Московский физико-технический институт

Машинное обучение.

Задание 1

Студент: Ивахненко Максим Дмитриевич

Группа: 491

Долгопрудный
2017

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

Доказательство. Так как все априорные вероятности одинаковы, то наивный байесовский классификатор максимизирует по y выражение:

$\prod_{k=1}^{n} P(x^{(k)}|y) = \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x^{(k)}-\mu_{yk})^2}{2\sigma^2}} = (\prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}}) e^{-\sum_{k=1}^{n} \frac{(x^{(k)}-\mu_{yk})^2}{2\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{\frac{-\rho(x,\mu_y)}{2\sigma^2}}$,

где $\rho(x, \mu_y)$ - евклидово "расстояние"от центра класса до объекта. Чем меньше расстояние тем больше экспонента и больше условная плотность, которую максимизирует классификатор. □

Задача 2. *Покажите, что "треугольный ROC-AUC"(см.лекцию 2) 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.*

Доказательство. Пусть $n_0 + n_1 = n$, где n_0 – количество объектов класса "0 n_1 – класса "1 n – всего объектов. Тогда

$E(TPR) = \sum_{k=1}^{n_1} \frac{k}{n_1} P(tp = k) = \sum_{k=1}^{n_1} \frac{k}{n_1} C_{n_1}^{k} p^k (1-p)^{n_1-k} = \frac{n_1 p}{n_1} = p$

$E(FPR) = \sum_{k=1}^{n_0} \frac{k}{n_0} P(fp = k) = \sum_{k=1}^{n_0} \frac{k}{n_0} C_{n_0}^{k} p^k (1-p)^{n_0-k} = \frac{n_0 p}{n_0} = p$

Следовательно $E(TPR) = E(FPR)$. То есть можем считать, что в среднем TPR = FPR, поэтому площадь под ROC кривой равна площади по кривой $y = x$ на отрезке $[0; 1]$, а именно 0,5. □

Задача 3. *Утверждается, что метод одного ближайшего соседа асимптотически (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. матожидание ошибок на выборке размера V будет просто произведением V на эту вероятность. Байесовский классификатор ошибается на объекте x с вероятностью:*

$$E_B = min(P(1|x), P(0|x))$$

Условные вероятности будем считать непрерывными функциями от $x \in R^m$, 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.

Доказательство.

$$E_B = min(p(0|x), p(1|x))$$

$$P(y \neq y_i) = P(x = 1, y_i = 0) + P(x = 0, y_i = 1) = P(1|x)P(0|x_i) + P(0|x)P(1|x)$$

$$\lim_{x_i \to x} P(y \neq y_i) = 2P(1|x)P(0|x) = 2E_B(1 - E_B) \leq 2E_B$$

□