Задание 1. Теория

Александр Захаров, 494

27 февраля 2017 г.

1 Задача 4.1 Наивный байес и центроидный классификатор

Байесовский классификатор работает по принципу: $\hat{y} = argmax_y P(y|x)$. При этом в наивном байесовском классификаторе $P(y|x) = \frac{P(x|y)\cdot P(y)}{P(x)} = \frac{P(y)\cdot \prod\limits_{k=1}^{n} P(x^{(k)}|y)}{P(x)}$. Так как априорные веротности классов совпадают, то $P(y)$ совпадают для всех классов. Выбор y также не зависит от $P(x)$, поэтому $argmax_y P(y|x) = argmax_y \prod\limits_{k=1}^{n} P(x^{(k)}|y)$.

$\prod\limits_{k=1}^{n} P(x^{(k)}|y) = C \cdot e^{-\frac{\sum\limits_{k=1}^{n} (x^{(k)}-\mu_{yk})^2}{2\sigma^2}}$, где C - константа.

Как видно из формулы максимум будет для класса, центр которого ближе. Ведь можно считать, что минимизируется $\sum\limits_{k=1}^{n} (x^{(k)} - \mu_{yk})^2$ – расстояние до центра класса. □

2 Задача 4.2 ROC-AUC случайных ответов

Для того, чтобы в среднем ROC-AUC = 0.5 достаточно показать, что TPR = FPR для промежуточной точки "треугольного"ROC-AUC. Найдем их матожидания:

При этом принадлежность классу независима с выставелнием метки. Размер выборки n, k элементов в 1 классе.

$\mathbb{E}(TPR) = \frac{p\cdot k}{k} = p$

$\mathbb{E}(FPR) = \frac{p\cdot(n-k)}{n-k} = p$ □

3 Задача 4.3 Ошибка 1NN и оптимального байесовского классификатора

Запишем ошибку 1NN через вероятности, где x_n - ближайший сосед, тогда $P(y|x_n)$ - индикатор принадлежности x_n определенному классу.

$E_{1NN} = P(0|x)\cdot P(1|x_n) + P(1|x)\cdot P(0|x_n) \to 2\cdot P(0|x)\cdot P(1|x) \leq 2\cdot \min(P(0|x), P(1|x))$. Предельный переход осуществлен в силу непрерывности вероятности по x □