Наивный байес и центроидный классификатор

Наивный байесовый классификатор равен
$a(X) = argmax_y P(y|X) = argmax_y P(X|y)P(y)$

По условию, априорная вероятность $P(y) = const$, а $P(x^{(k)}|y)$ -- плотность распределения $N(\mu_{y,k}, \sigma^2)$, следовательно,

$$a(X) = argmax_y \Pi_{i=1}^n P(x^{(i)}|y)$$

Что влечет

$$a(X) = argmin_y \sum_{i=1}^n (x^{(i)} - \mu_{y,k})^2$$

Таким образом, $a(x)$ является минимизатором расстояния (в смысле L_2 -нормы) от X до $\mu_y = (\mu_{y,1}, \dots, \mu_{y,n})^T$, что и требовалось показать.

ROC-AUC случайных ответов

ROC задается парами точек

$$(\eta_1, \eta_2) = (\frac{\sum I(y_i = 0, a_\omega(x_i) = 1)}{\sum I(y_i = 0)}; \frac{\sum I(y_i = 1, a_\omega(x_i) = 1)}{\sum I(y_i = 1)}),$$

где $a_\omega(x_i) = I(\xi - \omega > 0)$, $\xi \sim Bin(1, p)$. Зафиксируем y и найдем мат. ожидание (η_1, η_2) для $\omega \in [-1; 1]$

$$\mathrm{E}(\eta_1, \eta_2) = \mathrm{E}(\frac{\sum I(y_i = 0, a_\omega(x_i) = 1)}{\sum I(y_i = 0)}; \frac{\sum I(y_i = 1, a_\omega(x_i) = 1)}{\sum I(y_i = 1)}) =$$

$$= (\frac{\sum_{i=1}^m P(a_\omega(x_i = 1)}{m}; \frac{\sum_{i=1}^{n-m} P(a_\omega(x_i = 1)}{n-m}) = (p, p)$$

Предпоследнее равенство верно в силу независимости $a(x_i)$. Таким образом, все значения ROC будут на диагонали и $\mathrm{E}[ROC - AUC] = 0.5$

Ошибка 1NN и оптимального байесовского классификатора

См. фото

In [ ]: