Домашнее задание по машинному обучению №1

Котельникова Дарья 491 группа

27.02.2017

1 Наивный байес и центроидный классификатор

Если в наивном байесовском классификаторе классы имеют одинаковые априорные вероятности, а плотность распределения признаков в каждом классе имеют вид $P(x^{(k)}|y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x^{(k)}-\mu_{yk})^2}{2\sigma^2}}$, где $x^{(k)}, k = 1, \cdots, n$ - признаки объекта x, тогда классификация сводится к отнесению объекта x к классу y, центр которого μ_y ближе всего к x.

Для доказательства максимизируем $argmax_y \prod_{k=1}^{n} P(x_i|y) =$

$$\prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x^k-\mu_{yk})^2}{2\sigma^2}}$$

Отсюда, видно, что необходимо максимизировать сумму:

$$e^{-\sum_{k=1}^{n}(x^k-\mu_{ky})^2}$$

$\sum_{k=1}^{n}(x^k - \mu_{ky})^2 = \rho(\bar{x}, \mu_y)$
Относим $\bar{x}$ к классу y, а это то же, что и минимизировать расстояние $\rho(\bar{x}, \mu_y)$ (экспонента максимальна, когда расстояние будет минимально), которое минимально, когда $\bar{x}$ ближе всего к центру распредления μ_y.

2 ROC-AUC случайных ответов

Покажем, что "треугольный $ROC - AUC$"в случае, когда классификатор дает случайные ответы - $a(x) = 1$ с вероятностью p и $a(x) = 0$ с вероятностью $1 - p$, будет в среднем равен 0.5, независимо от p и доли класса 1 в обучающей выборке.
Проверяем это условие в среднем, значит на графике должна получиться прямая $y = x$.
За x - обозначим $TruePositiveRate(TPR)$, а за y - $FalsePositiveRate(FPR)$, от которых и строится ROC-кривая.
На графике для подсчета $ROC - AUC$ используем формулу:

$$S = \frac{1}{2}xy + (1-x)y + \frac{1}{2}(1-x)(1-y)$$

$$S = \frac{1}{2} + \frac{1}{2}(x-y)$$

$$\Rightarrow ES = \frac{1}{2} + \frac{1}{2}E(x-y)$$

γ - доля объектов из traindataset, которая имеет класс 1. Пусть $\gamma n = k$

$$Ex = \frac{TP}{TP + FN} = \frac{pk}{pk + p(1-k)} \frac{pk}{pk + (1-p)k} = p$$

$$Ey = \frac{FP}{FP + TN} = \frac{p(n-k)}{p(n-k) + (1-p)(n-k)} = p$$

$$ES = \frac{1}{2} + \frac{1}{2}(Ex - Ey) = \frac{1}{2}$$

$$Q.E.D$$

3 Ошибка 1NN и оптимального байесовского классификатора

$E_N -$ усреднениепообучающимнаборам. $P(0|x) = p_\pi$ и $P(1|x) = 1 - p_\pi$ - непрерывно по x. Байесовский классификатор в точке $x \longrightarrow 0$, если

$$P(0|x) > P(1|x)$$

Байесовский классификатор в точке $x \longrightarrow 1$, если

$$P(0|x) < P(1|x)$$

$E_B(x) = min(P(0|x), P(1|x))$ - вероятность ошибки байесовского классификатора в точке x.
$(x_n, y_n) \in T, (x, y)$ - независимые в T случайные величины с распределением π в точке пространства $X \times Y$.
$P(y \neq y_n) = P(y = 0, y_n = 1) + P(y = 1, y_n = 0) = P(0|x)P(1|x_n) + P(1|x)P(0|x_n) = 2P(0|x)P(1|x)$ (т.к плотности распределений непрерывны и $x \approx x_n$)
$\Rightarrow E_N(x) = P(y \neq y_n) < 2P(0|x)P(1|x) = 2(1 - E_B(x))E_B(x)$
Предположим, что $E_B(x) \leq E_n(x)$ при $N \longrightarrow \infty$ (по свойству непрерывности вероятности)

$$E_B \leq 2(1 - E_B)E_B \leq 2E_B$$