# Machine Learning

## Naive Bayes

### 1. Basics

**Bayes' theorem**:

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

**independence assumptions**:

$$\begin{aligned} P(X|Y) &= P(X_1, X_2|Y) \\ &= P(X_1|X_2, Y)P(X_2|Y) \\ &= P(X_1|Y)P(X_2|Y) \end{aligned}$$

where $X = \langle X_1, X_2 \rangle$ .

**Normal distribution**:

$$p(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

### 2. Model

**input**:

$$(x^{(1)}, y^{(1)}), \cdots, (x^{(m)}, y^{(m)}) \text{ where } x^{(i)} \in \mathbb{R}^n, y^{(i)} \in \{1, \ldots, k\}$$

**hypothesis**:

$$p(x|y) = \prod_{i=1}^{n} p(x_i|y)$$

$$
\begin{aligned}
\hat{y} &= \underset{y}{\operatorname{argmax}}\, p(y|\hat{x}) = \underset{y}{\operatorname{argmax}} \frac{p(\hat{x}|y)p(y)}{p(\hat{x})} \\
&= \underset{y}{\operatorname{argmax}}\, p(\hat{x}|y)p(y) = \underset{y}{\operatorname{argmax}}\, p(y)\prod_{i=1}^{n} p(\hat{x}_i|y)
\end{aligned}
$$

**parameter estimation(discrete)**:

$$
\begin{aligned}
p(\hat{x}_j|y) &= \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = \hat{x}_j \wedge y^{(i)} = y\}}{\sum_{i=1}^{m} 1\{y^{(i)} = y\}} \\
p(y) &= \frac{\sum_{i=1}^{m} 1\{y^{(i)} = y\}}{m}
\end{aligned}
$$

**smooth(discrete)**:

$$
\begin{aligned}
p(\hat{x}_j|y) &= \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = \hat{x}_j \wedge y^{(i)} = y\} + l}{\sum_{i=1}^{m} 1\{y^{(i)} = y\} + l \cdot \operatorname{distinct}\{x_j^{(1)}, \ldots, x_j^{(m)}\}} \\
p(y) &= \frac{\sum_{i=1}^{m} 1\{y^{(i)} = y\} + l}{m + lk}
\end{aligned}
$$

**parameter estimation(continuous)**:

$$
\begin{aligned}
p(\hat{x}_j|y) &= \frac{1}{\sigma_{jy}\sqrt{2\pi}} e^{-\frac{(\hat{x}_j - \mu_{jy})^2}{2\sigma_{jy}^2}} \\
\mu_{jy} &= \frac{\sum_{i=1}^{m} x_j^{(i)} \cdot 1\{y^{(i)} = y\}}{\sum_{i=1}^{m} 1\{y^{(i)} = y\}} \\
\sigma_{jy} &= \frac{\sum_{i=1}^{m} (x_j^{(i)} - \mu_{jy})^2 \cdot 1\{y^{(i)} = y\}}{\sum_{i=1}^{m} 1\{y^{(i)} = y\}} \\
p(y) &= \frac{\sum_{i=1}^{m} 1\{y^{(i)} = y\}}{m}
\end{aligned}
$$

**unbiased estimation**:

$$
\sigma_{jy} = \frac{\sum_{i=1}^{m} (x_j^{(i)} - \mu_{jy})^2 \cdot 1\{y^{(i)} = y\}}{(\sum_{i=1}^{m} 1\{y^{(i)} = y\}) - 1}
$$

**output**:

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, p(y|\hat{x}) = \underset{y}{\operatorname{argmax}}\, p(y) \prod_{i=1}^{n} p(\hat{x}_i|y)$$