机器学习常用数学公式（二）：概率统计

李新春 lxcnju@163.com

2017 年 5 月 20 日

概率论和统计学的知识在机器学习中应用非常广泛。比如，朴素贝叶斯分类算法基于各个类别的先验概率、特征属性值的条件概率求使得后验概率最大类，逻辑斯蒂回归利用极大似然估计进行求解最优化,概率图模型更是和概率计算、参数估计息息相关，贝叶斯网络、马尔可夫随机场属于概率图模型，二者的重要性不言而喻。本文总结一下经常用到的概率统计的相关知识。主要分为统计量、分布（族）、常用概率不等式这三个方面展开。

1 统计量

1、样本均值（sample mean）、样本方差（sample variance）

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,\ S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{1.1}$$

2、样本k阶原点矩（moment）、样本k阶中心距

$$a^k = \frac{1}{n}\sum_{i=1}^{n} X_i^k,\ m_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k \tag{1.2}$$

3、样本偏度（skewness）、样本峰度（kurtosis）

$$\frac{m_3}{m_2^{\frac{3}{2}}} = \frac{\sqrt{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{(\sum_{i=1}^{n}(X_i - \bar{X})^2)^{\frac{3}{2}}},\ \frac{m_4}{m_2^2} - 3 = \frac{n\sum_{i=1}^{n}(X_i - \bar{X})^4}{(\sum_{i=1}^{n}(X_i - \bar{X})^2)^2} - 3 \tag{1.3}$$

4、样本相关系数（correlation coefficient）

$$r(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{1.4}$$

5、次序统计量（order statistic）

$$X_1, X_2, \cdots, X_n \Rightarrow X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)} \tag{1.5}$$

6、极大值、极小值、极差（range）、样本p分位数（quantile）

$$X_{(n)},\ X_{(1)},\ X_{(n)} - X_{(1)},\ X_{([np])}\ or\ X_{([np]+1)} \tag{1.6}$$

7、样本变异系数（coefficient of variation）

$$\frac{S_n}{\bar{X}} \tag{1.7}$$

2 分布（族）

1、0-1分布（伯努利分布）

$$X \sim b(1,p), P(X=1)=p \tag{2.1}$$

$$E(X)=p,\ VAR(X)=p(1-p) \tag{2.2}$$

2、二项分布

$$X \sim b(n,p), P(X=k)=C_n^k p^k (1-p)^{n-k} \tag{2.3}$$

$$E(X)=np,\ VAR(X)=np(1-p) \tag{2.4}$$

3、泊松分布

$$X \sim P(\lambda), P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda} \tag{2.5}$$

$$E(X)=\lambda,\ VAR(X)=\lambda \tag{2.6}$$

对于二项分布，n趋近于无穷时可转化为泊松分布。取$p=\frac{\lambda}{n}$，有：

$$\begin{aligned}
p(X=k) &= C_n^k p^k (1-p)^{n-k} \\
&= \frac{n!}{k!(n-k)!}\frac{\lambda^k}{n^k}(1-\frac{\lambda}{n})^(n-k) \\
&= \frac{\lambda^k}{k!}\frac{n!}{(n-k)!n^k}(1-\frac{\lambda}{n})^n(1-\frac{\lambda}{n})^{-\lambda} \\
&= \frac{\lambda^k}{k!}\frac{n!}{(n-k)!n^k}(1-\frac{\lambda}{n})^n(1-\frac{\lambda}{n})^{-\lambda} \\
&= \frac{\lambda^k}{k!}\frac{n-k+1}{n}\cdots\frac{n}{n}(1-\frac{\lambda}{n})^n(1-\frac{\lambda}{n})^{-\lambda}
\end{aligned}$$

$$\lim_{n\to\infty} p(X=k) = \lim_{n\to\infty} C_n^k p^k (1-p)^{n-k} = \frac{\lambda^k}{k!}e^{-\lambda} \tag{2.7}$$

4、超几何分布（M件正品、N-M件次品，无放回抽样n件产品，有X件正品）

$$X \sim h(n,M,N), P(X=k)=\frac{C_M^k C_{N-M}^{n-k}}{C_N^n} \tag{2.8}$$

$$E(X)=n\frac{M}{N},\ VAR(X)=n\frac{M(N-M)(N-n)}{N^2(N-1)} \tag{2.9}$$

对于超几何分布，当$n \ll N$时转化为二项分布。取$p=\frac{M}{N}$，有：

$$\begin{aligned}
p(X=k) &= \frac{C_M^k C_{N-M}^{n-k}}{C_N^n} \\
&= \frac{M!}{k!(M-k)!}\frac{(N-M)!}{(n-k)!(N-M-n+k)!}\frac{n!(N-n)}{N!} \\
&= \frac{n!}{k!(n-k)!}\frac{M!}{(M-k)!}\frac{(N-M)}{(N-M-n+k)!}\frac{(N-n)}{N!} \\
&= C_n^k \frac{((M-k+1)\cdots M)((N-M-n+k+1)\cdots(N-M))}{(N-n+1)\cdots N} \\
&= C_n^k \frac{(\frac{M}{N}-\frac{k-1}{N})\cdots(\frac{M}{N})(1-\frac{M}{N}-\frac{n-k-1}{N})\cdots(1-\frac{M}{N})}{(1-\frac{n-1}{N})\cdots(\frac{N}{N})}
\end{aligned}$$

$$\lim_{n\ll N} p(X=k) = C_n^k(\frac{M}{N})^k(1-\frac{M}{N})^{n-k} = C_n^k p^k(1-p)^{n-k} \tag{2.10}$$

5、几何分布（首次击中时的试验次数X）

$$X \sim Ge(p), P(X=k) = (1-p)^{k-1}p \tag{2.11}$$

$$E(X) = \frac{1}{p},\ VAR(X) = \frac{1-p}{p^2} \tag{2.12}$$

几何分布的无记忆性：

$$\begin{aligned} p(X>m+n|X>n) &= \frac{p(X>m+n, X>n)}{p(X>n)} \\ &= \frac{p(X>m+n)}{p(X>n)} \\ &= \frac{\sum_{k=m+n+1}^{\infty}(1-p)^{k-1}p}{\sum_{k=n+1}^{\infty}(1-p)^{k-1}p} \\ &= \frac{(1-p)^{m+n}}{(1-p)^n} \\ &= (1-p)^m = p(X>m) \end{aligned}$$

6、均匀分布

$$f(x) = \begin{cases} \frac{1}{b-a}, & a<x<b \\ 0, & otherwise \end{cases} \tag{2.13}$$

$$F(x) = \begin{cases} 0, & x<a \\ \frac{x-a}{b-a}, & a\le x<b \\ 1, & x\ge b \end{cases} \tag{2.14}$$

$$X \sim U(a,b),\ E(X) = \frac{a+b}{2},\ VAR(X) = \frac{(b-a)^2}{12} \tag{2.15}$$

7、正态分布

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.16}$$

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\,dt \tag{2.17}$$

$$X \sim N(\mu, \sigma^2),\ E(X) = \mu,\ VAR(X) = \sigma^2 \tag{2.18}$$

多元正态分布：

$$f(x_1, x_2, \cdots, x_p) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\exp(-(\mathbf{x}-\mu)^{\mathbf{T}}\mathbf{\Sigma^{-1}}(\mathbf{x}-\mu)) \tag{2.19}$$

特别地，取协方差矩阵为$\begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$，则得到二元正态分布，$\rho$为相关系数：

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)\right) \tag{2.20}$$

8、指数分布

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x\ge 0 \\ 0, & x<0 \end{cases} \tag{2.21}$$

$$F(x)=\begin{cases}1-e^{-\lambda x}, & x\geq 0\\ 0, & x<0\end{cases} \tag{2.22}$$

$$X\sim E(\lambda),\ E(X)=\frac{1}{\lambda},\ VAR(X)=\frac{1}{\lambda^2} \tag{2.23}$$

指数分布和泊松分布的关系：泊松分布$P(\lambda)$的均值λ可解释为泊松流在单位时间内出现的质点数，那么在时间段$(0,t]$时间内出现的质点数X(t)也服从泊松分布，均值为$t\lambda$，则$X(t)\sim P(t\lambda), P(X(t)=k)=\frac{(\lambda t)^k e^{-\lambda t}}{k!}$。考察泊松流第一个质点出现的时间，记为随机变量$T_1$，则得到$P(T_1>t)=P(X(t)=0)=e^{-\lambda t}$。则有：

$$F(x)=p(T_1\leq x)=1-p(T_1>x)=1-e^{-\lambda x},\ x\geq 0 \tag{2.24}$$

从上可知，泊松分布描述的是单位时间内出现的质点数X的分布，故是离散型分布；指数分布刻画的是第一个质点出现的时刻，故为连续型分布。同时，利用同样地方法可得到，第n个质点出现的时刻服从Gamma分布。同时，指数分布和几何分布一样具有无记忆性。

9、对数正态分布

$$f(x)=\frac{1}{x\sqrt{2\pi}\sigma}e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}} \tag{2.25}$$

$$X\sim LN(\mu,\sigma^2),\ E(X)=e^{\mu+\sigma^2/2},\ VAR(X)=(e^{\sigma^2}-1)e^{2\mu+\sigma^2} \tag{2.26}$$

10、柯西分布（当$\mu=0,\gamma=1$时为标准柯西分布）

$$f(x;\mu,\gamma)=\frac{1}{\pi\gamma\left(1+(\frac{x-\mu}{\gamma})^2\right)} \tag{2.27}$$

$$F(x)=\frac{1}{2}+\frac{1}{\pi}\arctan\left(\frac{x-\mu}{\gamma}\right) \tag{2.28}$$

$$X\sim Cau(\mu,\lambda),\ E(X)=None,\ VAR(X)=None \tag{2.29}$$

11、χ^2分布

$$f(x)=\begin{cases}\frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}e^{-\frac{x}{2}}x^{\frac{n}{2}-1}, & x>0\\ 0, & x\leq 0\end{cases} \tag{2.30}$$

$$F(x)=\frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}\int_0^x e^{-\frac{t}{2}}t^{\frac{n}{2}-1}\,dt \tag{2.31}$$

$$X\sim \chi^2(n),\ E(X)=n,\ VAR(X)=2n \tag{2.32}$$

求解χ^2分布的均值和方差需要借助于特征函数$\psi(t)$的技巧。

$$\begin{aligned}\psi(t) &= E(e^{ixt})\\ &= \int_0^\infty e^{ixt}\frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}e^{-\frac{x}{2}}x^{\frac{n}{2}-1}\,dx\\ &= \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}\int_0^\infty e^{-(\frac{1}{2}-it)x}x^{\frac{n}{2}-1}\,dx\\ &= \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}\frac{\Gamma(\frac{n}{2})}{\left(\frac{1}{2}-it\right)^{\frac{n}{2}}}\\ &= (1-2it)^{-\frac{n}{2}}\\ E(x^k) &= \frac{\psi^{(n)}(0)}{i^n}\end{aligned}$$

其中，Γ函数为$\Gamma(n)=\int_0^{\infty}e^{-x}x^{n-1}\,dx$。

12、瑞利分布

$$f(x)=\frac{x\exp(-\frac{x^2}{2\sigma^2})}{\sigma^2} \tag{2.33}$$

$$F(x)=1-\exp(-\frac{x^2}{2\sigma^2}) \tag{2.34}$$

$$X\sim R(\sigma),\ E(X)=\sqrt{\frac{\pi}{2}}\sigma,\ VAR(X)=\frac{4-\pi}{2}\sigma^2 \tag{2.35}$$

13、Γ分布族

$$f(x;\alpha,\lambda)=\frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x},\ x>0 \tag{2.36}$$

$$X\sim Ga(\alpha,\lambda),\ \psi(t)=\left(1-\frac{it}{\lambda}\right)^{-\alpha},\ E(X)=\frac{\alpha}{\lambda},\ VAR(X)=\frac{\alpha}{\lambda^2} \tag{2.37}$$

可以看出，$\chi^2(n)=\Gamma(\frac{n}{2},\frac{1}{2}),E(\lambda)=\Gamma(1,\lambda)$。

14、β分布族

$$f(x;a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1},\ 0\le x\le 1 \tag{2.38}$$

$$X\sim\beta(a,b),(a>0,b>0)\ E(X)=\frac{a}{a+b},\ VAR(X)=\frac{ab}{(a+b)^2(a+b+1)} \tag{2.39}$$

可以看出，$U(0,1)=\beta(1,1)$。

15、指数型分布族

$$f(x,\theta)=c(\theta)\exp\left(\sum_{i=1}^{k}c_i(\theta)T_i(x)\right)h(x),\ c(\theta)>0,h(x)>0 \tag{2.40}$$

or

$$f(x,\eta)=\exp\left(\sum_{i=1}^{k}\eta_iT_i(x)-a(\eta)\right)h(x),\ \eta\in\Xi=\{\eta(\theta):\theta\in\Theta\} \tag{2.41}$$

Γ分布族、β分布族、泊松分布、二项分布、正态分布等都是指数型分布。拿二项分布举例说明如下：

$$\begin{aligned} f(x) &= C_n^xp^x(1-p)^{n-x} && (2.42)\\ &= C_n^x(1-p)^n\left(\frac{p}{1-p}\right)^x && (2.43)\\ &= C_n^x(1-p)^n\exp\left(x\ln\frac{p}{1-p}\right) && (2.44)\end{aligned}$$

取$c(p)=(1-p)^n,c_1(p)=\ln\frac{p}{1-p},T_1(x)=x,h(x)=C_n^x$，则可得二项分布为指数型分布族。

3 常用概率不等式

1、马尔可夫不等式

$$P(Z\ge t)\le\frac{E[Z]}{t} \tag{3.1}$$

证明：$P(Z \geq t) = E[\mathbf{I}\{Z \geq t\}]$，如果$Z \geq t$，$\frac{Z}{t} \geq 1 \geq \mathbf{I}\{Z \geq t\}$；如果$Z < t$，$\frac{Z}{t} \geq 0 = \mathbf{I}\{Z \geq t\}$。$\mathbf{I}$为指示函数。所以：

$$P(Z \geq t) = E[\mathbf{I}\{Z \geq t\}] \leq E\left[\frac{Z}{t}\right] = \frac{E[Z]}{t} \tag{3.2}$$

2、车比雪夫不等式

$$P(|Z - E[Z]| \geq t) \leq \frac{VAR(Z)}{t^2} \tag{3.3}$$

证明：

$$P(|Z - E[Z]| \geq t) = P((Z - E[Z])^2 \geq t^2) \leq \frac{E[(Z - E[Z])^2]}{t^2} = \frac{VAR(Z)}{t^2} \tag{3.4}$$

推论：

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i\right| \geq t\right) \leq \frac{VAR(Z_1)}{nt^2} \tag{3.5}$$

3、切诺夫界

$$P(Z - E[Z] \geq t) \leq \min_{\lambda \geq 0} E[e^{\lambda(Z - E[Z])}]e^{-\lambda t} = \min_{\lambda \geq 0} M_{Z-E[Z]}(\lambda)e^{-\lambda t} \tag{3.6}$$

and

$$P(Z - E[Z] \leq -t) \leq \min_{\lambda \geq 0} E[e^{\lambda(E[Z] - Z)}]e^{-\lambda t} = \min_{\lambda \geq 0} M_{E[Z]-Z}(\lambda)e^{-\lambda t} \tag{3.7}$$

其中$M_Z(\lambda) = E[\exp(\lambda Z)]$是矩量母函数。

证明（只证明第一式，且式子对于$\lambda \geq 0$的选取为任意的，所以在下面证明的结果里面取最小下界对结果没有影响）：

$$P(Z - E[Z] \geq t) = P(\exp(\lambda(Z - E[Z])) \geq \exp(t)) \leq \frac{E[e^{\lambda(Z - E[Z])}]}{e^{\lambda t}} \tag{3.8}$$

推论：

$$P(\sum_{i=1}^{n} Z_i \geq t) \leq \frac{M_{Z_1+Z_2+\cdots+Z_n}(\lambda)}{e^{\lambda t}} = \frac{\prod_{i=1}^{n} E[\exp(\lambda Z_i)]}{e^{\lambda t}} = \frac{(E[e^{\lambda Z_1}])^n}{e^{\lambda t}} \tag{3.9}$$

4、矩量生成函数结合切诺夫界

$$M_Z(\lambda) = E[e^{\lambda Z}] \leq \exp\left(\frac{C^2\lambda^2}{2}\right), for\ all \quad \lambda \in \mathbf{R} \tag{3.10}$$

$$M_Z(\lambda) = E[e^{\lambda Z}] = \exp\left(\frac{\sigma^2\lambda^2}{2}\right) \quad Z \sim N(0, \sigma^2) \tag{3.11}$$

$$E[e^{\lambda S}] \leq \exp\left(\frac{\lambda^2}{2}\right) \quad P(S = 1) = P(S = -1) = \frac{1}{2} \tag{3.12}$$

下面利用（3.12）式，记$Z = \sum_{i=1}^{n} S_i, P(S_i = 1) = P(S_i = -1) = \frac{1}{2}$:

$$P(Z \geq t) \leq E[e^{\lambda Z}]e^{-\lambda t} = E[e^{\lambda S_1}]^n e^{-\lambda t} \leq e^{\frac{n\lambda^2}{2} - \lambda t} \tag{3.13}$$

根据切诺夫界对于$\lambda \geq 0$的任意性，取$\lambda = \frac{t}{n}$得到：

$$P(Z \geq t) \leq e^{-\frac{t}{2n}} \tag{3.14}$$

5、Hoeffding引理

$$E[\exp(\lambda(Z-E[Z]))] \le \exp\left(\frac{\lambda^2(b-a)^2}{8}\right) \quad Z\in[a,b], for\ all\ r\in\mathbf{R} \tag{3.15}$$

6、Hoeffding不等式

$$P\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i-E[Z_i])\ge t\right)\le \exp\left(-\frac{2nt^2}{(b-a)^2}\right) \tag{3.16}$$

and

$$P\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i-E[Z_i])\le -t\right)\le \exp\left(-\frac{2nt^2}{(b-a)^2}\right) \tag{3.17}$$

参考文献

1.常见分布 $https://wenku.baidu.com/view/ab59abb8c77da26925c5b0a3.html$

2.常见离散型分布 $https://wenku.baidu.com/view/3f4360d380eb6294dd886c54.html$

3、CS229 Supplemental Lecture notes Hoeffding's inequality