生成模型必然考虑：

$$P(c|x)=\frac{P(x,c)}{P(x)}$$

$$P(c|x)=\frac{P(c)\,P(x|c)}{P(x)}$$

$P(c)$：样本空间中各类样本所占的比例

极大似然估计：源自频率主义学派，根据数据条样来估计概率分布参数.

$$P(D_c|\theta_c)=\prod_{x\in D_c}P(x|\theta_c)$$ （使用θ_c找到D_c中所有样本的概率）

在所有θ_c的可能取值中，找到一个使$P(D_c|\theta_c)$最大的值.

$$LL(\theta_c)=\log P(D_c|\theta_c)=\sum_{x\in D_c}\log P(x|\theta_c)$$

$$p(x|c)\sim N(\mu_c,\sigma_c^2)$$

11

朴素贝叶斯分类器

采用"属性条件独立性假设"：对已知类别，假设所有属性相互独立，每个属性独立地对分类结果发生影响.

$$P(c|x)=\frac{P(c)\,P(x|c)}{P(x)}$$

x=青绿，大，西瓜

x_1=青绿

x_2=大

x_3=西瓜

$$=\frac{P(c)}{P(x)}\prod_{i=1}^{d}P(x_i|c)$$

$$P(x|c)=P(x_1|c)\,P(x_2|c)\,P(x_3|c)$$

$$=P(x_1,x_2,x_3|c)$$

朴素贝叶斯判定准则：

$$h_{nb}(x)=\arg\max_{c\in y}P(c)\prod_{i=1}^{d}P(x_i|c)$$

基于训练集D，估计类先验概率$P(c)$

$$P(c)=\frac{|D_c|}{|D|}$$

D_{c,x_i}表示D_c中在第i个属性上取值为x_i的样本组成的集合.

$$P(x_i|c)=\frac{|D_{c,x_i}|}{}$$