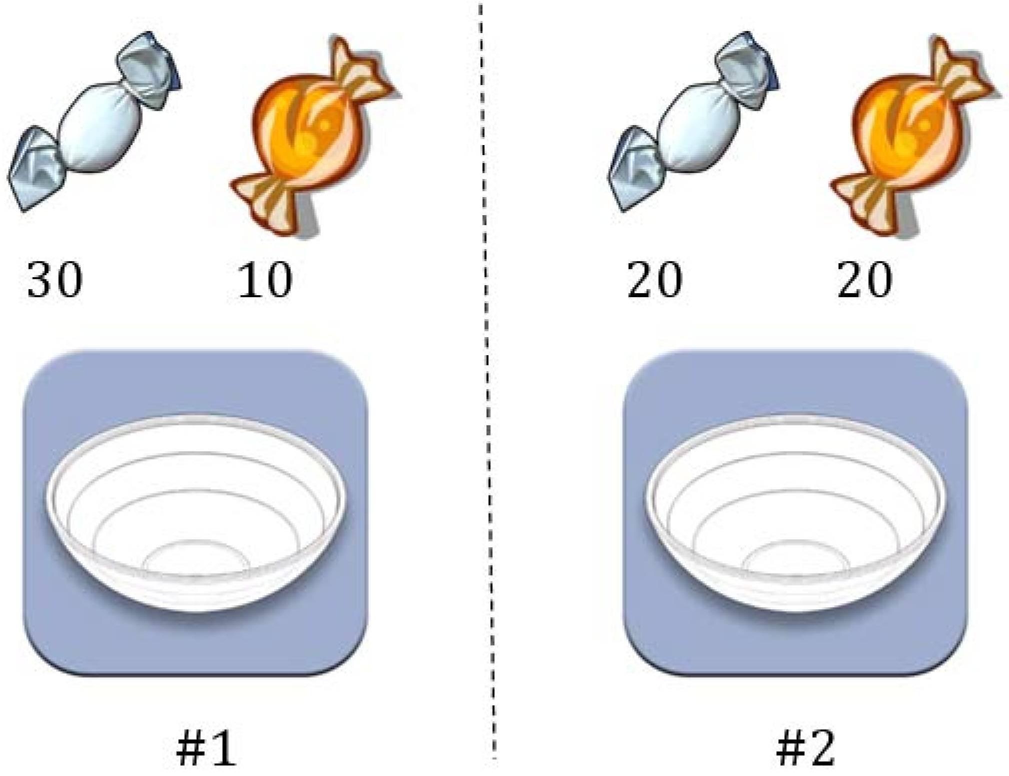
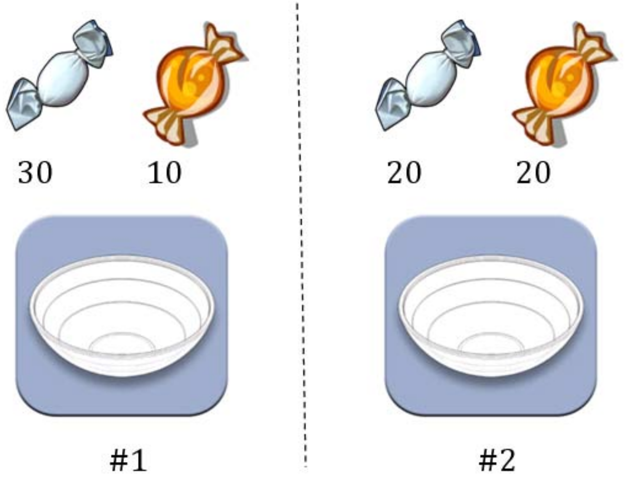
朴素贝叶斯算法

L先生AI课堂

贝叶斯公式的应用

- 有两个碗，第一个碗中装有30个水果糖和10个巧克力糖，第二个碗中装有20个水果糖和20个巧克力糖
- 现在随机选择一个碗，从中取出一颗糖，发现是水果糖，请问这颗糖属于哪个碗？





X1(水果糖)	X2(巧克力糖)	个数	Y(属于哪个碗)
1	0	30	1
1	0	20	2
0	1	10	1
0	1	20	2

x=1➔y=?

y=max(p(y=1|x=1，x2=0),p(y=2|x=1,x2=0))

贝叶斯公式的应用单个特征到多个特征

$$P(A_j|B)=\frac{P(A_jB)}{P(\mathrm{B})}=\frac{P(A_j)*P(\mathrm{B}|A_j)}{P(B)}=\frac{P(A_j)*P(\mathrm{B}|A_j)}{\mathop{?}\limits_{i=1}^{n}P(\mathrm{B}|\mathrm{A}_i)*P(\mathrm{A}_i)}$$

对应给定的样本X的特征向量x_1,x_2,...,x_m；该样本X的类别y的概率可以由贝叶斯公式得到：

$$P(y|x_1,x_2,\ldots x_m)=\frac{P(y)P(x_1,x_2,\ldots x_m|y)}{P(x_1,x_2,\ldots x_m)}$$

朴素贝叶斯算法

- 朴素贝叶斯(Naive Bayes， NB)是基于“**特征之间是独立的**”这一朴素假设，应用贝叶斯定理的监督学习算法。

朴素贝叶斯算法推导

- **特征属性之间是独立的**，所以得到：

$$P(x_i \mid y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m) = P(x_i \mid y)$$

- 公式优化得到对于多个特征：

$$P(y \mid x_1, x_2, \ldots x_m) = \frac{P(y)P(x_1, x_2, \ldots x_m \mid y)}{P(x_1, x_2, \ldots x_m)} = \frac{P(y)\prod_{i=1}^{m} P(x_i \mid y)}{P(x_1, x_2, \ldots x_m)}$$

- 推导公式类比：

根据贝叶斯定理

$$P(x_1, x_2, x_3 \mid y) = \frac{P(x_1, x_2, x_3, y)}{P(y)} = \frac{\mathrm{P}(x_1 \mid y)\square\mathrm{P}(x_2 \mid y)\,\mathrm{P}(x_3 \mid y)?P(y)}{P(y)} = \overset{3}{\underset{i=1}{?}} P(x_i \mid y)$$

x=1➔y=?

y=max(p(y=1|x=1，x2=0),p(y=2|x=1,x2=0))

朴素贝叶斯算法

$$P(y \mid x_1, x_2, \dots x_m) = \frac{P(y)P(x_1, x_2, \dots x_m \mid y)}{P(x_1, x_2, \dots x_m)} = \frac{P(y)\prod_{i=1}^{m} P(x_i \mid y)}{P(x_1, x_2, \dots x_m)}$$

在给定样本的情况下，$P(x_1,x_2,\dots,x_m)$是常数，所以得到：

从而：

$$P(y \mid x_1, x_2, \dots x_m) \propto P(y)\prod_{i=1}^{m} P(x_i \mid y)$$

$$\hat{y} = \arg\max_{y} P(y) ?_{i=1}^{m} P(x_i \mid y)$$

朴素贝叶斯算法流程

- 朴素贝叶斯算法流程/定义如下:
  - 设$x=\{x_1,x_2,\ldots,x_m\}$为待分类样本，其中x_i为x的一个特征属性
  - 类别集合为$C=\{y_1,y_2,\ldots,y_n\}$
  - 分别计算$P(y_1|x),P(y_2|x),\ldots.,P(y_n|x)$的值（贝叶斯公式）
  - 如果$P(y_k|x)=max\{P(y_1|x),P(y_2|x),\ldots.,P(y_n|x)\}$,那么认为x为$y_k$类型