概率论

机器学习最重要的概率知识：最大似然 和 最大后验，模式识别，本质上就是找出众多事件的关联关系

第一部分：回顾概率核心理论

随机事件的概念，概率的定义与计算方法;常用的概率分布，联合概率、边缘概率、条件概率

1.事件的关系与运算

(1) 子事件： $A \subset B$ ，若 A 发生，则 B 发生。

(2) 相等事件： $A = B$ ，即 $A \subset B$ ，且 $B \subset A$ 。

(3) 和事件： $A \bigcup B$ （或 $A + B$ ）， A 与 B 中至少有一个发生。

(4) 差事件： $A - B$ ， A 发生但 B 不发生。

(5) 积事件： $A \bigcap B$ （或 AB ）， A 与 B 同时发生。

(6) 互斥事件（互不相容）： $A \bigcap B = \varnothing$ 。

(7) 互逆事件（对立事件）：
$A \bigcap B = \varnothing, A \bigcup B = \Omega, A = \bar{B}, B = \bar{A}$

2.运算律

(1) 交换律：$A\bigcup B=B\bigcup A, A\bigcap B=B\bigcap A$

(2) 结合律：$(A\bigcup B)\bigcup C=A\bigcup(B\bigcup C)$ ；

$(A\bigcap B)\bigcap C=A\bigcap(B\bigcap C)$

(3) 分配律：$(A\bigcup B)\bigcap C=(A\bigcap C)\bigcup(B\bigcap C)$

3.德·摩根律

$$\overline{A\bigcup B}=\bar{A}\bigcap\bar{B}\quad\overline{A\bigcap B}=\bar{A}\bigcup\bar{B}$$

4.完全事件组

$A_1A_2\cdots A_n$ 两两互斥，且和事件为必然事件，即

$$A_i\bigcap A_j=\varnothing, i\neq j, \bigcup_{i=1}^{n}=\Omega$$

5.概率的基本公式

(1)条件概率:

$P(B|A)=\dfrac{P(AB)}{P(A)}$,表示 A 发生的条件下， B 发生的概率。

(2)全概率公式:

$$P(A)=\sum_{i=1}^{n}P(A|B_i)P(B_i),B_iB_j=\varnothing,i\neq j,\bigcup_{i=1}^{n}B_i=\Omega$$

(3) **Bayes**公式:

$$P(B_j|A)=\frac{P(A|B_j)P(B_j)}{\sum\limits_{i=1}^{n}P(A|B_i)P(B_i)},j=1,2,\cdots,n$$

注：上述公式中事件 B_i 的个数可为可列个。

(4)乘法公式： $P(A_1A_2)=P(A_1)P(A_2|A_1)=P(A_2)P(A_1|A_2)$

$$P(A_1A_2\cdots A_n)=P(A_1)P(A_2|A_1)P(A_3|A_1A_2)\cdots P(A_n|A_1A_2\cdots A_{n-1})$$

6.事件的独立性

(1) A 与 B 相互独立 $\Leftrightarrow P(AB)=P(A)P(B)$

(2) A ， B ， C 两两独立

$\Leftrightarrow P(AB)=P(A)P(B)$ ； $P(BC)=P(B)P(C)$ ； $P(AC)=P(A)P(C)$ ；

(3) A ， B ， C 相互独立

$\Leftrightarrow P(AB)=P(A)P(B)$ ； $P(BC)=P(B)P(C)$ ；

$P(AC)=P(A)P(C)$ ； $P(ABC)=P(A)P(B)P(C)$

7.独立重复试验

将某试验独立重复 n 次，若每次实验中事件 A 发生的概率为 p ，则 n 次试验中 A 发生 k 次的概率为:

$P(X=k)=C_n^kp^k(1-p)^{n-k}$

第二部分：核心的几种随机变量的分布以及变量之间的关系

离散型随机变量的几种主要的分布，而连续型随机变量主要就是掌握正态分布即可

分布的期望、方差等数字特征，协方差以及相关性的意义和计算方法

5.常见分布

(1) 0-1分布: $P(X=k)=p^k(1-p)^{1-k}, k=0,1$

(2) 二项分布: $B(n,p)$ ： $P(X=k)=C_n^k p^k(1-p)^{n-k}, k=0,1,\cdots,n$

(3) **Poisson**分布: $p(\lambda)$ ： $P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}, \lambda>0, k=0,1,2\cdots$

(4) 均匀分布 $U(a,b)$ ： $f(x)=\begin{cases}\frac{1}{b-a}, a<x<b\\ 0,\end{cases}$

(5) 正态分布: $N(\mu,\sigma^2)$ ： $\varphi(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \sigma>0, \infty<x<+\infty$

(6)指数分布: $E(\lambda): f(x)=\begin{cases}\lambda e^{-\lambda x}, x>0, \lambda>0\\ 0,\end{cases}$

(7)几何分布: $G(p): P(X=k)=(1-p)^{k-1}p, 0<p<1, k=1,2,\cdots.$

(8)超几何分布:

$H(N,M,n): P(X=k)=\frac{C_M^k C_{N-M}^{n-k}}{C_N^n}, k=0,1,\cdots,min(n,M)$

第三部分：参数估计理论

最小偏差无偏估计、最大似然估计和贝叶斯估计，EM 算法

第四部分：假设检验

第五部分：建立随机理论的相关概念

蒙特卡罗方法的基本思想

第六部分：信息论

关于熵的一些理论，联合熵、条件熵、相对熵、互信息等概念，以及最大熵模型

第七部分：随机过程初步理论和应用

马尔科夫链是必须学习的，了解状态转移矩阵、多步转移、几种不同的状态分类、平稳分布等

第八部分：时间序列分析

pandas 工具

数理统计

数理统计是关于抽样、统计和假设检验的科学。以数据为基础，利用数学方程式来探究变量变化规律的一套规范化流程。

总结来说，我们可以认为机器学习和统计建模是预测建模领域的两个不同分支。这两者之间的差距在过去的 10 年中正在不断缩小，而且它们之间存在许多相互学习和借鉴的地方。未来，它们之间的联系将会更加紧密。

1.基本概念

总体：研究对象的全体，它是一个随机变量，用 X 表示。

个体：组成总体的每个基本元素。

简单随机样本：来自总体 X 的 n 个相互独立且与总体同分布的随机变量 $X_1, X_2 \cdots, X_n$ ，称为容量为 n 的简单随机样本，简称样本。

统计量：设 $X_1, X_2 \cdots, X_n$，是来自总体 X 的一个样本， $g(X_1, X_2 \cdots, X_n)$ ）是样本的连续函数，且 $g()$ 中不含任何未知参数，则称 $g(X_1, X_2 \cdots, X_n)$ 为统计量。

样本均值： $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

样本方差： $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$

样本矩：样本 k 阶原点矩： $A_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k, k = 1, 2, \cdots$

样本 k 阶中心矩： $B_k = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^k, k = 1, 2, \cdots$

2.分布

χ^2 分布： $\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2 \sim \chi^2(n)$ ，其中 $X_1, X_2 \cdots, X_n$，相互独立，且同服从 $N(0,1)$

t 分布： $T = \dfrac{X}{\sqrt{Y/n}} \sim t(n)$ ，其中 $X \sim N(0,1), Y \sim \chi^2(n)$，且 X ， Y 相互独立。

F 分布： $F = \dfrac{X/n_1}{Y/n_2} \sim F(n_1, n_2)$ ，其中 $X \sim \chi^2(n_1), Y \sim \chi^2(n_2)$，且 X ， Y 相互独立。

分位数：若 $P(X \le x_\alpha) = \alpha$，则称 x_α 为 X 的 α 分位数

3.正态总体的常用样本分布

(1) 设 $X_1, X_2 \cdots, X_n$ 为来自正态总体 $N(\mu, \sigma^2)$ 的样本，

$\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i, S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$，则：

1) $\overline{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$ 或者 $\dfrac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$

2) $\dfrac{(n-1)S^2}{\sigma^2} = \dfrac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \sim \chi^2(n-1)$

3) $\dfrac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 \sim \chi^2(n)$

4) $\dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$

4.重要公式与结论

(1) 对于 $\chi^2 \sim \chi^2(n)$ ，有 $E(\chi^2(n)) = n, D(\chi^2(n)) = 2n;$

(2) 对于 $T \sim t(n)$ ，有 $E(T) = 0, D(T) = \frac{n}{n-2}(n > 2)$ ；

(3) 对于 F~$F(m,n)$ ，有 $\frac{1}{F} \sim F(n,m), F_{a/2}(m,n) = \frac{1}{F_{1-a/2}(n,m)};$

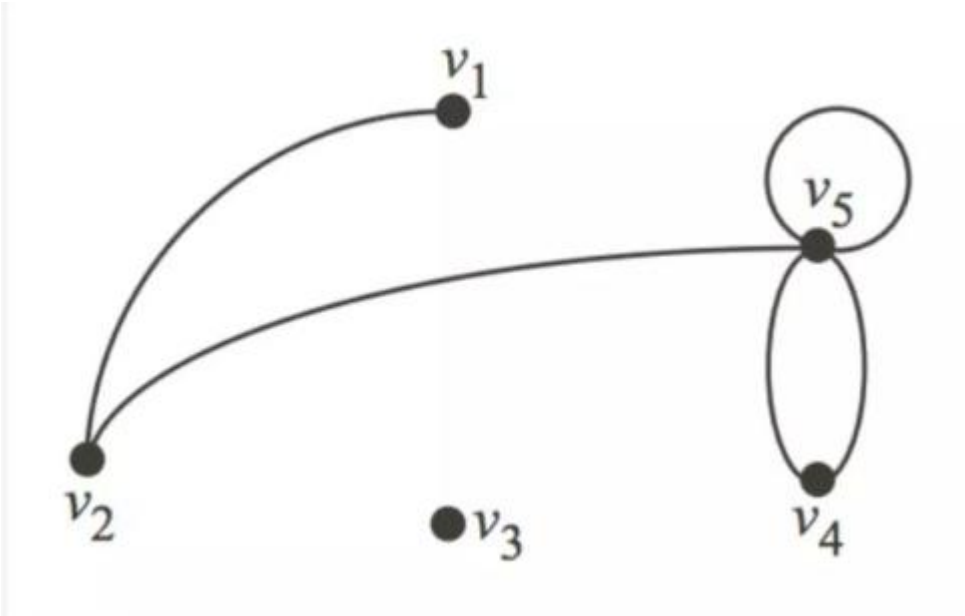
(4) 对于任意总体 X ，有 $E(\overline{X}) = E(X), E(S^2) = D(X), D(\overline{X}) = \frac{D(X)}{n}$

图论

图提供了一种处理关系和交互等抽象概念的更好的方法。它还提供了直观的视觉方式来思考这些概念。图很自然地成了分析社会关系的基础。

图及其应用

让我们看一个简单的图（Graph）来理解这个概念。如下图所示：



假设此图代表某个城市的热门景点位置，以及游客所遵循的路径。我们把 V 视为景点位置，将 E 视为从一个地方到另一个地方的路径。

具体而言，**图（Graph）是用于研究对象和实体之间成对关系的数学结构。**它是离散数学的一个分支，在计算机科学，化学，语言学，运筹学，社会学等领域有多种应用。

数据科学和分析领域也使用图来模拟各种结构和问题。作为一名数据科学家，你应该能以有效的方式解决问题，如果数据是以特定方式排列的，则图可以提供一种解决问题的机制。

形式上看，

- 图是一对集合。G = (V, E)，V 是顶点集合，E 是边集合。 E 由 V 中的元素对组成（无序对）
- 有向图（DiGraph）也是一对集合。D = (V, A)，V 是顶点集合，A 是弧集合。A 由 V 中的元素对组成（有序对）

在有向图的情况下，（u，v）和（v，u）之间存在区别。通常在这种情况下，边被称为弧，以指示方向的概念。

图的应用范围

- **营销分析**

图可用于找出社交网络中最有影响力的人。广告商和营销人员可以通过社交网络中最有影响力的人员传达他们的信息，从而估算最大的营销价格。

- **银行交易**

图可用于查找有助于减少欺诈交易的异常模式。有一些例子可以通过分析银行网络的资金流动来侦测恐怖主义活动。

- **供应链**

图有助于确定送货卡车的最佳路线以及识别仓库和交付中心的位置。

- **制药公司**

制药公司可以使用图论优化销售人员的路线。这有助于降低成本并缩短销售人员的行程时间。

- **电信行业**

电信公司通常使用图（Voronoi图）来了解基站的数量和位置，以确保最大的覆盖范围。

图的历史

1840 年，A.F Mobius 提出了完全图（complete graph）和二分图（bipartite graph）的概念，Kuratowski 通过趣味谜题证明它们是平面的。树的概念（没有环的连通图）由 Gustav Kirchhoff 于 1845 年提出，他在计算电网或电路中的电流时使用了图论思想。

1852 年，Thomas Gutherie 发现了著名的四色问题。然后在 1856 年，Thomas P. Kirkman 和 William R.Hamilton 研究了多面体的循环，并通过研究仅访问某些地点一次的旅行，发明了称为哈密顿图的概念。1913 年，H.Dudeney 提到了一个难题。尽管发明了四色问题，但 Kenneth Appel 和 Wolfgang Haken 在一个世纪后才解决了这个问题。这一次被认为是图论真正的诞生。

Caley 研究了微分学的特定分析形式来研究树。这在理论化学中有许多含义。这也导致了枚举图论（enumerative graph theory）的发明。不管怎么说，“图”这个术语是由 Sylvester 在 1878 年引入的，他在“量子不变量”与代数和分子图的协变量之间进行了类比。

1941 年，Ramsey 致力于着色问题，这产生了另一个图论的分支 - 极值图论（Extremal graph theory）。1969 年，Heinrich 使用计算机解决了四色问题。对渐近图连通性的研究产生了随机图论。图论和拓扑学的历史也密切相关，它们有许多共同的概念和定理。

图的好处

- 图提供了一种处理关系和交互等抽象概念的更好的方法。它还提供了直观的视觉方式来思考这些概念。图很自然地成了分析社会关系的基础。

- 图数据库已成为一种常用的计算工具，并且是 SQL 和 NoSQL 数据库的替代方案。
- 图用于以 DAG（定向非循环图）的形式建模分析工作流。
- 一些神经网络框架还使用 DAG 来模拟不同层中的各种操作。
- 图理论用于研究和模拟社交网络，欺诈模式，功耗模式，社交媒体的病毒性和影响力。社交网络分析（SNA）可能是图理论在数据科学中最著名的应用。
- 它用于聚类算法 - 特别是 K-Means。
- 系统动力学也使用一些图理论 - 特别是循环。
- 路径优化是优化问题的一个子集，它也使用图的概念。
- 从计算机科学的角度来看，图提供了计算效率。某些算法的 Big O 复杂度对于以图形式排列的数据更好（与表格数据相比）。

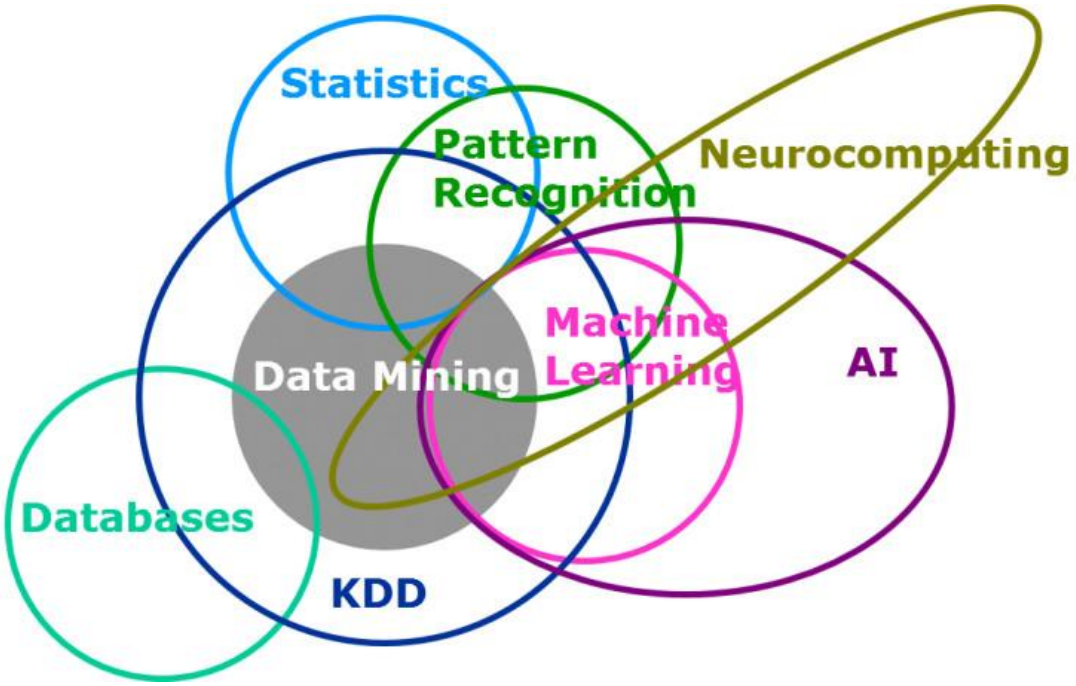
必备术语

- 顶点 u 和 v 称为边（u，v）的末端顶点。
- 如果两条边具有相同的末端顶点，则它们是平行的。
- 形式为（v，v）的边是循环。
- 如果图没有平行边和循环，则图被称为简单图。
- 如果图没有边，则称其为 Empty，即 E 是空的。
- 如果图没有顶点，则称其为 Null，即 V 和 E 是空的。
- 只有 1 个顶点的图是一个 Trivial graph。
- 具有共同顶点的边是相邻的。具有共同边的顶点是相邻的。

- 顶点 v 的度，写作 d（v），是指以 v 作为末端顶点的边数。按照惯例，我们把一个循环计作两次，并且平行边缘分别贡献一个度。
- 孤立顶点是度数为 1 的顶点。d（1）顶点是孤立的。
- 如果图的边集合包含了所有顶点之间的所有可能边，则图是完备的。
- 图 G =（V，E）中的步行（Walk）是指由图中顶点和边组成的一个形如 ViEiViEi 的有限交替序列。
- 如果初始顶点和最终顶点不同，则 Walk 是开放的（Open）。如果初始顶点和最终顶点相同，则 Walk 是关闭的（Closed）。
- 如果任何边缘最多遍历一次，则步行是一条 Trail。
- 如果任何顶点最多遍历一次，则 Trail 是一条路径 Path（除了一个封闭的步行）。
- 封闭路径（Closed Path）是一条回路 Circuit，类似于电路。

附录

算法或理论	用到的数学知识点
贝叶斯分类器	随机变量，贝叶斯公式，随机变量独立性，正态分布，最大似然估计
决策树	概率，熵，Gini 系数
KNN 算法	距离函数
主成分分析	协方差矩阵，散布矩阵，拉格朗日乘数法，特征值与特征向量
流形学习	流形，最优化，测地线，测地距离，图，特征值与特征向量
线性判别分析	散度矩阵，逆矩阵，拉格朗日乘数法，特征值与特征向量
支持向量机	点到平面的距离，Slater 条件，强对偶，拉格朗日对偶，KKT 条件，凸优化，核函数，Mercer 条件
logistic	概率，随机变量，最大似然估计，梯度下降法，凸优化，牛顿法
随机森林	抽样，方差
AdaBoost 算法	概率，随机变量，极值定理，数学期望，牛顿法
隐马尔科夫模型	概率，离散型随机变量，条件概率，随机变量独立性，拉格朗日乘数法，最大似然估计
条件随机场	条件概率，数学期望，最大似然估计
高斯混合模型	正态分布，最大似然估计，Jensen 不等式
人工神经网络	梯度下降法，链式法则
卷积神经网络	梯度下降法，链式法则
循环神经网络	梯度下降法，链式法则
生成对抗网络	梯度下降法，链式法则，极值定理，Kullback-Leibler 散度，Jensen-Shannon 散度，测地距离，条件分布，互信息
K-means 算法	距离函数
贝叶斯网络	条件概率，贝叶斯公式，图
VC 维	Hoeffding 不等式


Statistics
Pattern Recognition
Neurocomputing
Machine Learning
AI
Data Mining
Databases
KDD