Условие 1. *Наивный байес и центроидный классификатор*
Покажите, что если в наивном байесовском классификаторе классы имеют одинаковые априорные вероятности, а плотность распределения признаков в каждом классе имеет вид
$P(x^{(k)}|y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x^{(k)}-\mu_{y_k})^2}{2\sigma^2}}$*, где x(k) , k = 1, ..., n — признаки объекта x, классификация сводится к отнесению объекта x к классу y, центр которого* μ_y *ближе всего к x*

Решение. Нужно максимизировать выражение $p(Y) \cdot \prod_{k=1}^n P(x^{(k)}|y)$. Так как априорные вероятности каждого класса равны, необходимо и достаточно минимизировать $\prod_{k=1}^n P(x^{(k)}|y)$. Последнее выражение имеет вид $C_1 \cdot e^{-\sum_{i=1}^n f(x^{(i)},y)C_2}$, где C_1, C_2 не зависят от y.

Следовательно, минимизируем $\sum_{i=1}^n f(x^{(i)}, y) = \sum_{i=1}^n (x^{(i)} - \mu_{y_i})^2$. Последняя величина в точности есть расстояние от x до центра класса y. Значит, утверждение задачи верно. □

Условие 2. *ROC-AUC случайных ответов*
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.

Решение. «треугольный ROC-AUC» равен площади, огранчинной сверху отрезками AC, BC, снизу- осью OX, где A - начало координат, B - точка с координатами (1,1), C - точка с координатами (ξ, η), где ξ- False Positive rate, η - True Negative rate.

Тогда искомая площадь равна 0,5+S, где S - ориентированная площадь трегольника ABC (с минусом, если C ниже AB, с плюсом в противном случае).
$S = 0,5 \cdot AB \cdot$ (ориентированное расстояние от C до AB).
Ориентированное расстояние - линейная функция. Значит, ориентированная площадь - линейная от C функция, равная $f(C)$. При этом $E\ f(C) = f(E\ )$ в силу линейности f. При этом A, B фиксированы. Таким образом, чтобы проверить, что «треугольный ROC-AUC» в среднем равен 0,5, необходимо и достаточно проверить, что $f(E\ ) = 0$, то есть что среднее положение точки C - на прямой AB.
Это равносильно тому, что $E\ \xi = E\ \eta$.

Последнее верно, так как обе указанных случайных величины порождаются случайными ответами, которые равны единице с верояностью p. То есть обе величины равны по распределению (они бернулииевские с параметром p), а потому имеют равные матожидания.

□

Условие 3. *Ошибка 1NN и оптимального байесовского классификатора*
Утверждается, что метод одного ближайшего соседа асимптотически (при условии, что макси- мальное 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 ве- роятность ошибки на фиксированном объекте x, т.к. матожидание ошибок на выборке раз- мера V будет просто произведением V на эту

вероятность. Байесовский классификатор оши- бается на объекте x с вероятностью: $E_B = minP(1|x), P(0|x)$ *Условные вероятности будем считать непрерывными функциями от* $x \in R^m$*, чтобы иметь возможность делать предельные переходы. Метод ближайшего соседа ошибается с вероятно- стью:* $E_N = P(y \neq y_n)$ *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*

Решение. Напишем выражение для E_N
$E_N = P(y = 1, y_n = 0) + P(y = 0, y_n = 1) =$
$=$ (так как принадлежность к классам 0 и 1 для объектов x и
x_n считаем независимыми событиями) $=$
$= P(y = 1)P(y_n = 0) + P(y = 0)P(y_n = 1) = P(1|x)P(0|x_n) + P(0|x)P(1|x_n)$

При стремлении n к бесконечности, правая часть, по условию о непрерывности условных вероятностей как функций от точки, стремится к
$P(1|x)P(0|x) + P(0|x)P(1|x) = 2P(0|x)P(1|x)$
Так как $P(1|x), P(0|x) \leq 1$, имеем $2P(0|x)P(1|x) \leq 2P(0|x)$; $2P(0|x)P(1|x) \leq 2P(1|x)$, откуда $E_N \leq 2\min(P(1|x), P(0|x)) = 2E_B$, что и требовалось показать.

□

Зайко Александр, 499 группа