

矩阵分解与贝叶斯模型



胡俊峰 2020/11/04

LSI, SVD, & Eigenvectors

➤ SVD decomposes:

➤ Term x Document matrix X as

$$➤ X = U\Sigma V^T$$

➤ Where U, V left and right singular vector matrices, and

➤ Σ is a diagonal matrix of singular values

➤ Corresponds to eigenvector-eigenvalue decomposition: $Y = VL V^T$

➤ Where V is orthonormal and L is diagonal

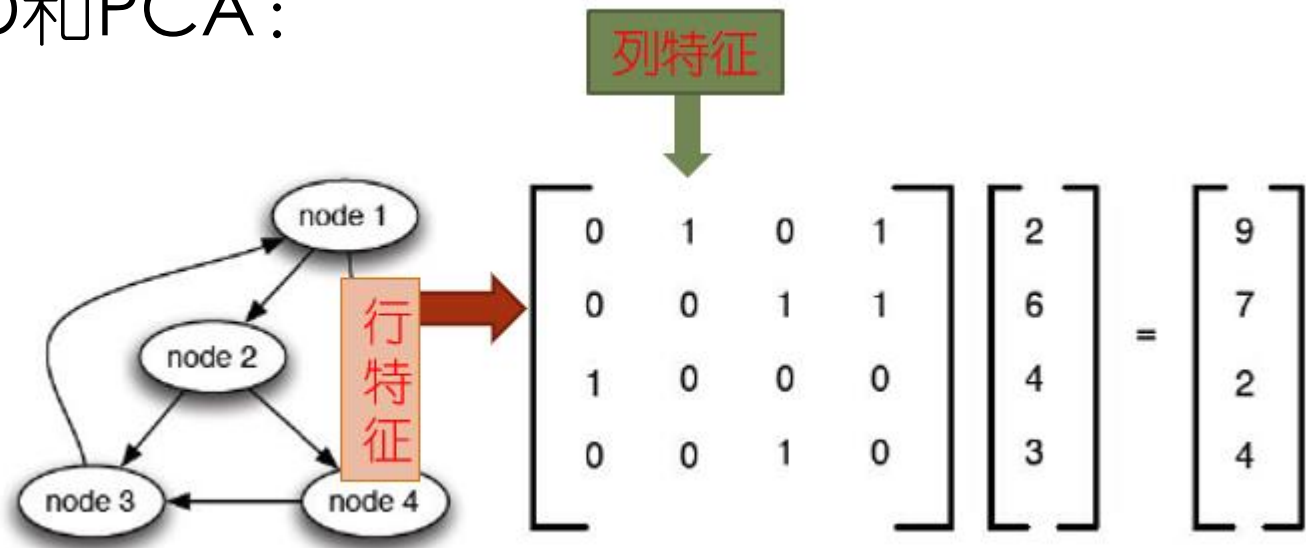
➤ U : matrix of eigenvectors of $Y = XX^T$

➤ V : matrix of eigenvectors of $Y = X^T X$

➤ Σ : diagonal matrix L of eigenvalues

$$\begin{aligned} XX^T &= (U\Sigma V^T)(U\Sigma V^T)^T = (U\Sigma V^T)(V^T \Sigma^T U^T) = U\Sigma V^T V \Sigma^T U^T = U\Sigma \Sigma^T U^T \\ X^T X &= (U\Sigma V^T)^T (U\Sigma V^T) = (V^T \Sigma^T U^T)(U\Sigma V^T) = V \Sigma U^T U \Sigma V^T = V \Sigma^T \Sigma V^T \end{aligned}$$

Hits算法与SVD和PCA:



$$a^k = (M^T M)^{k-1} h^0 \quad \leftarrow \text{对 } M^T M, MM^T \text{ 分别求出一组正交基}$$

$$h^k = (MM^T)^k h^0$$


定理: $n \times n$ 的实对称矩阵有 n 个特征值, 且不同特征值的特征向量彼此正交 (即特征向量构成线性空间的一组基底)

h_0 正规化后连乘迭代收敛于 Z_1 同理 a_0 也会收敛于 Z_1'

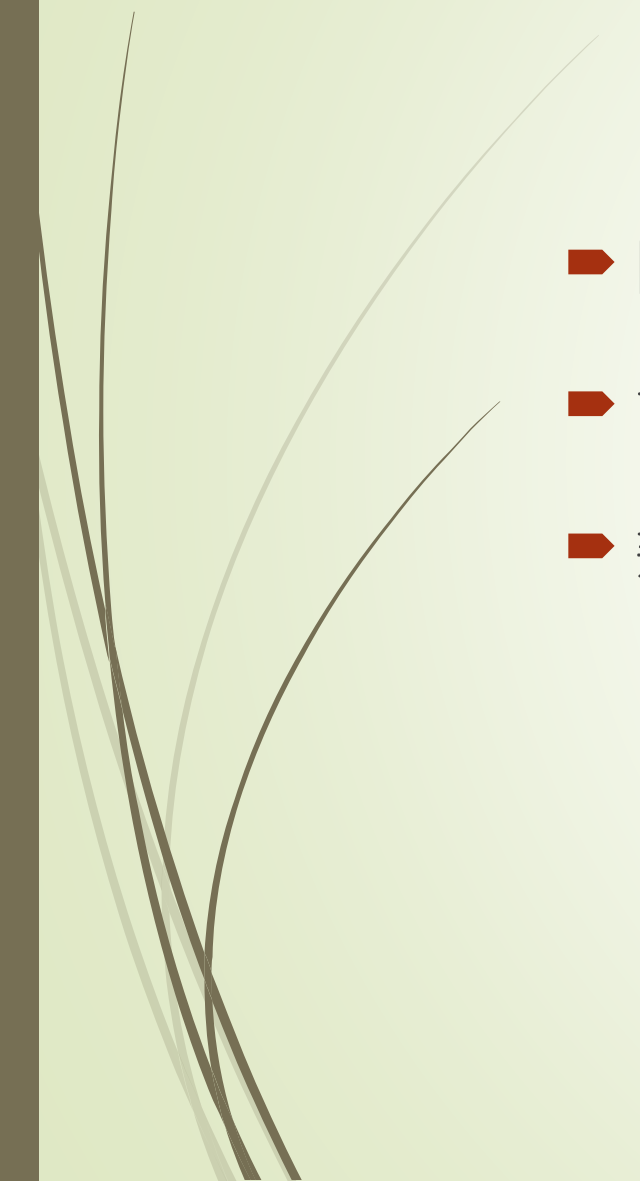
$$h_0 = q_1 z_1 + q_2 z_2 + \dots + q_n z_n$$

M矩阵的SVD

与M矩阵的PCA
差了一个标准化
和中心化



SVD在数据挖掘领域的应用

- ➡ 隐含语义挖掘 (LSI 或称 LSA)
 - ➡ 协同过滤
 - ➡ 数据降噪
- 

LSI (LSA)

- Singular Value Decomposition (SVD) used for the word-document matrix
 - A least-squares method for dimension reduction

	Term 1	Term 2	Term 3	Term 4
Query	user	interface		
Document 1	user	interface	HCI	interaction
Document 2			HCI	interaction

Classic LSI Example (Deerwester)

Titles

- c1: *Human machine interface for Lab ABC computer applications*
- c2: *A survey of user opinion of computer system response time*
- c3: *The EPS user interface management system*
- c4: *System and human system engineering testing of EPS*
- c5: *Relation of user-perceived response time to error measurement*
- m1: *The generation of random, binary, unordered trees*
- m2: *The intersection graph of paths in trees*
- m3: *Graph minors IV: Widths of trees and well-quasi-ordering*
- m4: *Graph minors: A survey*

Terms

Documents

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

2-D Plot of Terms and Docs from Example

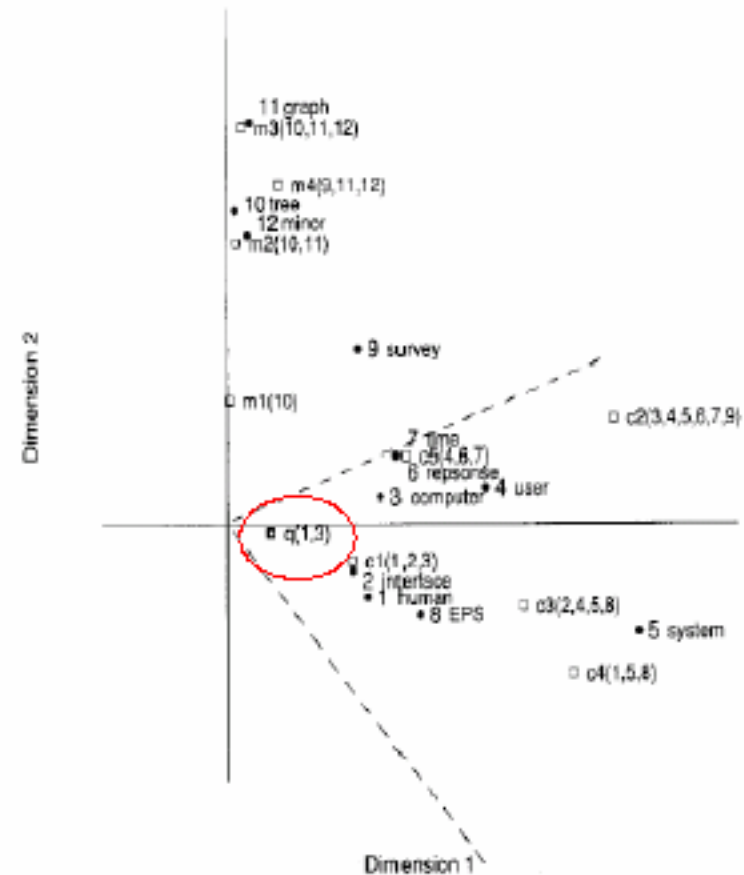


FIG. 1. A two-dimensional plot of 12 Terms and 9 Documents from the sample TM set. Terms are represented by filled circles. Documents are shown as open squares, and component terms are indicated parenthetically. The query ("human computer interaction") is represented as a pseudo document at point q . Axes are scaled for Document-Document or Term-Term comparisons. The dotted cone represents the region whose points are within a cosine of .9 from the query q . All documents about human-computer (c1-c5) are "near" the query (i.e., within this cone), but none of the graph theory documents (m1-m4) are nearby. In this reduced space, even documents c3 and c5 which share no terms with the query are near it.

样本聚类 与非负矩阵分解 $X \approx F * G^T$

—— 参考 <https://www.nature.com/articles/44565.pdf>

- 分解后的G矩阵的每一列对F中的向量进行编码来表达X中的一个样本
- 如果 G^T 中的每一列且仅有一个非0元
- 则X中每个元素被F中每一列近似表出
- 同时自然得出 G^T 的每一行之间是正交的
- G^T 同一行中所有非零元对应的 X_i 被影射到同一个点（同一类）



贝叶斯分析基础

陈桐飞、胡俊峰



贝叶斯统计与Naïve Bayes classifier

- Bayes 模型基础
 - Naïve Bayes 分类器
- 

概率论相关知识

- 随机变量及其分布:
 - 二项分布 $B(n, p)$
 - 正态分布 $N(\mu, \sigma^2)$
- 期望, 方差

二项分布

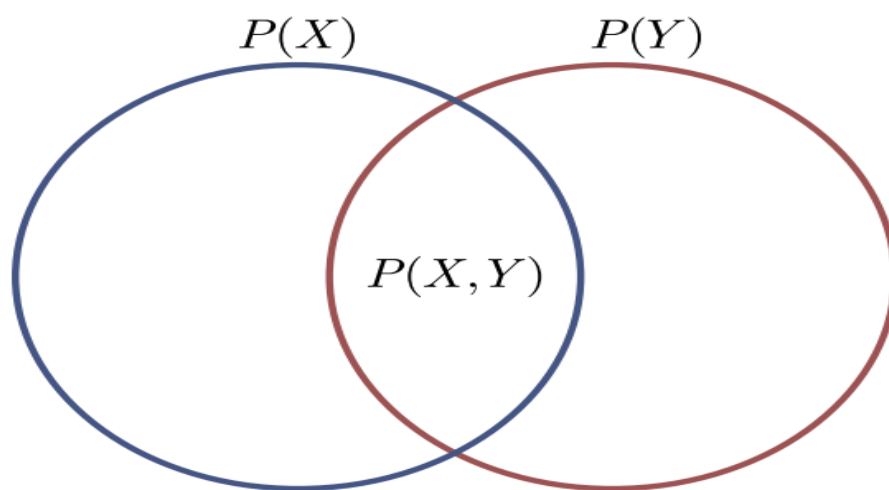
- 只有两种可能结果的随机试验，当成功的概率（ π ）是恒定的二项分布公式且各次试验相互独立。如果进行 n 次试验，取得成功次数为 k （ $k=0,1,\dots, n$ ）的概率可用下面的二项分布概率公式来描述：

- $P=C(k,n)*\pi^k*(1-\pi)^{n-k}$

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}=b(k;n,p)$$
$$(k=0,1,\dots,n),$$

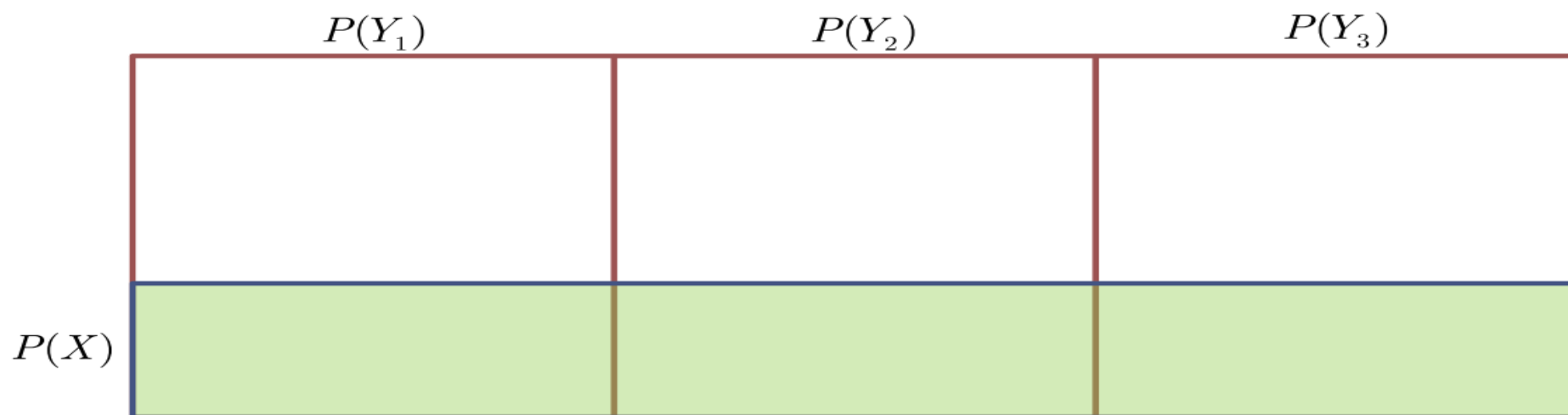
联合概率与条件概率

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



全概率公式


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



贝叶斯公式

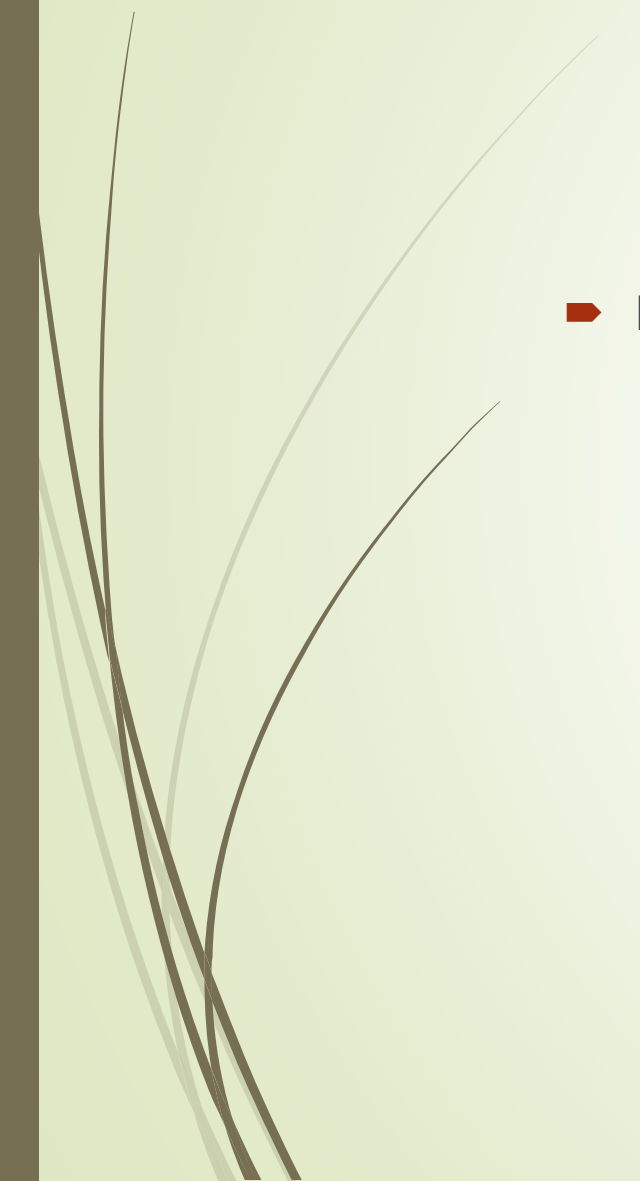

$$\begin{aligned}\therefore P(X, Y) &= P(X | Y)P(Y) \\ \therefore P(X, Y) &= P(Y | X)P(X)\end{aligned}$$

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



举个例子：

- 小A如果精神好，80%可能会起来跑步。
- 如果精神不好，40%可能会起来跑步
- 总体观察小A精神好的概率为60%
- 看到小A在跑步，问：精神好的概率？


$$\Rightarrow P(G | R) = \frac{P(R | G) * P(G)}{P(R)}$$

抛硬币

- 硬币若公平, 则其出现正面的概率为 0.5.
- 拿到一枚硬币, 如何知道它是否公平?
 - 做实验!
- 最大似然估计方法:
 - 抛硬币 n 次, 得到 m 次正面.
 - 故其出现正面的概率的最大似然估计为 m/n .

参数估计

- 抛硬币 4 次得到 4 次正面.
 - 你相信正面的概率为 1 吗?
- 魔术师拿出一枚道具硬币. 硬币抛 4 次得到 2 次正面.
 - 你相信正面的概率为 0.5 吗?
- 实验开始前, 我们对问题已有自己的认识.
 - 贝叶斯统计学承认这种认识的存在, 并将其用“先验”表述.

重新审视概率的定义

- 频率 (frequency):
 - 多次实验后频率的极限
- 信念 (belief):
 - 主观对事件发生与否的信念
- 频率学派统计学 (Frequentist statistics) 使用第一种定义.
- 贝叶斯学派统计学 (Bayesian statistics) 使用第二种定义.

重新审视贝叶斯公式

后验 (posterior) 似然 (likelihood) 先验 (prior)

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

$$\text{后验} \propto \text{先验} \times \text{似然}$$

参数的分布化

- 有时最大后验估计不能满足我们的要求
- 我们更希望先验和后验都以分布的形式给出
- 将参数也视为随机变量

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

- 我们需要通过先验分布 $P(Y)$ 计算后验分布 $P(Y|X)$.

后验的分布 (续)

$$P(Y | X) = \frac{P(X | Y)P(Y)}{\int_y P(X | y)P(y) \mathrm{d} y}$$

- 例: 连续投掷骰子 n 次, 得到 m 次正面. 显然, 每次骰子的状态服从二项分布 $x \sim B(1, p)$. 已知参数 p 的先验分布为 F , 求 p 的后验分布.

- 解: 显然, p 的取值范围为 $[0, 1]$. 代入贝叶斯公式:

$$\begin{aligned} P(p | X) &= \frac{P(X | p)P(p)}{\int_0^1 P(X | t)P(t)\mathrm{d} t} \\ &= \frac{p^m(1-p)^{n-m} F(p)}{\int_0^1 t^m(1-t)^{n-m} F(t)\mathrm{d} t} \end{aligned}$$

- 我们遇到了一些困难.

后验的分布

$$P(p | X) = \frac{p^m (1 - p)^{n-m} F(p)}{\int_0^1 t^m (1 - t)^{n-m} F(t) dt}$$

- 这个式子是否能容易地解出, 依赖于先验分布 F 的选择.
 - 当然, 使用数值计算的方法去近似积分的值也是可行的.
- 我们希望选择满足下列条件的先验分布 F :
 - (1) 无需使用数值计算的方法去近似所求的结果;
 - (2) 我们希望后验和先验是同种分布, 方便我们的后续计算.
- 我们称这样的先验分布为似然函数的**共轭先验(conjugate prior)**.
- 存在这样的先验吗?

共轭先验

$$P(p | X) = \frac{p^m (1 - p)^{n-m} F(p)}{\int_0^1 t^m (1 - t)^{n-m} F(t) dt}$$

- 答案是肯定的. 对于刚才的例子:
- 选择 $F(p) \propto p^{\alpha-1} (1 - p)^{\beta-1}$. 这个分布叫做 Beta 分布.
- 代入上式:
$$P(p | X) \propto \frac{p^m (1 - p)^{n-m} p^{\alpha-1} (1 - p)^{\beta-1}}{\int_0^1 t^m (1 - t)^{n-m} t^{\alpha-1} (1 - t)^{\beta-1} dt}$$
$$\propto p^{m+\alpha-1} (1 - p)^{n-m+\beta-1}$$
- p 的后验仍然为 Beta 分布.

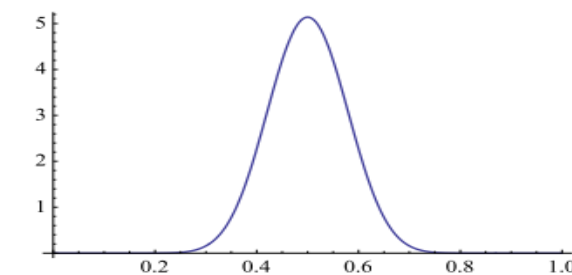
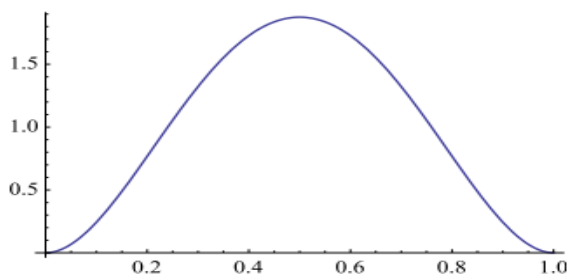
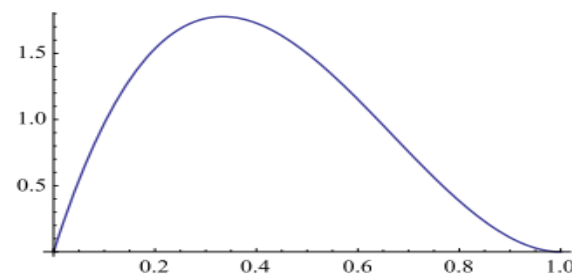
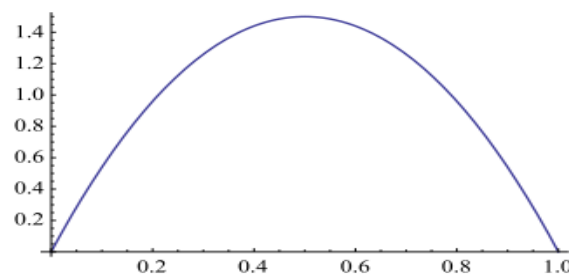
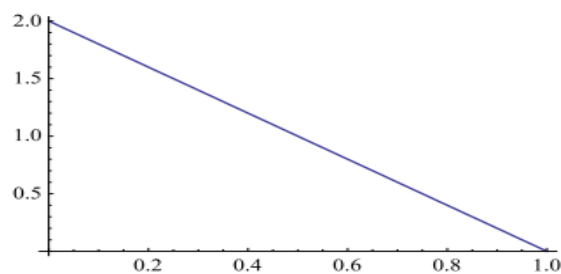
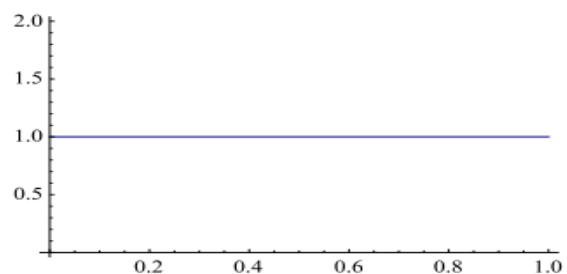
共轭先验

$$\begin{aligned} p &\sim \text{Beta}(\alpha, \beta) \\ m &\sim B(n, p) \\ p_{|m} &\sim \text{Beta}(m + \alpha, n - m + \beta) \end{aligned}$$

- Beta 分布为伯努利分布及二项分布的共轭先验分布.
- 对于其他分布, 我们也能找到其共轭先验, 这里不再推导.
- **问题: 为了计算方便而选择先验真的没问题吗?**
 - Beta 分布的表达能力
 - 参数的物理意义: 伪计数

继续硬币的例子

- 先验 $\text{Beta}(1, 1)$, 观测 { 反正反正 ... 反正 } (40 次观测).



抛硬币问题总结

- 抛硬币 n 次, 得到 m 次正面. 求得到正面的概率 p .
- 频率学派统计学的解答:
 - 最大似然估计值 $\hat{p} = \frac{m}{n}$. 该估计值的方差可用大数定律计算.
- 贝叶斯学派统计学的解答:
 - 在先验分布 $p \sim \text{Beta}(\alpha, \beta)$ 的条件下, 后验分布为
$$p \sim \text{Beta}(\alpha + m, \beta + n - m).$$
 - 后验期望估计值 $\hat{p} = \frac{m + \alpha}{n + \alpha + \beta}$.

贝叶斯方法总结

后验
(posterior)

似然
(likelihood)

先验
(prior)

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

The diagram shows the Bayes' theorem formula enclosed in a blue box. Three labels with arrows point to specific parts of the formula: '后验 (posterior)' in red points to $P(Y | X)$; '似然 (likelihood)' in green points to $P(X | Y)$; and '先验 (prior)' in blue points to $P(Y)$.

- 模型参数也视为随机变量(而非传统的一个值)
- 为参数设置先验分布
 - 先验的参数称为**超参数(hyperparameter)**
- 求模型参数的后验分布

机器学习

- 从已知数据中分析得到规律, 并对未知数据进行预测.
- 预测函数 $y = F(\mathbf{x})$.
- 通过已知数据学习函数 F .
- 线性回归就是一种机器学习:
 - 预测函数 $F(x) = ax + b$.
 - 通过数据集 $\{(x, y)\}$ 学习 F , 也即学习参数 a, b .

分类器

- 根据样本, 得到它的分类.
- 每个样本 \mathbf{x} 为一个多维向量. 每一维都称为“特征”.
- 预测函数 $F(\mathbf{x})$. k 分类器值域为类型集 $\{1, \dots, k\}$.
- 分类器: $\mathbb{R}^d \rightarrow \{1, \dots, k\}$.
- 对比回归器: $\mathbb{R}^d \rightarrow \mathbb{R}$.

朴素贝叶斯分类器 (Naïve Bayes)

- 文本分类.
- 统计文本中每个互异的词的频数作为这个文本的向量.
 - 词袋模型 (bag-of-words model)

• *To be, or not to be, that is the question.*

(2	2	1	1	1	1	1	1)
	to	be	or	not	that	is	the	question	

朴素贝叶斯分类器

- 基本思想: 最大后验估计.

$$\begin{aligned} F(\mathbf{x}) &= \arg \max_y P(y \mid \mathbf{x}) \\ &= \arg \max_y \frac{P(\mathbf{x} \mid y)P(y)}{P(\mathbf{x})} \\ &= \arg \max_y P(\mathbf{x} \mid y)P(y) \\ &= \arg \max_y P(y) \prod_{i=1}^d P(x_i \mid y) \end{aligned}$$

- 朴素 (naïve): 做了特征之间互相独立的假设

参数估计

$$F(\mathbf{x}) = \arg \max_y P(y) \prod_{i=1}^d P(x_i | y)$$

- 模型所需的参数有 $P(y)$, $P(x_i | y)$.
- 最大似然估计:

$$\hat{P}(y) = \frac{|N(y)|}{Total}$$

$$\hat{P}(x_i | y) = \frac{n_{x_i, y}}{n_y}$$

- 问题?
 - 概率为 0 的情况. 若类 1 中出现词 x , 类 2 中没有.
 - 则 $P(x|2) = 0$. 一个含有 x 的词永远无法被分入类 2.
 - 这是我们不希望看到的.

平滑 (smoothing)

- 拉普拉斯 (+1) 平滑:

$$\hat{P}(y) = \frac{|N(y)|}{Total}$$

$$\hat{P}(x_i | y) = \frac{n_{x_i, y} + 1}{n_y + |V|}$$

- 带系数:

$$\hat{P}(x_i | y) = \frac{n_{x_i, y} + \alpha}{n_y + \alpha |V|}$$

运用朴素贝叶斯分类器

- 体育/政治/科技/... 20 类新闻语料分类, 训练集为 18000 篇文档, 测试集 2000 篇.
- 文档预处理: 依据停用词表去掉停用词.
 - a, the, is, on, 等等
- 使用朴素贝叶斯分类器进行训练/分类.
 - 84.56% 准确率.

朴素贝叶斯分类器总结

- 多项分布朴素贝叶斯分类器
- 复杂度: 分类 $O(k)$
- 基本思想: 最大后验估计
- 平滑问题

分类与回归

- 对于一个样本集，如果能找到一个合理的分类函数，使得：
 $F(X) \approx 1$ (当 $Y = 1$); $F(X) \approx 0$ (当 $Y = 0$)
- 则可以称 我们找到了一个原样本集的一个‘似然’函数。
- 如果 F 是以最大概率符合样本数据，则 F 称为最大似然函数。



本次作业稍晚一些会发布

