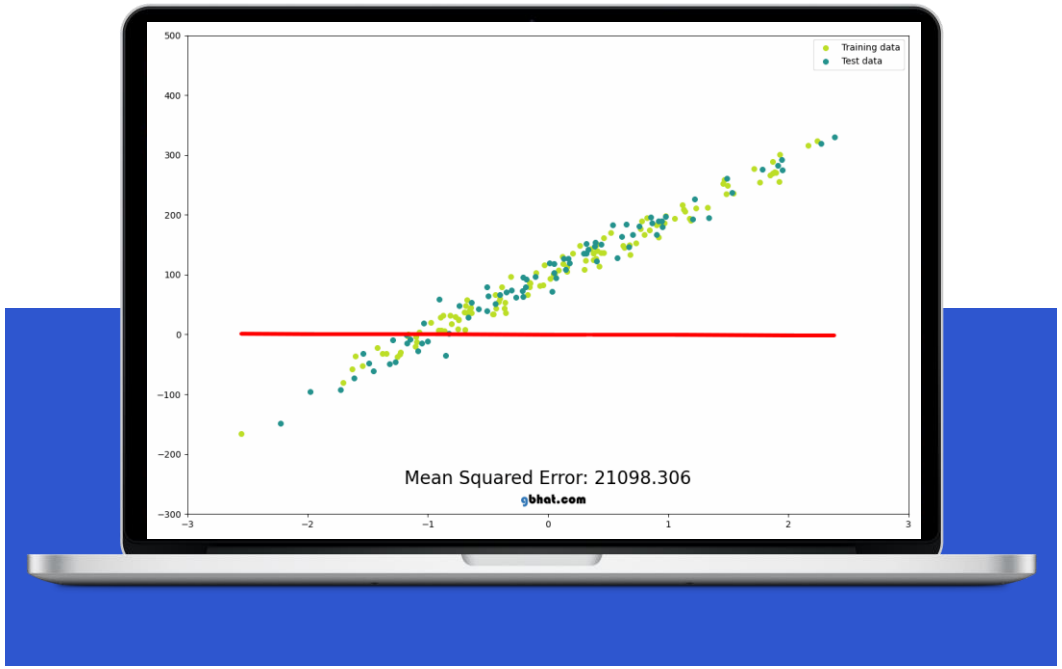
《机器学习》

第四章 朴素贝叶斯

机器学习就是对计算机一部分数据进行学习，然后对另外一些数据进行预测与判断。核心是“使用算法解析数据，从中学习，然后对新数据做出决定或预测”。

2023年10月

上节回顾



监督学习

回归 Regression

线性回归

$$h(x) = w^T x + b$$

分类 Classification

逻辑回归

$$h(x) = \text{Sigmoid}(w^T x + b)$$

支持向量机

决策树

分类问题

线性模型 不适用于文本等高维度特征数据

(a) 公共安全　　(b) 金融风控　　(c) 媒体监管

目录/CONTENTS

01 贝叶斯方法

02 朴素贝叶斯原理

03 朴素贝叶斯案例

04 朴素贝叶斯代码实现

目录/CONTENTS

01 贝叶斯方法

02 朴素贝叶斯原理

03 朴素贝叶斯案例

04 朴素贝叶斯代码实现

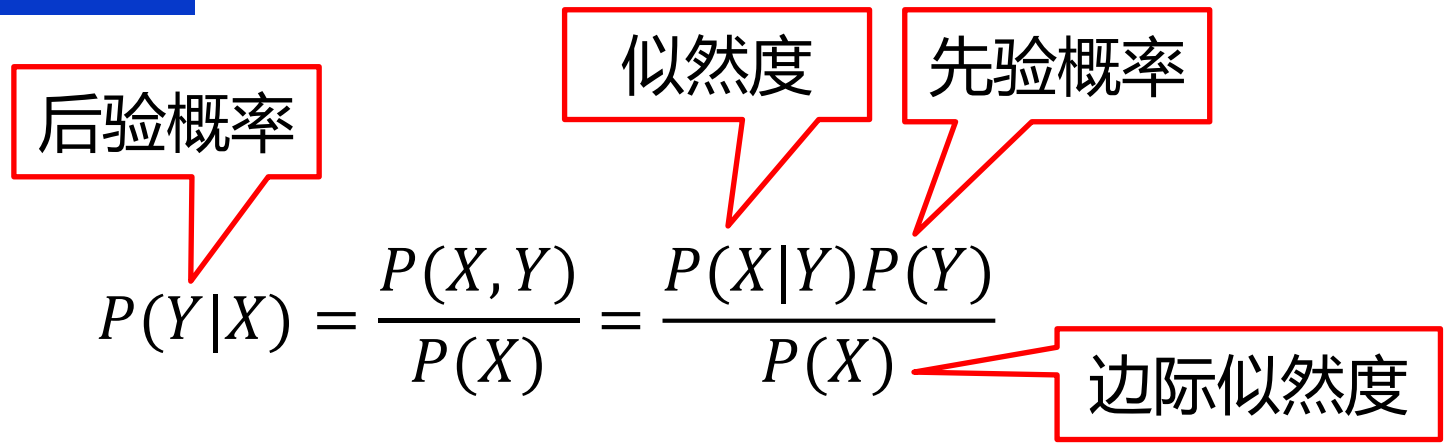
背景知识

贝叶斯分类：贝叶斯分类是一类分类算法的总称，这类算法均以贝叶斯定理为基础。

- 先验概率：根据以往经验和分析得到事件 Y 的概率 $P(Y)$ 。
- 后验概率：根据已经发生的事件X来分析得到事件 Y 的概率 $P(Y|X)$。

贝叶斯方法

贝叶斯公式



- 朴素贝叶斯法是典型的生成学习方法。生成方法由训练数据学习联合概率分布 $P(X,Y)$，然后求得后验概率分布$P(Y|X)$。
- 具体来说，利用训练数据学习$P(X|Y)$和$P(Y)$的估计，得到联合概率分布：

  $P(X,Y) = P(X|Y)\ P(Y)$

Thomas Bayes

LII. *An Eſſay towards ſolving a Problem in the Doctrine of Chances. By the late Rev. Mr.* Bayes, *F. R. S. communicated by Mr.* Price, *in a Letter to* John Canton, *A. M. F. R. S.*

Dear Sir,

Read Dec. 23, 1763. I Now ſend you an eſſay which I have found among the papers of our deceaſed friend Mr. Bayes, and which, in my opinion, has great merit, and well deſerves to be preſerved. Experimental philoſophy, you will find, is nearly intereſted in the ſubject of it; and on this account there ſeems to be particular reaſon for thinking that a communication of it to the Royal Society cannot be improper.

T. Bayes.

1701-1761

目录/CONTENTS

01 贝叶斯方法

02 朴素贝叶斯原理

03 朴素贝叶斯案例

04 朴素贝叶斯代码实现

朴素贝叶斯原理

贝叶斯法是典型的生成学习方法

生成方法由训练数据学习联合概率分布 $P(X,Y)$，然后求得后验概率分布 $P(Y|X)$。

具体来说，利用训练数据学习 $P(X|Y)$ 和 $P(Y)$ 的估计，得到联合概率分布：

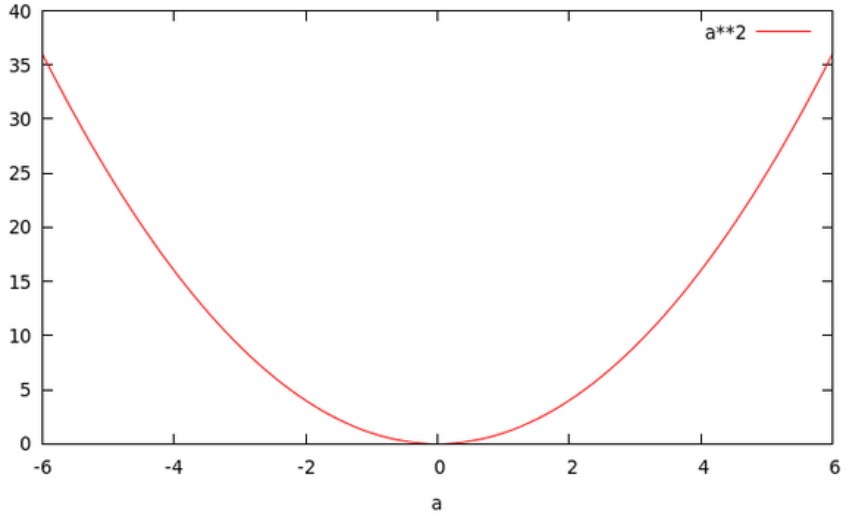
$$P(X,Y)=P(Y)P(X|Y)$$

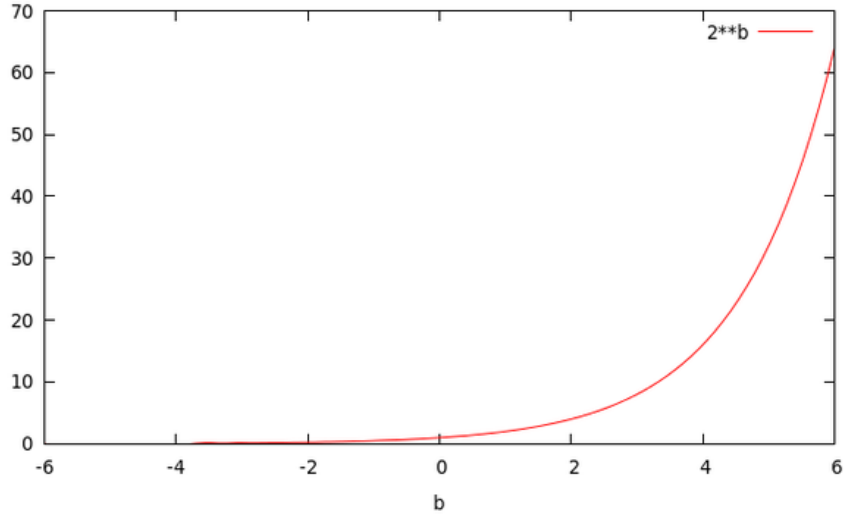
概率估计方法可以是极大似然估计或贝叶斯估计。

概率 vs. 似然

- 假设有个基于参数θ的概率模型 p(x | θ)：
  - 给定θ 下 **x** 的概率 probability,
  - 观察到事件**x**条件下**θ**的**似然 likelihood**。

- 举例：函数a^b的两种形态





概率 vs. 似然

- 假设有个基于参数θ的概率模型 p(x | θ)：
  - 给定θ 下 **x** 的概率,
  - 观察到事件**x**条件下**θ的似然。**

- **极大似然估计** (Maximum Likelihood Estimation, **MLE**) :频率学派认为 θ 是一个值，选择最大化 p(x | θ)的θ.

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{\theta}) = f(x_1, x_2, \ldots, x_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})$$

$$\hat{\boldsymbol{\theta}}_{mle} = arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\boldsymbol{x})$$

- **最大后验估计(**Maximum A Posteriori estimation, **MAP):**贝叶斯学派认为θ 是一个变量，选择最大化p(θ | x) 的θ。

$$\hat{\boldsymbol{\theta}}_{map} = arg\max_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta}|\boldsymbol{x}) = arg\max_{\boldsymbol{\theta}} \frac{f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{m(\boldsymbol{x})} = arg\max_{\boldsymbol{\theta}} f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

概率 vs. 似然

有一个硬币，它有θ的概率会正面向上，有$1-\theta$的概率反面向上。θ是存在的，但不知道是多少。为了获得θ的值，你做了一个实验：

- 将硬币抛10次，得到了一个正反序列：

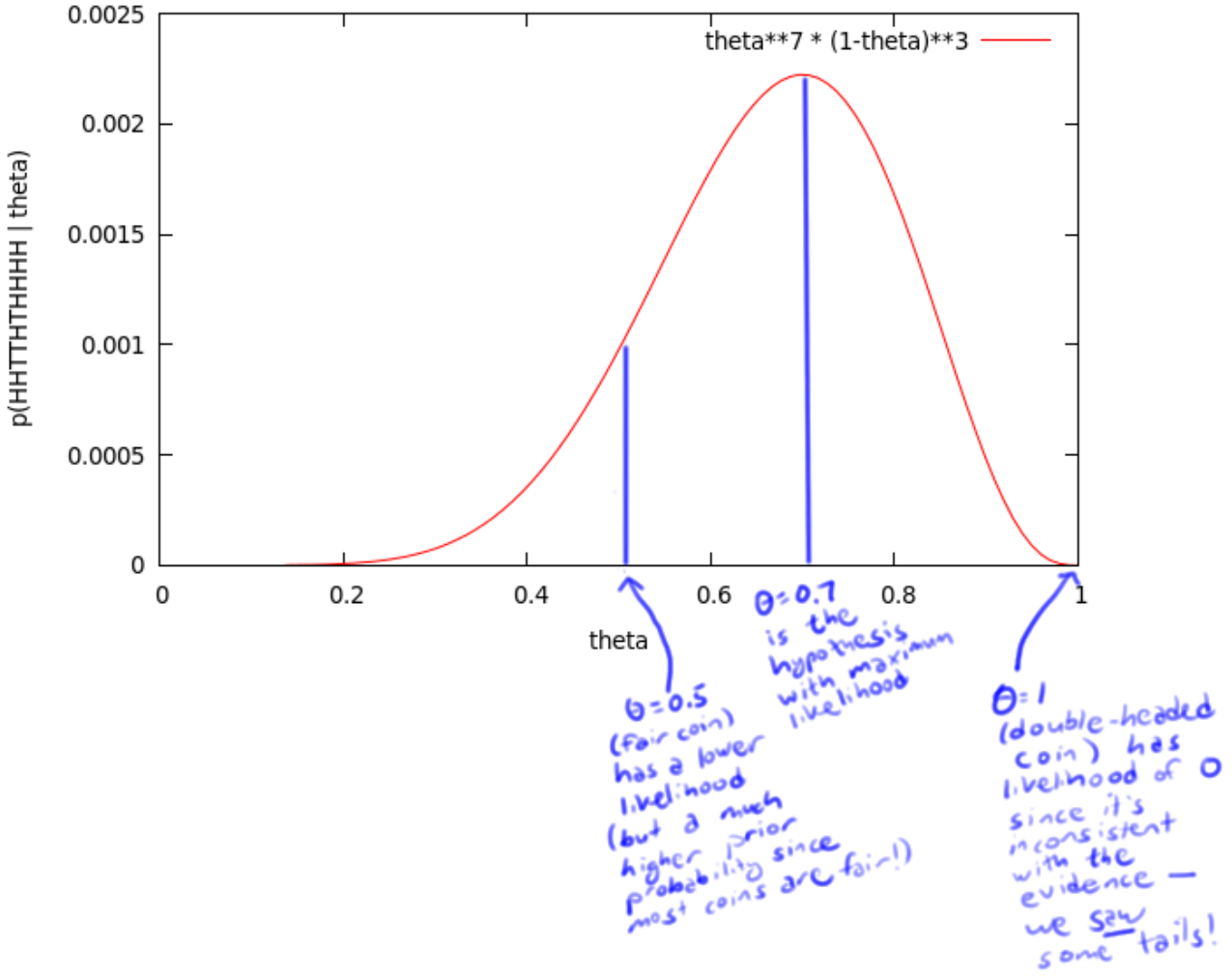
  $\boldsymbol{x}$ = 正正反反正反正正正正

- 无论θ的值是多少，这个序列的概率值为

  $\theta \cdot \theta \cdot (1-\theta) \cdot (1-\theta) \cdot \theta \cdot (1-\theta) \cdot \theta \cdot \theta \cdot \theta \cdot \theta = \theta^7 (1-\theta)^3$

  - 如果θ值为0，则得到这个序列的概率值为0。
  - 如果θ值为1/2，概率值为1/1024。

- 右图显示最大似然估计的取值为0.7



贝叶斯模型预测

贝叶斯法分类时，对给定的输入x，需要预测类别$\hat{y}$。根据贝叶斯定理:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

可以计算后验概率

$$P(Y = c_k|X = x) = \frac{P(X = x|Y = c_k)P(Y = c_k)}{P(X = x)} = \frac{P(X = x|Y = c_k)P(Y = c_k)}{\sum_{k=1}^{K} P\,(X = x|Y = c_k)P(Y = c_k)}$$

将后验概率最大的类c_k作为x的类输出：

$$\hat{y} = \arg\max_{c_k} P(Y = c_k|X = x)$$

朴素贝叶斯原理

朴素贝叶斯法的基本假设是条件独立性

$$P(X=x|Y=c_k)=P\left(x^{(1)},\cdots,x^{(n)}|Y=c_k\right)=\prod_{j=1}^{n}P\left(x^{(j)}|Y=c_k\right)$$

c_k代表类别，k代表类别个数

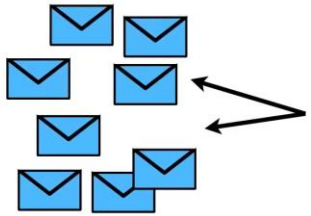
研表究明
汉字的序顺并不定一会影阅响读
比如当看你完这句话后
才发这现里的字全是都乱的

- 优点：高效，且易于实现。
- 缺点：分类的性能不一定很高。

朴素贝叶斯原理

朴素贝叶斯法的基本假设是条件独立性

$$P(X=x|Y=c_k)=P\left(x^{(1)},\cdots,x^{(n)}|Y=c_k\right)=\prod_{j=1}^{n}P\left(x^{(j)}|Y=c_k\right)$$



现在想象一下，我们收到了
来自朋友和家人的正常消息。

这次让我们尝试对这条消息
进行分类

目录/CONTENTS

01 贝叶斯方法

02 朴素贝叶斯原理

03 朴素贝叶斯案例

04 朴素贝叶斯代码实现

朴素贝叶斯案例

假设我们正在构建一个分类器，该分类器说明文本是否与运动(Sports)有关。

我们的训练数据有5句话：

文本	标签
A great game	Sports
The election was over	Not Sports
Very clean match	Sports
A clean but forgettable game	Sports
It was a close election	Not Sports

我们想要计算句子 “A very close game” 是 Sports 的概率以及它不是 Sports 的概率。

即 P(Sports | a very close game) 这个句子的类别是 Sports 的概率

朴素贝叶斯案例

特征：单词的频率

已知贝叶斯定理 $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$，则：

$$P(\text{ Sports } | \text{ a very close game })$$

$$= \frac{P(\text{ a very close game } | \text{ Sports }) \times P(\text{ Sports })}{P(\text{ a very close game })}$$

由于我们只是试图找出哪个类别有更大的概率，可以舍弃除数，只是比较

$P(\text{ a very close game } | \text{ Sports }) \times P(\text{ Sports })$ 和

$P(\text{ a very close game } | \text{ Not Sports }) \times P(\text{ Not Sports })$

朴素贝叶斯案例

朴素Naive：我们假设一个句子中的每个单词都与其他单词无关。

$$P(\text{ a very close game }) = P(a) \times P(\text{ very }) \times P(\text{ close }) \times P(\text{ game })$$

$$P(\ a\ very\ close\ game|Sports\ )$$
$$= P(a|\ Sports\ ) \times P(\ very\ |\ Sports) \times P(\ close\ |\ Sports\ ) \times P(\ game\ |\ Sports\ )$$

朴素贝叶斯案例

计算每个类别的先验概率

对于训练集中的给定句子，$P(\text{Sports})$的概率为⅗。$P(\text{Not Sports})$是⅖。

文本	标签
A great game	Sports
The election was over	Not Sports
Very clean match	Sports
A clean but forgettable game	Sports
It was a close election	Not Sports

在计算 $P(game|Sports)$ 就是“game”有多少次出现在 Sports 的样本，然后除以 sports 为标签的文本的单词总数（3+3+5=11）。

$$P(game|Sports) = \frac{2}{11}$$

朴素贝叶斯案例

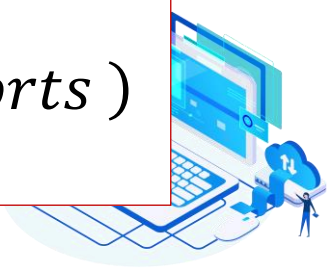
计算每个类别的先验概率

文本	标签
A great game	Sports
The election was over	Not Sports
Very clean match	Sports
A clean but forgettable game	Sports
It was a close election	Not Sports

“close”不会出现在任何 sports 样本中！那就是说

$$P(close|Sports) = 0$$

$$P(\ a\ very\ close\ game|Sports\ )$$
$$= P(a|\ Sports\ ) \times P(\ very\ |\ Sports) \times P(\ close\ |\ Sports\ ) \times P(\ game\ |\ Sports\ )$$
$$= 0$$



朴素贝叶斯案例

通过使用一种称为拉普拉斯平滑的方法：我们为每个计数加1，因此它永远不会为零。为了平衡这一点，我们将可能单词的数量添加到除数中，因此计算结果永远不会大于1。

在这里的情况下，可能单词是

a	great	very	over
it	but	game	election
clean	close	the	was
forgettable	match		

由于可能的单词数是14，因此应用平滑处理可以得到$P(\text{game} \mid \text{sports}) = \frac{2+1}{11+14}$

朴素贝叶斯案例

Word	P (word | Sports)	P (word | Not Sports)
a	$(2+1) \div (11+14)$	$(1+1) \div (9+14)$
very	$(1+1) \div (11+14)$	$(0+1) \div (9+14)$
close	$(0+1) \div (11+14)$	$(1+1) \div (9+14)$
game	$(2+1) \div (11+14)$	$(0+1) \div (9+14)$

$$P(a|\text{ Sports }) \times P(\text{ very }|\text{ Sports }) \times P(\text{ close }|\text{ Sports }) \times$$

$$P(\text{ game }|\text{ Sports }) \times P(\text{ Sports }) = 2.76 \times 10^{-5} = 0.0000276$$

$$P(a|\text{ Not Sports }) \times P(\text{ very }|\text{ Not Sports }) \times P(\text{ close }|\text{ Not Sports }) \times$$

$$P(\text{ game }|\text{ Not Sports }) \times P(\text{ Not Sports}) = 0.572 \times 10^{-5} = 0.00000572$$

目录/CONTENTS

01 贝叶斯方法

02 朴素贝叶斯原理

03 朴素贝叶斯案例

04 朴素贝叶斯代码实现

朴素贝叶斯代码实现

最常用的GaussianNB是高斯贝叶斯分类器

它假设特征的条件概率分布满足高斯分布：

数学期望(mean)

$$P\left(X^{(j)}\middle|y=c_k\right)=\frac{1}{\sqrt{2\pi\sigma_k^2}}exp\left(-\frac{(X^{(j)}-\boldsymbol{\mu_k})^2}{2\sigma_k^2}\right)$$

方差：　$\sigma^2=\frac{\sum(X-\mu)^2}{N}$

朴素贝叶斯代码实现

其他贝叶斯分类器：

- **MultinomialNB**是多项式贝叶斯分类器，它假设特征的条件概率分布满足多项式分布；
- **BernoulliNB**是伯努利贝叶斯分类器。它假设特征的条件概率分布满足二项分布。

朴素贝叶斯代码实现

最常用的GaussianNB是高斯朴素贝叶斯分类器的scikit-learn实现。

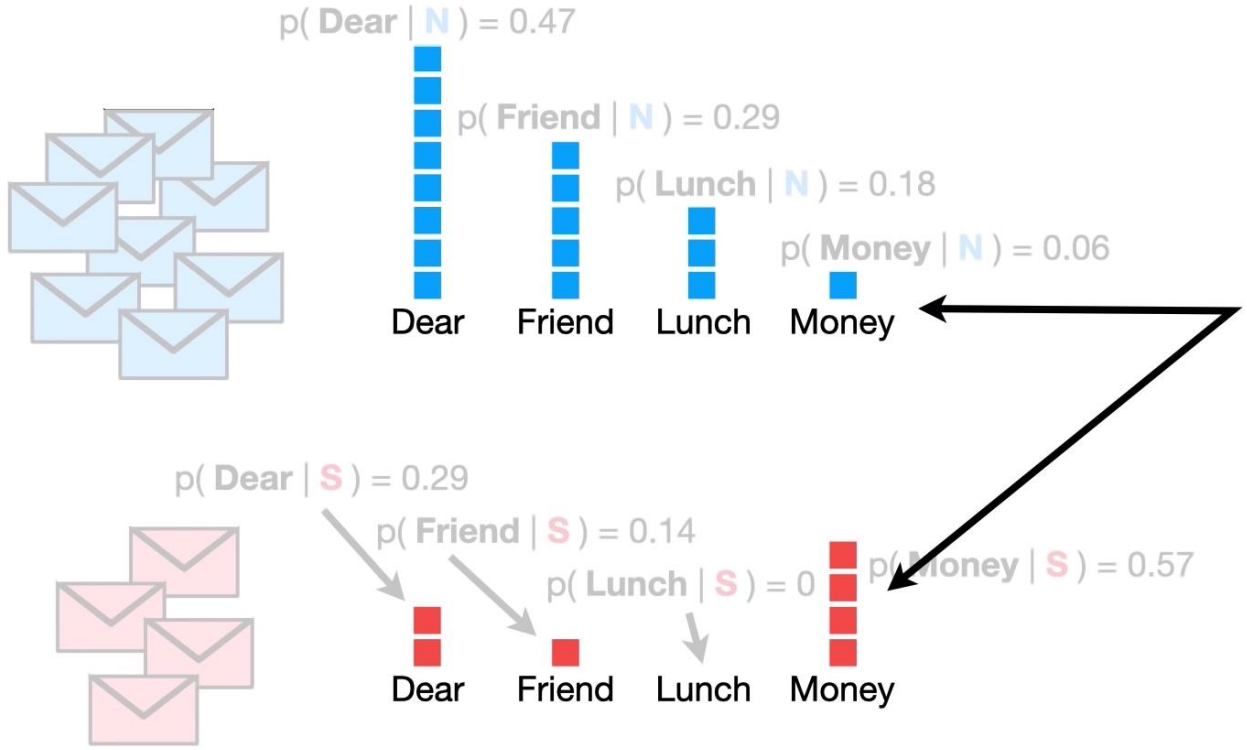
```
import numpy as np
X = np.array([[-1, -1], [-2, -1], [-3, -2], [1, 1], [2, 1], [3, 2]])
Y = np.array([1, 1, 1, 2, 2, 2])
#引入高斯朴素贝叶斯
from sklearn.naive_bayes import GaussianNB
#实例化
clf = GaussianNB()
#训练数据 fit相当于train
clf.fit(X, Y)
#输出单个预测结果
print "==Predict result by predict=="
print(clf.predict([[-0.8, -1]]))
print "==Predict result by predict_proba=="
print(clf.predict_proba([[-0.8, -1]]))
print "==Predict result by predict_log_proba=="
print(clf.predict_log_proba([[-0.8, -1]]))
```

小组PK

V.S

知识小结

1. 贝叶斯理论
2. 朴素贝叶斯分类的基本假设
3. 拉普拉斯修正方法
4. 朴素贝叶斯代码实现



作业与拓展

1. 课后练习：用朴素贝叶斯方法实现文本预测。

   训练数据集见学习通APP。

文本	标签
A great game	Sports
The election was over	Not Sports
Very clean match	Sports
A clean but forgettable game	Sports
It was a close election	Not Sports

2. 拓展思考：调研和探索Thomas Bayes的工作历史背景和动机?

《机器学习》

THANKS

感谢您的观看

2023年10月