Машинное обучение.
Задание 1

Пашментов Никита

1 Задание 1

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

1.1 Решение

$\hat{y} = \arg\max_y \big(\prod_{i=1}^{n} P(x^{(i)} \mid y) P(y) \big) = \arg\max_y \big(\prod_{i=1}^{n} P(x^{(i)} \mid y) \big)$

Будем искать максимум прологарифмированной функции:

$\hat{y} = \arg\max_y \big(\sum_{i=1}^{n} \ln\left(P(x^{(i)} \mid y)\right) \big)$

Подставим плотность из условия:

$\hat{y} = \arg\max_y \big(-\frac{1}{2}\ln(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(x^{(i)}-\mu_{yi})^2}{2\sigma^2} \big) = \arg\min_y \sum_{i=1}^{n} (x^{(i)} - \mu_{yi})^2$

Нетрудно заметить, что:

$\hat{y} = \arg\min_y (\rho(x, \mu_y))$

2 Задание 3

Утверждается, что метод одного ближайшего соседа асимтотически(при условии, что максимальное по всем точкам выборки расстояние до ближайшего соседа стремится к нулю) имеет матожидание ошибки не более чем вдвое больше по сравнению с оптимальным байесовским классификатором(который это матожидание минимизирует).

Покажите это, рассмотрев задачу бинарной классификации. Достаточно рассмотреть вероятность ошибки на фиксированном объекте x, т.к. математическое ожидание ошибок на выборке размера V будет просто произведением V на эту вероятность. Байесовский классификатор ошибается на объекте x с вероятностью:

$$E_B = min\{P(1 \mid x), P(0 \mid x)\}$$

Условные вероятности будем считать непрерывными функциями от $x \in \mathbb{R}^m$, чтобы иметь возможность делать предельные переходы. Метод ближайшего соседа ошибается с вероятностью:

$$E_N = P(y \neq y_n)$$

Здесь y - настоящий класс x, а y_n - класс ближайшего соседа x_n к объекту x в предположении, что в обучающей выборке n объектов, равномерно заполняющих пространство.
Докажите исходное утверждение, выписав выражение для E_N (принадлежность к классам 0 и 1 для объектов x и x_n считать независимыми событиями) и осуществив предельный переход по n.

2.1 Решение

$E_N = P(y_n = 1 \mid x_n)P(0 \mid x) + P(y_n = 0 \mid x_n)P(1 \mid x) \simeq 2P(1 \mid x)P(0 \mid x)$ (по непрерывности $P(y \mid x)$)
$2P(1 \mid x)P(0 \mid x) \leq 2\min\{P(1 \mid x), P(0 \mid x)\}$
Следовательно:
$E_B \leq 2E_N$ - ч.т.д.