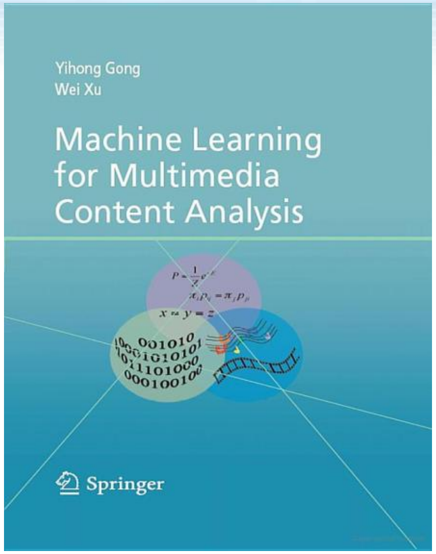
计算机视觉的
统计方法与机器学习

第一课

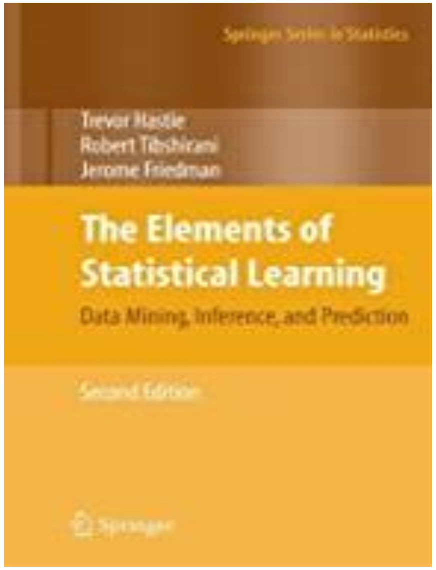
龚怡宏

西安交通大学

本课程教科书



Yihong Gong and Wei Xu, “Machine Learning for Multimedia Content Analysis”, Springer Publishers, 2008.



Trevor Hastie, Robert Tibshirani, Jerome Friedman, “The Elements of Statistical Learning: Data Mining, Inference, and Prediction”, Second Edition, Springer Publishers, 2017.

课程特点

- 强调方法的全局观，不追求严密的数学证明。
- 聚焦主流、实用的机器学习算法，不追求大而全。
- 培养既掌握系统性理论知识，又具有实战能力的全面型人工智能专业人才。

课程评分方式

A计划（疫情在3月底前解除）

- 课程没有期末考试，以大作业做为课程考察。
- 大作业需参加中国智慧城市技术挑战赛中的视频内容解析比赛（教育部、西安交大指定全国A类技术大赛）。

B计划（疫情解除时间延后，挑战赛无法按期举办）

- 平时成绩：包括线上教学中穿插的即时答题成绩+课后作业成绩，占比40%。
- 期末考试成绩：占比60%。

中国智慧城市技术挑战赛

- 本团队5年间的参赛成果：特等奖1项，一等奖7项，二等奖6项，三等奖4项，共获奖18项，61人次获奖。
- 打分规则：
  - 进入决赛就得“优”。
  - 没进入决赛但在最终排名中成绩高于倒数10%的得“良”。
  - 没进入决赛，同时在最终排名中成绩是倒数10%的得“中”。
  - 没有提交比赛结果的得“不及格”。

课程提纲

1、机器学习算法综述

2、数据聚类1（k-means，谱聚类基础）

3、数据聚类2（单线性NMF模型 ）

4、线性分类器与支持向量机（SVM）

5、Boosting

6、深度卷积神经网络

7、马尔科夫链

8、马尔科夫随机场和吉布斯采样（时间足够时）

9、HMM与EM算法

第一课内容

一、机器学习算法的应用

二、什么是机器学习算法

三、机器学习算法的分类

四、机器学习算法的重要构成要素

五、概率论基本知识回顾

一、机器学习算法的应用

- 机器学习算法在哪些方面已经接近或超越了人类？
  - 几乎所有棋牌类游戏
  - 人脸识别
  - 语音识别
  - 机器翻译
  - 知识问答
  - 无人驾驶
  - 复杂病情诊断

美国加州无人驾驶汽车路测统计

California Autonomous Testing Disengagements						
Company	**Total Miles Driven (2014.09-2017.12)**	**Total DE* (2014.09-2017.12)**	**Miles per DE**	**Miles per DE in 2015**	**Miles per DE in 2016**	**Miles per DE in 2017**
Waymo (aka Google)	1,412,743.60	528	2675.65	1244.37	5127.97	5595.95
Mercedes-Benz	3,500.20	2209	1.58	1.69	2	1.29
Delphi	21597.9	664	32.53	41.14	17.56	22.35
Bosch	3372.1	2665	1.27	1.5	0.68	2.43
Nissan	10591.4	158	67.03	14.01	146.39	208.63
Cruise(GM)	141691.05	389	364.24	2.32	54.01	1254.91
VW/Audi	5,531	85	65.07	65.07	-	-
Tesla Motors	550	182	3.02	-	3.02	-
Baidu USA	395.33	26	15.21	-	4.52	17.75
Drive.ai	6572.4	151	43.53	-	9.44	65.38
Telenav	1697	58	29.26	-	40.5	27.97
Valeo	574.1	215	2.67	-	7.23	2.61
BMW	638	1	638	-	638	-
Ford	590	3	196.67	-	196.67	-
Zoox	2244.6	14	160.33	-	-	160.33
NVIDIA	505	109	4.63	-	-	4.63

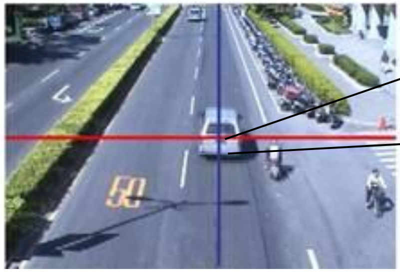
1. *DE = Disengagements
2. 统计起止时间： 2015年 (2014.09 - 2015.11) 2016年 (2015.12 - 2016.11) 2017年 (2016.12 - 2017.11)

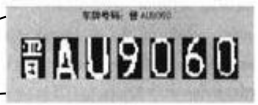
复杂病情诊断

- IBM押注“沃森智慧医疗系统”，希望人工智能技术诊断和治疗疑难疾病的能力超越人类最好医生，但却以失败而告终。
- 挑战性在于：难以搜集大量患者的完整病例。
- 人体是一个极为复杂系统，同样症状不一定对应同样疾病，同样疾病不一定拥有同样症状。
- 同样疾病的最佳治疗方案也因人而异。

一、机器学习算法的应用

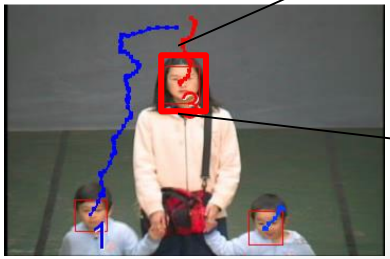
- 机器学习算法的其它成功应用
  - 车辆检测、跟踪、车牌识别
  - 行人检测、跟踪、人脸识别
  - 大规模群体事件自动监控
  - 森林火灾自动监测
  - ……
- 总之，凡是需要高智能的地方都需要机器学习算法去实现

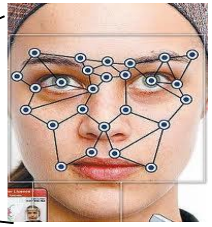




车牌识别

车辆检测、跟踪





人脸识别

行人检测、跟踪

森林火灾自动检测

大规模群体事件自动监控

二、什么是机器学习算法

- **Machine learning**, a branch of artificial intelligence, is about the construction and study of systems that can learn from data (Wikipedia).
- 机器学习算法的一般应用框架
  - 定义需要实现的功能。
  - 采集足够多的正例与负例样本: $T=\{\mathbf{x}_i, y_i\}_1^N$
  - 利用训练样本$T=\{\mathbf{x}_i, y_i\}_1^N$通过迭代训练，得到模型 $y=f(\mathbf{x}, \Theta)$。
  - 如果标签$y_i \in \{-1, +1\}$是离散的，这就是一个分类问题；如果y_i是连续的，这就是一个回归问题。

二、什么是机器学习算法

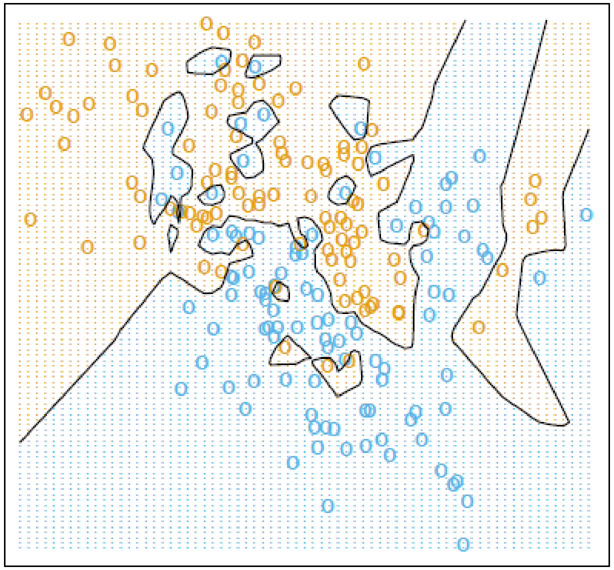
- 模型是什么？
  - 模型是用来描述某个特定现象或事务的。
    - ✓ 牛顿万有引力定律是描述宇宙中所有宏观物体运动规律的模型。
    - ✓ 薛定谔方程是描述微观粒子运动规律的模型。
- 模型的种类
  - 归纳模型（Inductive inference）：由一个数学公式构成，公式中的每个变量都具有明确物理意义，能够真正描述目标系统的规律。
  - 预测模型（Predictive inference）：往往是由一个万能函数构成，由许多参数组成，每个参数一般不具备任何物理意义。一般只能模拟或预测目标系统的输出。
  - 直推模型（Transductive inference）：没有明确的模型或函数，但可计算出模型在特定点的值。

二、什么是机器学习算法

- 直推模型（Transductive inference）。
  - 每个数据都是对目标世界的取样。
  - 当对所在世界的取样足够全面与密集时，我们就获得了对这个世界的完整描述。
  - 当一个未知数据到来时，我们只要找到与它最相似的样本就能知道这个数据的属性与含义。

Ask not “what is this?”, ask “what is this like?”

– Moshe Bar, 哈佛大学神经学家



样本足够多时，简单的k-NN就能对未知数据做精准分类

三种模型的总结与比较

	Inductive inference（归纳模型）	Predictive inference（预期模型）	Transductive inference（直推模型）
目标	发现事物的真正规律	发现预测规则	评估未知预测函数在某些点的值
复杂度	比较困难	相对容易	最容易
适用性	少数变量就能描述的简单世界	需多个变量描述的复杂世界	需多个变量描述的复杂世界
计算成本	低	高	最高
泛化能力	低	高	最高

三、机器学习算法的分类

- 非监督学习与监督学习
  - 非监督学习：不需要训练样本的机器学习算法，如数据聚类算法。
  - 监督学习：需要训练样本$T=\{\mathbf{x}_i, y_i\}_1^N$的机器学习算法，如大多数分类、回归算法。
- 生成模型与判别模型（generative vs. discriminative）
  - 生成模型计算数据x与标签y的联合概率$P(x,y)$，用下列公式计算分类概率：$P(y|x) = P(x,y)/P(x)$
  - 判别模型直接计算分类概率$P(y|x)$
- 简单数据模型与复杂数据模型
  - 简单数据模型：被用来处理相互独立的简单数据
  - 复杂数据模型：被用来处理具有时空关联性的复杂数据

四、机器学习算法的三个重要方面

- **Structural model:** 我们选择哪一类函数 $f(\mathrm{x}, \Theta)$来建立模型？
- **Error model:** 我们选择哪一类损失函数（loss function）$L(y, f(\mathrm{x}, \Theta))$ 来做训练？损失函数相当于为模型的选择制定考核标准。
- **Optimization procedure:** 我们选择哪一种数值计算方法来获取最优模型 $f^*(\mathrm{x}, \Theta)$？

针对简单数据的常用 Structural Models

- Gaussian or Gaussian Mixture Model (GMM):

$$\mathrm{f}(\mathbf{x},\Theta)=\sum_i w_i \mathrm{N}(\mathbf{x};\mu_i,\Sigma_i),\ \text{where } \Theta=\{w_i,\mu_i,\Sigma_i\}_{i=1}^N$$

- Linear Model:

$$f(\mathbf{x},\Theta)=\mathbf{w}\cdot\mathbf{x}+b,\ \text{where } \Theta=\{\mathbf{w},b\}$$

SVM 是线性模型与hinge损失函数的组合。

$$L(y,f(\mathbf{x};\Theta))=\max\left(0,(1-yf(\mathbf{x};\Theta)\right)$$

- Logistic Regression Model:

$$\ln\left(\frac{p}{1-p}\right)=\beta_0+\boldsymbol{\beta}^T\cdot\mathbf{x}$$

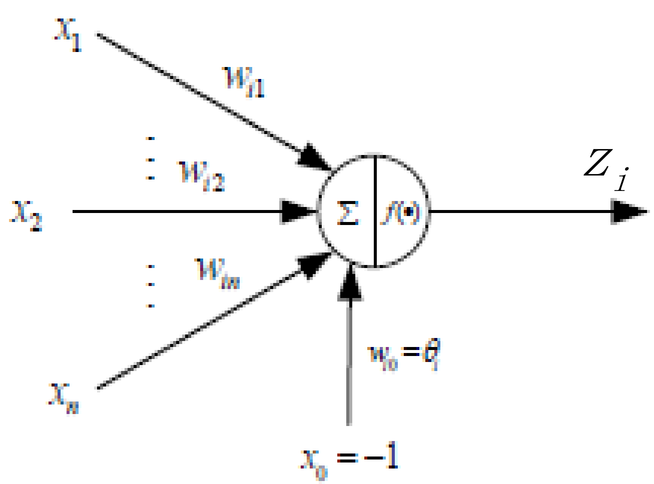
Continue

- Additive Model:

$$F_M(\mathbf{x},\Theta)=\sum_{m=1}^{M}\beta_m b_m(\mathbf{x};\gamma_m),\ \text{where}\ \ \Theta=\{\beta_m,\gamma_m\}_{m=1}^{M}$$

  - 当 $b(\mathrm{x};\gamma_m)$ 是sigmoid函数时（例如tanh($\gamma_m\cdot\mathrm{x}$)），$F_M(\mathrm{x},\Theta)$ 代表单隐蔽层神经网络。

$$z_i=\tanh(\sum_{j=1}^{n}w_{ij}x_j+b_i),\qquad y_k=\sum_l \beta_{kl}z_l=\sum_l \beta_{kl}\tanh(\sum_j w_{lj}x_j+b_l)$$



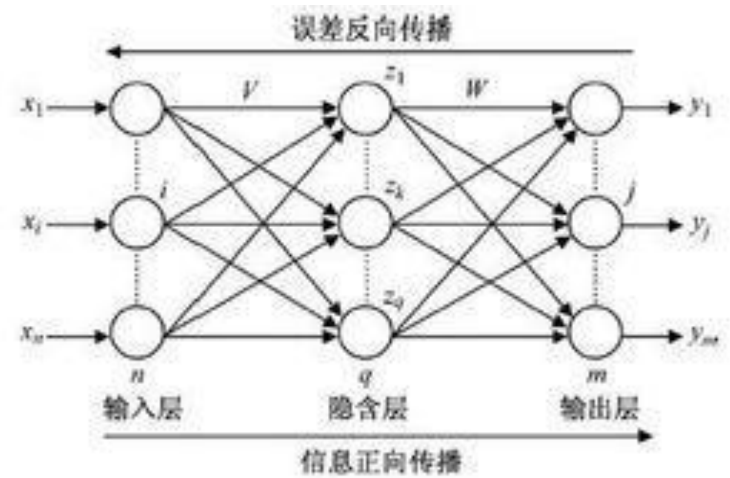


图 1 3层 BP神经网络结构图

Continue

- Additive Model:

$$F_M(\mathbf{x},\Theta)=\sum_{m=1}^{M}\beta_m b_m(\mathbf{x};\gamma_m),\ \text{where}\ \ \Theta=\{\beta_m,\gamma_m\}_{m=1}^{M}$$

  - 当 $b(\mathrm{x};\gamma_m)$ 是sigmoid函数时（例如tanh($\gamma_m\cdot$x)）， $F_M(\mathrm{x},\Theta)$ 代表单隐蔽层（single-hidden layer）神经网络。
  - 当 $b(\mathrm{x};\gamma_m)$ 是高斯函数时，$F_M(\mathrm{x},\Theta)$ 代表 GMM。
  - 当 $b(\mathrm{x};\gamma_m)=\psi_{a,b}(\mathrm{x})$ 是小波函数时，$F_M(\mathrm{x},\Theta)$ 代表小波变换。

误差模型（Error Models）

- 平方差误差（Squared error）：

$$L(y, f(\mathbf{x};\Theta)) = (y - f(\mathbf{x};\Theta))^2$$

这种损失函数不适用于分类任务。

- 绝对误差（Absolute error）：

$$L(y, f(\mathbf{x};\Theta)) = \| y - f(\mathbf{x};\Theta) \|$$

这种损失函数不适用于分类任务。

- 指数误差（Exponential loss）：

$$L(y, f(\mathbf{x};\Theta)) = \exp(-yf(\mathbf{x};\Theta))$$

误差模型 (Error Models)

- Hinge loss:

$$L(y, f(\mathbf{x};\Theta)) = \max\left(0, (1 - yf(\mathbf{x};\Theta))\right)$$

- Log-likelihood:

$$L(y, p(\mathbf{x})) = \sum_{k=1}^{K} I(y=k)\log p(y=k \mid \mathbf{x})$$

For two class problems:

$$L(y, p(\mathbf{x})) = y\log p(\mathbf{x}) + (1-y)\log(1-p(\mathbf{x})) = -\log(1+e^{-yF(\mathbf{x})})$$

where $p(\mathbf{x}) = \dfrac{e^{F(\mathbf{x})}}{e^{F(\mathbf{x})} + e^{-F(\mathbf{x})}}$

误差模型（Error Models)

- Softmax loss：深度学习卷积神经网络常用损失函数，就是 归一化+Cross entropy

Cross Entropy的定义:

给定两个概率分布：$\mathbf{P} = [p_1, p_2, \ldots, p_n]$, $\mathbf{Q} = [q_1, q_2, \ldots, q_n]$,

$$H(\mathbf{P};\mathbf{Q}) = -\sum_{i=1}^{n} p_i \log q_i$$

给定神经网络输出层向量**x**：$\mathbf{x} = [x_1, \ldots, x_k, \ldots, x_n]$,

标签向量**y**：$\mathbf{y} = [0, \ldots, 1, \ldots 0]$, $y_k \neq 0$

归一化

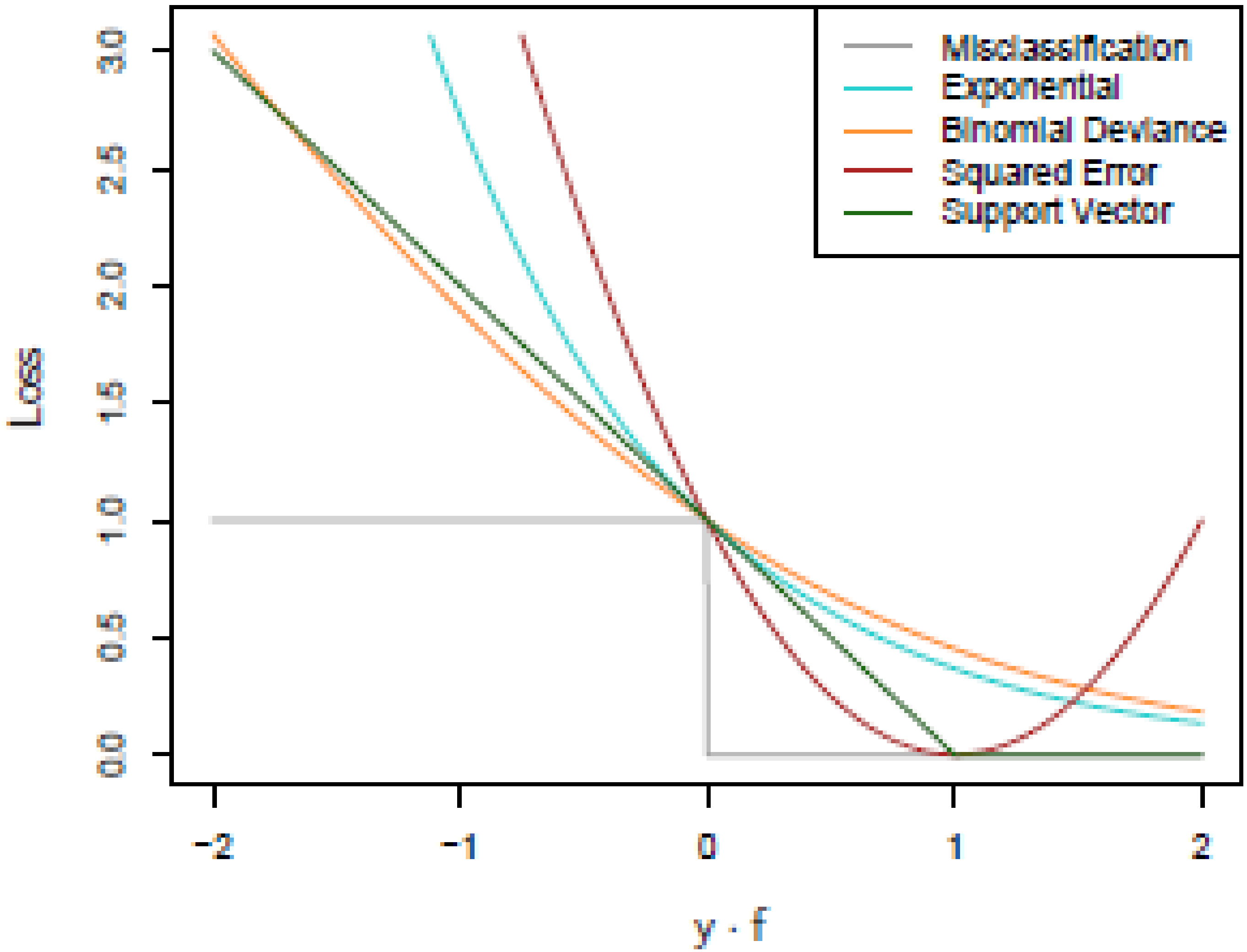
$$\hat{\mathbf{x}} = \left[\frac{e^{x_1}}{\sum_{i=1}^{n} e^{x_i}}, \cdots, \frac{e^{x_k}}{\sum_{i=1}^{n} e^{x_i}}, \cdots, \frac{e^{x_n}}{\sum_{i=1}^{n} e^{x_i}}\right]$$

Cross entropy

$$-1 * \log \frac{e^{x_k}}{\sum_{i=1}^{n} e^{x_i}} = -\log \frac{e^{x_k}}{\sum_{i=1}^{n} e^{x_i}}$$

softmax loss

损失函数曲线图



Misclassification
Exponential
Binomial Deviance
Squared Error
Support Vector
Loss
0.0
0.5
1.0
1.5
2.0
2.5
3.0
−2
−1
0
1
2
y · f

什么是好的损失函数?

- 平滑、易求导
- 必须是 $yF(\mathrm{x})$ 的单调函数
- 不应该对错误惩罚太大（鲁棒性）

Optimization Methods

- Steepest gradient decent
- Forward stagewise algorithm
- Newton-Raphson algorithm (needs to compute second-order derivative)
- E-M algorithm (Expectation, Maximization)

主观题 10分

设置

1、当前主流机器学习方法的本质是什么？

2、损失函数的作用是什么？

正常使用主观题需2.0以上版本雨课堂

作答

五、概率论基本知识回顾

概率定义(definition)

任何满足以下三个条件的函数 $\mathrm{P}(a)$统称为 **概率分布函数**：

公理1：对任意的 a，均有 $\mathrm{P}(a)\geqslant 0$。

公理2：$\mathrm{P}(\Omega)=1$，其中 Ω代表全集。

公理3：若事件 $a_1, a_2, \ldots$ 是不相交的，则

$$\mathrm{P}\left(\bigcup_{i=1}^{\infty} a_i\right)=\sum_{i=1}^{\infty}\mathrm{P}(a_i)$$

概率

■从上述定义，我们不难看出函数 P 的一些性质，比如：

$$\mathrm{P}(\varnothing)=0$$

$$a\subset b \;\Rightarrow\; \mathrm{P}(a)\leqslant \mathrm{P}(b)$$

$$0\leqslant \mathrm{P}(a)\leqslant 1$$

$$\mathrm{P}(a^{c})=1-\mathrm{P}(a)$$

$$a\cap b=\varnothing \;\Rightarrow\; \mathrm{P}(a\cup b)=\mathrm{P}(a)+\mathrm{P}(b)$$

■引理(lemma)

$$\mathrm{P}(a\bigcup b)=\mathrm{P}(a)+\mathrm{P}(b)-\mathrm{P}(a\bigcap b)$$

独立事件

■定义(definition)

若事件 a 和 b满足如下条件，则我们称其相互独立：

$$\mathrm{P}(a,b)=\mathrm{P}(a)\,\mathrm{P}(b)$$

同理，一系列事件$\{a_i : i \in I\}$相互独立，则需对 I 的任意有限子集J，均有：

$$\mathrm{P}\left(\bigcap_{i\in J} a_i\right)=\prod_{i\in J}\mathrm{P}(a_i)$$

条件概率

■定义(definition)

假设P(b)>0，我们将 b 已经发生的前提下 a 发生的概率称为**条件概率**：

$$\mathrm{P}(a \mid b) = \frac{\mathrm{P}(a,b)}{\mathrm{P}(b)}$$

■引理(lemma)

结合事件a与b相互独立的条件，进而可以得出：

$$\mathrm{P}(a \mid b) = \frac{\mathrm{P}(a,b)}{\mathrm{P}(b)} = \frac{\mathrm{P}(a)\,\mathrm{P}(b)}{\mathrm{P}(b)} = \mathrm{P}(a)$$

关于条件概率的几点总结

1. 若P(a)>0，则有

$$\mathrm{P}(a \mid b)=\frac{\mathrm{P}(a, b)}{\mathrm{P}(b)}$$

2. 对于固定的b，P(·|b)满足概率的三条公理，而对于固定的a，P(a|·)则并不满足公理。

3. 一般来说，$\mathrm{P}(a \mid b) \neq \mathrm{P}(b \mid a)$。

4. 当且仅当P(a|b)=P(a)时，a与b相互独立。

贝叶斯概率理论

■定理(Bayes' Theorem)

将全集划分为*k*个部分$a_1,\ldots,a_k$，且对于每个*i*都有$\mathrm{P}(a_i)>0$。若P(*b*)>0，则对每个*i*=1,…,*k*，均有

$$\mathrm{P}(a_i\,|\,b)=\frac{\mathrm{P}(b\,|\,a_i)\,\mathrm{P}(a_i)}{\sum_j \mathrm{P}(b\,|\,a_j)\,\mathrm{P}(a_j)}$$

End