机器学习之概率统计基础

ML&DM 机器学习 概率论 数理统计

julyedu四月机器学习算法班第一次课中关于概率论和第二次课中关于数理统计、参数估计的笔记，也有自己的一些补充。

前言

概率论与数理统计的关注点不同：

- 概率论：已知总体的分布情况，求某个事件发生的概率。
  例如装箱问题：将12件正品和3件次品随机的装在3个箱子中，每箱装5件，则每箱中恰有1件次品的概率是多少?
- 数理统计：已知样本，估计总体。是概率论的逆向工程
  例如正态分布的矩估计：在正态分布总体中采样得到n个样本$X_1, X_2, \cdots, X_n$，估计该总体的均值和方差。

概率统计与机器学习：

- 训练：从样本估计总体（模型），就是数理统计
- 预测：已知总体（模型），求某个事件发生的概率，就是概率论

对于n个特征 $X_1, X_2, \cdots, X_n$ 的样本，每个特征 X_i 的值可以看作是一个分布，对于有监督学习来说每个样本的标签 Y 的值也是一个分布。可以基于各个分布的特性来评估模型和样本

概率基础

概率：$P(X) \in [0, 1]$

累积分布函数：$\Phi(x) = P(X \leq x)$

- $\Phi(x)$ 一定为**单增函数**
- $\min(\Phi(x)) = 0$，$\max(\Phi(x)) = 1$
- 将值域为 $[0, 1]$ 的某函数 $yf(x)$ **看成** x **事件的累积概率**
- 若 $y = f(x)$ 可导，则 $p(x) = f'(x)$ 为某概率密度函数

概率密度函数：连续随机变量X的分布函数为 $\Phi(x) = \int_{-\infty}^{x} f(t)dt$，$f(x)$ 称为 X 的概率密度函数

条件概率：$P(A \mid B) = \frac{P(AB)}{P(B)}$
全概率公式：$P(A) = \sum_i P(A \mid B_i)P(B_i)$
贝叶斯(Bayes)公式：$P(B_i \mid A) = \frac{P(AB_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)}$

先验概率、后验概率

例：对以往数据分析表明，当机器调整的好时，产品的合格率为98%，而当机器发生某种故障时，其合格率为55%。每天早上机器开动时，机器调整良好的概率为95%，试求已知某日早上第一件产品是合格品时，机器调整良好的概率是多少？

解：设A为事件“产品合格”，B为事件“机器调整良好”

则 $P(A \mid B)=0.98$，$P(A \mid \overline{B})=0.55$

$P(B)=0.95$，$P(\overline{B})=0.05$

由贝叶斯公式：

$$\begin{aligned} P(B \mid A) &= \frac{P(AB)}{P(A)} \\ &= \frac{P(A \mid B)P(B)}{P(A \mid B)P(B)+P(A \mid \overline{B})P(\overline{B})} \\ &= \frac{0.98 \times 0.95}{0.98 \times 0.95+0.55 \times 0.05} \\ &= 0.97 \end{aligned}$$

概率0.95是由以往数据分析得到的，叫作**先验概率**；

得到信息后（第一件产品是合格品）再重新加以修正的概率（0.97）叫**后验概率**。

两个学派

给定某系统的若干样本，求系统的参数

- 频率学派：假定参数是某个/某些未知的定值，求这些参数如何取值，能够使得某目标函数取极大、极小。如矩估计/MLE/MaxEnt/EM等。
- 贝叶斯学派：假定参数本身是变化的，服从某个分布，求在这个分布约束下使得某目标函数极大/极小。

常见分布

离散型随机变量的分布

0-1分布

$$P\{X=k\}=p^k(1-p)^{1-k},\ k=0,1 \quad (0<p<1)$$

期望：$E(X)=1 \cdot p+0 \cdot (1-p)=p$

方差：$D(X)=E(X^2)-[E(X)]^2=1^2 \cdot p+0^2 \cdot (1-p)-p^2=p(1-p)$

$$
\begin{aligned}
D(X) &= E\{[X-E(X)]^2\} \\
&= E\{X^2 - 2XE(X) + [E(X)]^2\} \\
&= E(X^2) - 2E(X)E(X) + [E(X)]^2 \\
&= E(X^2) - [E(X)]^2
\end{aligned}
$$

二项(Bernoulli)分布

试验 E 只有两个可能的结果：A 及 $\overline{A}$，则称 E 为伯努利（Bernoulli）试验。设 $P(A) = p\ (0 < p < 1)$，此时 $P(\overline{A}) = 1-p$。将 E 独立重复地进行n次，称为**n重伯努利试验**。n次中 A 出现k次的概率：

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

记 $X \sim b(n, p)$

设 X_i为第 i 次试验，显然 X_i服从参数为 p 的0-1分布，且有 $X = \sum_{i=1}^{n} X_i$，

期望：$E(X) = \sum_{i=1}^{n} E(X_i) = np$

方差：$D(X) = \sum_{i=1}^{n} D(X_i) = np(1-p)$

泊松(Possion)分布

$P\{X=k\} = \frac{\lambda^k e^{-\lambda}}{k!},\ k = 0, 1, 2, \cdots,$

记 $X \sim \pi(\lambda)$

期望：$E(X) = \sum_{k=0}^{\infty} k \frac{\lambda^k e^{-\lambda}}{k!} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} \cdot e^{\lambda} = \lambda$

$$
\begin{aligned}
E(X^2) &= E[X(X-1) + X] = E[X(X-1)] + E(X) \\
&= \sum_{k=0}^{\infty} k(k-1) \frac{\lambda^k e^{-\lambda}}{k!} + \lambda = \lambda^2 e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} + \lambda \\
&= \lambda^2 e^{-\lambda} e^{\lambda} + \lambda = \lambda^2 + \lambda
\end{aligned}
$$

方差：$D(X) = E(X^2) - [E(X)]^2 = \lambda$

泰勒展开与泊松分布

$$
e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^k}{k!} + R_k
$$

$$
\xrightarrow{\text{两边同时除以} e^x} 1 = 1 \cdot e^{-x} + x \cdot e^{-x} + \frac{x^2}{2!} \cdot e^{-x} + \frac{x^3}{3!} + \cdots + \frac{x^k}{k!} \cdot e^{-x} + R_k \cdot e^{-x}
$$

通项公式 $\frac{x^k}{k!} \cdot e^{-x} \to \frac{\lambda^k}{k!} \cdot e^{-\lambda}$

可以发现，每一项的总和为1（分布函数的最大值）

连续型随机变量的分布

均匀分布

> 概率密度为：
> $$f(x)=\begin{cases}\frac{1}{b-a}, & a<x<b\\ 0, & \text{其他}\end{cases}$$

期望：$E(X)=\int_{-\infty}^{\infty}xf(x)dx=\int_a^b\frac{1}{b-a}xdx=\frac{1}{2}(a+b)$
方差：$D(X)=E(X^2)-[E(X)]^2=\int_a^b\frac{1}{b-a}x^2dx-(\frac{a+b}{2})^2=\frac{(b-a)^2}{12}$

指数分布

> 概率密度函数：
> $$f(x)=\begin{cases}\frac{1}{\theta}e^{-x/\theta}, & x>0\\ 0, & \text{其他}\end{cases}$$

期望：$E(X)=\int_{-\infty}^{\infty}xf(x)dx=\int_0^{\infty}x\cdot\frac{1}{\theta}e^{-x/\theta}dx=-xe^{-x/\theta}\,|_0^{\infty}+\int_0^{\infty}e^{-x/\theta}dx=\theta$
方差：$D(X)=E(X^2)-[E(X)]^2=\int_0^{\infty}x^2\cdot\frac{1}{\theta}e^{-x/\theta}dx-\theta^2=2\theta^2-\theta^2=\theta^2$

指数分布的无记忆性：$P(x>s+t\mid x>s)=P(x>t)$
即，如果 x 是某电器元器件的寿命，已知元器件使用了 s 小时，则共使用至少 $s+t$ 小时的条件概率，与从未使用开始至少使用 t 小时的概率相等。

正态分布

> 连续型随机变量X 概率密度函数：
> $$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ \sigma>0,\ -\infty<x<+\infty$$
> 标准正态分布：
> $$\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$$

标准正态变量 $Z=\frac{X-\mu}{\sigma}$ 的期望和方差：
期望：$E(Z)=\dfrac{1}{\sqrt{2\pi}}\displaystyle\int_{-\infty}^{\infty}te^{-\frac{t^2}{2}}dt=\frac{-1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}\Big|_{-\infty}^{\infty}=0$
方差：$D(Z)=E(Z^2)-0^2=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}t^2e^{-\frac{t^2}{2}}dt=\frac{-1}{\sqrt{2\pi}}te^{-\frac{t^2}{2}}\Big|_{-\infty}^{\infty}+\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\frac{t\$2}{2}}dt=1$

正态分布的期望河方差：
期望：$E(X)=E(\mu+\sigma Z)=\mu$
方差：$D(X)=D(\mu+\sigma Z)=D(\sigma Z)=\sigma^2D(Z)=\sigma^2$

重要统计量

期望

- 离散型 $E(X)=\sum_i x_i p_i$
- 连续型 $E(X)=\int_{-\infty}^{\infty} xf(x)dx$

期望的性质

- $E(kX)=kE(X)$
- $E(X+Y)=E(X)+E(Y)$
- 若 X 和 Y 相互独立，$E(XY)=E(X)E(Y)$。反之不成立，若 $E(XY)=E(X)E(Y)$，只能说明 X 和 Y 不相关

> 独立：$P(AB)=P(A)P(B)$
> 互斥：$P(AB)=0, P(A+B)=P(A)+P(B)$

方差

$$D(X)=E\{[X-E(X)]^2\}=E(X^2)-[E(X)]^2$$

方差的性质

- $D(c)=0$
- $D(X+c)=D(X)$
- $D(kX)=k^2D(X)$
- 若 X 和 Y 独立，$D(X+Y)=D(X)+D(Y)$

协方差（机器学习中很重要）
评价两个随机变量的关系

$$Cov(X,Y)=E\{[X-E(X)][Y-E(Y)]\}$$

协方差的性质

- $Cov(X,Y)=Cov(Y,X)$
- $Cov(aX+b,cY+d)=acCov(X,Y)$
- $Cov(X_1+X_2,Y)=Cov(X_1,Y)+Cov(X_2,Y)$
- $Cov(X,Y)=E(XY)-E(X)E(Y)$
- X 和 Y独立，$Cov(X,Y)=0$
- 若$Cov(X,Y)=0$，称 X 和 Y 不相关

协方差的意义：是两个随机变量具有相同方向变化趋势的度量

- 若 $Cov(X,Y)>0$，它们的变化趋势相同
- 若 $Cov(X,Y)<0$，它们的变化趋势相反
- 若 $Cov(X,Y)=0$，称 X 和 Y 不相关

协方差可以用于降维（找出与结果不（线性）相关的特征）

相关系数

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{D(X)D(Y)}}$$

> 注意协方差的上界：$|Cov(X,Y)| \leq \sqrt{D(X)D(Y)} = \sigma_1\sigma_2$，即$|\rho| \leq 1$，当且仅当X与Y有线性关系时，等号成立。

可以看出，相关系数时标准尺度下的协方差。

协方差矩阵

对于n个随机向量 $(X_1, X_2, \cdots, X_n)$，任意两个元素 X_i 和 X_j 都可以得到一个协方差，从而形成 $n \times n$ 的矩阵，它是一个对称阵。

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix}$$

其中，$c_{ij} = E\{[X_i - E(X_i)][X_j - E(X_j)]\} = Cov(X_i, X_j)$

相关系数矩阵

矩

对随机变量X，X的k阶原点矩为

$$E(X^k)$$

X的k阶中心矩为

$$E\{[X - E(X)]^k\}$$

X和Y的k+l阶混合矩为

$$E(X^k Y^l)$$

X和Y的k+l阶混合中心矩为

$$E\{[X - E(X)]^k [Y - E(Y)]^l\}$$

- 期望E(X)是X的一阶原点矩
- 方差D(X)是X的二阶中心矩
- 协方差Cov(X,Y)是X和Y的二阶混合中心矩

重要定理和不等式

Jensen不等式

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$$

若$\theta_1, \cdots, \theta_k \geq 0,\ \theta_1 + \cdots + \theta_k = 1$，则

$$f(\theta_1 x_1 + \cdots + \theta_k x_k) \leq \theta_1 f(x_1) + \cdots + \theta_k f(x_k)$$

即 $f(Ex) = Ef(x)$

切比雪夫不等式

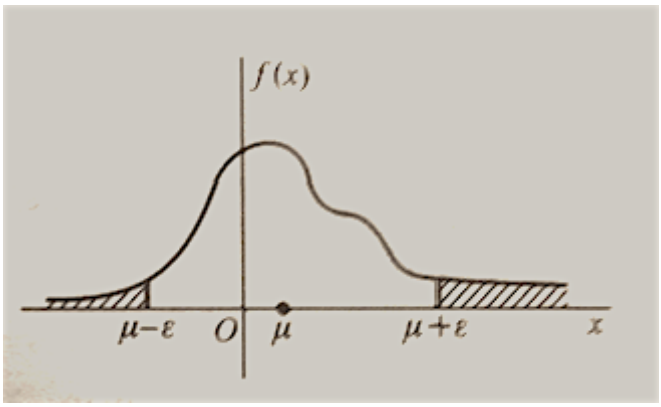
设随机变量X具有数学期望 $E(X) = \mu$，方差 $D(X) = \sigma^2$，则对于任意正数 ε，不等式

$$P\{|\ X - \mu\ | \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2}$$

成立。

$$\begin{aligned} P\{|\ X - \mu\ | \geq \varepsilon\} &= \int_{|x-\mu| \geq \varepsilon} f(x)dx \\ &\leq \int_{|x-\mu| \geq \varepsilon} \frac{|x-\mu|^2}{\varepsilon^2} f(x)dx \\ &\leq \frac{1}{\varepsilon^2} \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \\ &= \frac{1}{\varepsilon^2} \sigma^2 \end{aligned}$$

切比雪夫不等式说明，X的方差越小，事件 $\{|\ X - \mu\ | < \varepsilon\}$ 发生的概率越大。即：X取的值基本上集中在期望 μ 附近。



大数定理

设 $X_1, X_2, \cdots$ 是相互独立，服从同一分布的随机变量序列，且具有数学期望 $E(X_k) = \mu\ (k = 1, 2, \ldots)$。作前n个变量的算术平均 $Y_n = \frac{1}{n} \sum_{k=1}^{n} X_k$，则对于任意 $\varepsilon > 0$，有

$$\lim_{x \to \infty} P\{|\ Y_n - \mu\ | < \varepsilon\} = 1$$

$$E(Y)=E\left(\frac{1}{n}\sum_{k=1}^{n}X_k\right)=\frac{1}{n}\sum_{k=1}^{n}E(X_k)=\frac{1}{n}\cdot(n\mu)=\mu$$

$$D(Y)=D\left(\frac{1}{n}\sum_{k=1}^{n}X_k\right)=\frac{1}{n^2}\sum_{k=1}^{n}D(X_k)=\frac{1}{n^2}\cdot(n\sigma^2)=\frac{\sigma^2}{n}$$

由切比雪夫不等式有，$P\{|\ Y-\mu_Y\ |<\varepsilon\}\geq 1-\frac{\sigma_Y^2}{\varepsilon^2}$

即：$P\left\{\left|\frac{1}{n}\sum_{k=1}^{n}X_k-\mu\right|<\varepsilon\right\}\geq 1-\frac{\sigma^2/n}{\varepsilon^2}$

且，$\lim_{n\to\infty}1-\frac{\sigma^2/n}{\varepsilon^2}=1$，$P\left\{\left|\frac{1}{n}\sum_{k=1}^{n}X_k-\mu\right|<\varepsilon\right\}$ 始终小于等于1，

所以，$\lim_{x\to\infty}P\{|\ Y_n-\mu\ |<\varepsilon\}=1$

通俗的说就是样本均值收敛于总体均值。

也称弱大数定理、辛钦大数定理。

伯努利大数定理

设 n_A 是n次独立重复事件A发生的次数，P是事件A在每次试验中发生的概率，则对于任意正数 $\varepsilon>0$，有

$$\lim_{n\to\infty}P\left\{\left|\frac{n_A}{n}-p\right|<\varepsilon\right\}=1$$

中心极限定理

设随机变量 $X_1,X_2,\cdots,X_n,\cdots$ 独立同分布，且具有期望和方差：

$E(X_k)=\mu,\ D(X_k)=\sigma^2\ (k=1,2,\cdots)$，则随机变量之和 $\sum_{k=1}^{n}X_k$ 的标准化变量

$$Y_n=\frac{\sum_{k=1}^{n}X_k-n\mu}{\sqrt{n}\sigma}$$

的分布收敛到标准正态分布。

即n充分大时，$\frac{\sum_{k=1}^{n}X_k-n\mu}{\sqrt{n}\sigma}\overset{近似地}{\sim}N(0,1)$

或 $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\overset{近似地}{\sim}N(0,1)$

或 $\overline{X}\overset{近似地}{\sim}N(\mu,\frac{\sigma^2}{n})$

用样本估计总体参数

总体X的分布函数$F(x;\theta)$的形式已知，θ 无法求解，但可以利用X的样本 $X_1, X_2, \cdots, X_n$ 得到一个估计值 $\hat{\theta}$。常用的估计方法有矩估计和极大似然估计。

矩估计

k阶样本的原点矩为 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$

k阶样本的中心矩为 $M_k = \frac{1}{n}\sum_{i-1}^{n}(X_i - \overline{X})^k$

矩估计就是用样本矩作为相应的总体矩的估计量。

> **例：**设总体X在[a,b]上服从均匀分布，a,b未知。 $X_1, X_2, \cdots, X_n$ 是来自X的样本，试求a,b的矩估计量。
> **解：**
> $\mu_1 = E(X) = (a+b)/2$
> $\mu_2 = E(X^2) = D(X) + [E(X)]^2 = (b-a)^2/12 + (a+b)^2/4$
> 即 $\begin{cases} a+b = 2\mu_1 \\ b-a = \sqrt{12(\mu_2 - \mu_1^2)} \end{cases}$
> 解得 $a = \mu_1 - \sqrt{3(\mu_2 - \mu_1^2)},\ b = \mu_1 + \sqrt{3(\mu_2 - \mu_1^2)}$
> 用样本矩作为总体矩的估计就可以得到a,b的估计量
> $$\hat{a} = A_1 - \sqrt{3(A_2 - A_1^2)} = \overline{X} - \sqrt{3\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2\right)}$$
> $$\hat{b} = A_1 + \sqrt{3(A_2 - A_1^2)} = \overline{X} + \sqrt{3\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2\right)}$$

极大似然估计

设总体X的分布（概率密度）为$f(x;\theta)$， $X_1, X_2, \cdots, X_n$ 是来自X的样本，其联合密度为

$$L(x_1, x_2, \cdots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

$x_1, x_2, \cdots, x_n$ 是相应这个样本的样本值。

这里，θ 被看作固定但未知的参数；而样本已存在，$x_1, x_2, \cdots, x_n$ 是固定的，$L(x, \theta)$ 是关于 θ 的函数，即似然函数。

求参数 θ 的值，使得似然函数取极大值，这种方法就是极大似然估计。

不同的分布有不同的 θ 值，相同分布不同参数也是不同的 θ 值。这里讨论分布只有一个变量 θ ，更多情况下会是 $\theta_1, \theta_2, \cdots$ ，也是类似的。

从贝叶斯公式来看极大似然

给定样本D，在这些样本中计算某结论 $A_1, A_2, \cdots, A_n$ 出现的概率，即 $P(A_i|D)$

极大似然就是找出 $P(A_i|D)$ 最大时的 A_i 的值

$$\begin{aligned}\max P(A_i|D) &= \max \frac{P(D|A_i)P(A_i)}{P(D)} \\ &= \max(P(D|A_i)P(A_i)) \\ &\xrightarrow{P(A_i)\text{近似想等}} \max P(D|A_i)\end{aligned}$$

实际中，由于求导的需要，往往将似然函数取对数，得到对数似然函数。

$$\log L(x_1, x_2, \cdots, x_n; \theta) = \sum_{i=1}^{n} \log f(x_1, x_2, \cdots, x_n; \theta)$$

用导数等于0，求最大值：

$$\frac{d}{d\theta} \ln L(\theta) = 0$$

参考资料

julyedu四月机器学习班第1课、第2课
概率论与数理统计 浙江大学
机器学习 周志华