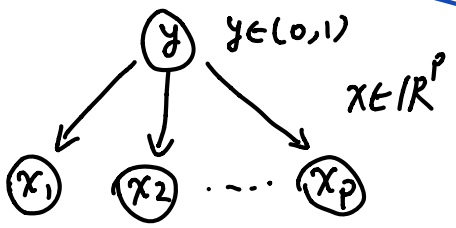
朴素贝叶斯 (Naive Bayes)

1. Introduction

↗最简单的概率图模型

- 思想：朴素贝叶斯假设(条件独立性假设)



→ 目的：为了简化运算. 相似的还有 Markov chain

$x_1 \to x_2 \to x_3 \cdots \to x_n$

$x_1 \perp x_3 \mid x_2$. 即节点只与其前一个节点有关.

↳ $x_i \perp x_j \mid y \quad (i \neq j)$

$$p(x|Y) = \prod_{j=1}^{p} p(x_j|Y)$$

2. 数学推导

问题: $Data = \{(x_i, y_i)\}_{i=1}^{N}$, $x_i \in R^p$. $y \in \{0,1\}$ (分类问题)

给定 x, y?

$$\hat{y} = \arg\max_y P(y|x)$$

$$= \arg\max_y \frac{P(x,y)}{P(x)}$$

$$= \arg\max \frac{P(x|y)\cdot P(y)}{P(x)} \propto P(x|y)\cdot P(y) \quad ①$$

由 Naive Bayes 的条件独立性假设, 可得

$$P(x|Y) = \prod_{i=1}^{p} P(x_i|Y) \quad ②.$$

联立①②两式,

得 $$\hat{y} = \arg\max_y P(y) \cdot \prod_{i=1}^{p} P(x_i|Y)$$

$$= \arg\max_y P(y) \cdot P(x_1|y) \cdot P(x_2|y) \cdots P(x_p|y)$$

此时 $P(y)$ 与 $P(x_i|y)$ 均可以通过对样本的直接统计得到

Exsmple

序号	驾龄	平均车速	性别
1	1	60	男
2	2	80	男
3	3	80	男
4	2	80	男
5	1	40	男
6	2	40	女
7	1	40	女
8	1	40	女
9	3	60	女
10	3	80	女

此时，当我们拿到一个样本，此人驾龄 2 年，平均车速 80，我们来用朴素贝叶斯分类器来推测此人的性别：

那么，此时分类 $y = \{男，女\}$，样本有两个特征，其中 x_1 表示驾龄，x_2 表示平均车速。

那么，我们来计算不同分类下的 $p(y)p(x_1|y)p(x_2|y)$，来看哪个大。

$$p(y = 男)p(x_1 = 2|y = 男)p(x_2 = 80|y = 男) = 0.5 * 0.4 * 0.6 = 0.12$$

$$p(y = 女)p(x_1 = 2|y = 女)p(x_2 = 80|y = 女) = 0.5 * 0.2 * 0.2 = 0.02$$

因此我们选择使得似然函数取值最大的 y 值，显然，我们推测此人为男性。