ДЗ №1

Даниил Дмитриев, 494

27 февраля 2017 г.

1 Задание 4.1

Для наивного байесовского классификатора (с учетом того, что априорные вероятности классов одинаковы):

$$P(y|X) \propto \prod_{i=1}^{n} P\left(x^{(k)}|y\right)$$

По условию

$$P\left(x^{(k)}|y\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(x^{(k)}-\mu_{yk}\right)^2}{2\sigma^2}}$$

Следовательно, получаем, что

$$P(y|X) \propto exp\left(-\sum_{k=0}^{n} \frac{\left(x^{(k)}-\mu_{yk}\right)^2}{2\sigma^2}\right)$$

То есть $P(y|X)$ максимально, когда $\sum_{k=0}^{n} \left(x^{(k)}-\mu_{yk}\right)^2$ минимально. А это как раз формула квадрата евклидового расстояния между μ_y и X.

2 Задание 4.2

В треугольном ROC-AUC у нас всего 3 порога: 0, 0.5 и 1. Давайте найдём соответствующие точки на графике. В первом и последнем случае это, очевидно, (0, 0) и (1, 1) соответственно. Для порога 0.5 вначале выпишем значения tp, tn, fp и fn.

Пусть α - доля класса "1"в выборке, а n - размер выборки. Тогда tp в среднем равно $n\alpha p$, так как всего "1"у нас $n\alpha$ и с вероятностью p мы даём правильный ответ. Аналогично,
tn $= n(1-\alpha)(1-p)$, fp $= n(1-\alpha)p$, fn $= n\alpha(1-p)$

По формуле для fpr и tpr:

$$fpr = \frac{fp}{fp+tn} = \frac{n(1-\alpha)p}{n(1-\alpha)p + n(1-\alpha)(1-p)} = p$$

$$tpr = \frac{tp}{tp + fn} = \frac{n\alpha p}{n\alpha p + n\alpha(1-p)} = p$$

То есть точка соответствующая порогу 0.5 - (p, p) вне зависимости от α. Так как эта точка лежит на диагонали, то ROC-AUC будет равен в среднем 0.5

3 Задание 4.3

(Разбирался на семинаре 494 группы)

$E_B = min\{P(0|X), P(1|X)\}$
$E_N = P(y \neq y_n) = P(y_n = 1|x_n)P(0|x) + P(y_n = 0|x_n)P(1|x)$

Мы предполагаем, что $P(y|x)$ непрерывна по x. Пусть l - размер выборки. Тогда $P(y_n|x_n) \to P(y_n|x)$ при $l \to \inf$. Значит

$$E_N \approx 2P(1|x)P(0|x) \leq 2min\{P(0|x), P(1|x)\} = 2E_B$$