1. Наивный байес и центроидный классификатор

Вероятность того, что элемент $x = (x_1, \dots, x_n)$ принадлежит классу y:

$$P(y|x_1,\dots,x_n) = \frac{P(y)P(x_1,\dots,x_n)}{P(x_1,\dots,x_n)} \overset{\text{предп. незав.}}{=} \frac{P(y)\prod\limits_{k=1}^{k=n} P(x_i|y)}{\prod\limits_{k=1}^{k=n} P(x_i)}$$

.

Найдем наиболее вероятный класс для данного x:

$$\hat{y} = \arg\max_y \frac{P(y)\prod\limits_{k=1}^{k=n} P(x_i|y)}{\prod\limits_{k=1}^{k=n} P(x_i)} \overset{\text{без констант}}{=} \arg\max_y P(y)\prod_{k=1}^{k=n} P(x_i|y) \overset{P(y)\text{ - константа}}{=} \arg\max_y \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-\sum\limits_{k=1}^{k=n}(x_k-\mu_{yk})^2}{2\sigma^2}} =$$

$$\arg\max_y e^{-\sum\limits_{k=1}^{k=n}(x_k-\mu_{yk})^2} = \arg\min_y \sum_{k=1}^{k=n}(x_k-\mu_{yk})^2$$

Легко заметить, что это то же самое, что евклидовое расстояние в квадрате между x и μ_y. Таким образом, классификация сводится к отнесению объекта x к классу y, центр которого μ_y ближе всего к x, что и требовалось доказать.

2. ROC-AUC случайных ответов

Пусть n – число объектов в выборке, k – число объектов первого класса. Тогда:

$\frac{TP}{n} = P(x \in \mathbf{1}) \cdot P(a(x) = 1) = \frac{k}{n}p$

$\frac{FP}{n} = P(x \notin \mathbf{1}) \cdot P(a(x) = 1) = \frac{n-k}{n}p$

$\frac{TN}{n} = P(x \notin \mathbf{1}) \cdot P(a(x) = 0) = \frac{n-k}{n}(1-p)$

$\frac{FN}{n} = P(x \in \mathbf{1}) \cdot P(a(x) = 1) = \frac{k}{n}(1-p)$

$TPR = \frac{k}{n}p/(\frac{k}{n}p + \frac{k}{n}(1-p)) = p$

$FPR = \frac{n-k}{n}p/(\frac{n-k}{n}p + \frac{n-k}{n}(1-p)) = p$

Таким образом, TRP = FPR не зависимо от p и k.

3. Ошибка 1NN и оптимального байесовского классификатора

$E_N = P(y \neq y_n) = P(y = 1, y_n = 0) + P(y = 0, y_n = 1) \overset{\text{независ.}}{=} P(y = 1)P(y_n = 0) + P(y = 0)P(y_n = 1) \to 2P(y = 1)P(y = 0) = 2P(1|x)P(0|x) \geq \min\{P(1|x), P(0|x)\}$,
так как $P(0|x) + P(1|x) = 1$, т.е. $2\max\{P(1|x), P(0|x)\} \geq 1$