机器学习中有关概率论知识的小结

一、引言

最近写了许多关于机器学习的学习笔记，里面经常涉及概率论的知识，这里对所有概率论知识做一个总结和复习，方便自己查阅，与广大博友共享，所谓磨刀不误砍柴工，希望博友们在这篇博文的帮助下，阅读机器学习的相关文献时能够更加得心应手！这里只对本人觉得经常用到的概率论知识点做一次小结，主要是 基本概念 ，因为机器学习中涉及概率论的地方，往往知道基本概念就不难理解，后面会不定期更新，希望博友们多留言补充。

二、贝叶斯（ Bayes）公式

通常把事件 A 的概率 P(A) 叫做实验前的假设概率，即先验概率 (prior probability) ，如果有另一个事件 B 与事件 A 有某种关系，即事件 A 和 B 不是互相独立的，那么当事件 B 确实发生之后，则应当重新估计事件 A 的概率，即 P(A | B)，这叫做条件概率或者试验后的假设概率，即后验概率 (posterior probability).

公式一：

$$P(A|B) = \frac{P(AB)}{P(B)}$$

再引入全概率公式： 设事件 A 当前仅当互不相容的事件 （即任意两个事件不可能同时发生的） B_i (i = 1, 2 , ... n) 中的任意一个事件发生时才可能发生，已知事件 B_i 的概率 $P(B_i)$ 及事件 A 在 B_i 已发生的条件下的条件概率，则事件 A 发生的概率为：

$$P(A) = \sum_{i=1}^{n} P(B_i)P(A|B_i)$$

这就是全概率公式 .

根据概率乘法定理：

$$P(AB) = P(A)P(B|A) = P(B)P(A|B)$$

我们可以得到：

$$P(A)P(B_i|A) = P(B_i)P(A|B_i)$$

于是：

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{P(A)}$$

再根据上面介绍的全概率公式，则可得到传说中的 贝叶斯公式 ：

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n} P(B_j)P(A|B_j)}$$

这些公式定理几乎贯穿整个机器学习，很基本，也很重要！

三、常用的离散随见变量分布

1. “ 0-1 ”分布 ": 设随机变量 X 只能取得两个数值： 0 与 1，而 概率函数 是：

$$f(x) = p^x q^{1-x}, x = 0, 1;$$

通常把这种分布叫做“ 0-1 ”分布或者两点分布，p 是分布参数 .

2. 二项分布 (binomial distribution) ： 设随机变量 X 可能的的值是 0, 1, 2, ..., n, 而概率函数是：

$$f(x) = C_n^x p^x q^{n-x}, x = 0, 1, 2, \dots, n,$$

其中 $0 < p < 1, p + q = 1.$，这种分布叫做二项分布，含有两个参数 n 和 p，通常把这种分布记作 $B(n, p)$，如果随见变量 X 服从二项分布 $B(n, p)$，记作 $X \sim B(n, p).$

3. 泊松 (Possion) 分布： 设随机变量 X 的可能值是一切非负整数，而 概率函数 是：

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}, x = 0, 1, 2, \dots$$

其中 λ > 0 为常数， 这种分布叫做泊松分布。 泊松分布就含有一个参数 λ ，记作 P(λ,)

如果随机变量 X 服从泊松分布，则记作 X~P(λ)

四、随机变量的分布函数

设 x 是任何实数，考虑随机变量 X 取得的值不大于 x 的概率，即事件 X ≤ x 的概率，记作 F(x) = P(X ≤ x) 这个函数叫做随机变量 X 的概率分布函数或者分布函数 ，注意区别于上面讲到的 概率函数 .

如果已知随机变量 X 的分布函数 F(X)， 则随见变量 X 落在半开区间 (x_1, x_2] 内的概率：

P(x_1 < X ≤ x_2) = F(x_2) - F(x_1)

五、连续随机变量的概率密度

连续随机变量的概率密度就是分布函数的导函数

六、随机变量的数学期望

如果随机变量 X 只能取得有限个值：

$$x_1, x_2, \dots, x_n$$

而取得有限个值得概率分别是：

$$p(x_1), p(x_2), \dots, p(x_n)$$

则数学期望：

$$E(X) = x_1 p(x_1) + x_2 p(x_2) + \cdots + x_n p(x_n)$$
$$= \sum_{i=1}^{n} x_i p(x_i)$$

如果连续随机变量 X 的概率密度为 $f(x)$，则连续随机变量的数学期望：

$$E(X) = \int_{-\infty}^{+\infty} x f(x) \mathrm{d}x$$

一个常数的的数学期望等于这个常数本身。

定理：两个独立随机变量的乘积的数学期望等于它们数学期望的乘积。证明如下：

对于 离散 随机变量 X 与 Y 独立：

$$E(XY) = \sum_i \sum_j x_i y_j p(x_i, y_j)$$
$$= \sum_i \sum_j x_i y_j p_X(x_i) p_Y(y_j)$$
$$= \sum_i x_i p_X(x_i) \sum_j y_j p_Y(y_j)$$
$$= E(X)E(Y)$$

对于 连续 随机变量 X 与 Y 独立：

$$E(XY)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}xyf(x,y)\,dx\,dy$$

$$=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}xyf_X(x)f_Y(y)\,dx\,dy$$

$$=\int_{-\infty}^{+\infty}xf_X(x)\,dx\int_{-\infty}^{+\infty}yf_Y(y)\,dy$$

$$=E(X)E(Y)$$

七、方差与标准差

随机变量　　X　　的方差记作　　D(X)，定义为：

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

下面证明一个很有用的公式　　(会用到性质：　一个常数的的数学期望等于这个常数本身　　)：

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

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

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

$$=E(X^2)-2E(X)E(X)+[E(X)]^2$$

$$=E(X^2)-[E(X)]^2$$

简而言之：随机变量的方差等于变量平方的期望减去期望的平方　　.

标准差就是方差的算术平方根。

常数的方差为　0.

八、协方差与相关系数

随机变量　　X　　与　随机变量　　Y　　的协方差记作　：

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

进一步推导可得：

$$cov(X,Y)=E(XY)-E(X)E(Y)$$

因为两个独立随机变量乘积的期望等于两个随机变量各自期望的乘积,　　于是当两个随机变量独立使，很容易得到它们的协方差为　　0.

两个随机变量　　X　　与　Y　　的　相关系数为：

$$R(X,Y)=\frac{cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

两个随机变量的相关系数的绝对值不大于 1.

当且仅当随机变量 Y 与 X 之间存在线性关系：

$$Y=a+bX$$

时，相关系数 $R(X,Y)$ 的绝对值等于 1，并且

$$R(X,Y)=\begin{cases}-1, & b<0\\ 1, & b>0\end{cases}$$

九、正态分布

正态分布又叫高斯分布，设连续随机变量 X 的概率密度

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x-\mu}{2\sigma^2}},-\infty<x<+\infty$$

其中 μ 及 σ>0 都是常数，这种分布就是正态分布 .

正态分布含有两个参数 μ 及 σ>0 其中μ等于正态分布的数学期望， 而 σ 等于正态分布的标准差 ，通常把这种分布记作 $N(\mu,\sigma^2)$，随机变量 X 服从正态分布 $N(\mu,\sigma^2)$，则记为：

$$X\sim N(\mu,\sigma^2)$$

定理 设随机变量 X 服从正态分布 $N(\mu,\sigma^2)$，则 X 的线性函数 Y= a + bX(b ≠0)也服从正态分布，且有

$$Y=a+bX\sim N(a+b\mu,b^2\sigma^2)$$

先总结这么多，以后遇到重要的概率论知识点会继续补充！