Теоретические задачи из задания 1 по курсу «Машинное обучение»

Александр Князев

February 27, 2017

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

Решение.
$y_{map} = argmax_{y\in Y}(P(y)\cdot\prod_{i=1}^{n} P(x^{(k)}|y)) = argmax_{y\in Y}[log(P(y)\cdot\prod_{i=1}^{n} P(x^{(k)}|y))] = argmax_{y\in Y}[logP(y)+\sum_{i=1}^{n} logP(x^{(k)}|y)]$ = |выкидываем отсюда logP(y) так как по условию все априорные вероятности равны| $= argmax_{y\in Y}[\sum_{i=1}^{n} logP(x^{(k)}|y)] = argmax_{y\in Y}\sum_{i=1}^{n}[-\frac{(x^{(k)}-\mu_y)^2}{2\sigma^2}] = argmin_{y\in Y}\sum_{i=1}^{n}[\frac{(x^{(k)}-\mu_y)^2}{2\sigma^2}] = argmin_{y\in Y}\sum_{i=1}^{n}(x^{(k)}-\mu_y)^2$
Последнее выражению и обозначает, что объект x будет отнесен к тому классу, центр которого ближе всего к нему.

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.*

Решение.
Первый вариант решения.
$a(x) = I[b(x) > t]$, где t - порог, $b(x)$ - какая-то функция, ставящая в соответствие объекту вероятность принятия ею значения 1, а $a(x)$ - наше предсказание.
Так как ответы случайны, то значение $b(x)$ - не зависит от x.
Вспомним, что по смыслу площадь под кривой ROC-AUC равна вероятности того, что случайно выбранный объекта класса 1 получит оценку принадлежности к классу 1 (то есть значение $b(x)$) выше, чем случайно выбранный объект класса 0.
В силу случайности ответов, для двух произвольных объектов вероятность, что значение $b(x)$ у первого больше, чем это значение у второго, равна вероятности, что значение $b(x)$ у второго больше, чем это значение у первого. То есть эти вероятности равны 0.5.
Таким образов, из смысла площади под кривой, мы понимаем, что в среднем она будет равна 0.5. И не зависит от p и доли класса 1 в обучающей выборке

Второй вариант.
Треугольный ROC-AUC определяется одной точкой. Найдем положение этой точки. Для этого нужно определить $\frac{TPR}{FPR}$.
Пусть n - доля элементов первого класса. Так как ответы случайны, то
TPR $= \frac{TP}{TP+FN} = \frac{\text{верные срабатывания}}{\text{размер класса 1}} = \frac{n\cdot p}{n} = p$.
FPR $= \frac{FP}{FP+TN} = \frac{\text{ложные срабатывания}}{\text{размер класса 0}} = \frac{(1-n)\cdot p}{1-n} = p$.
То есть $\frac{TPR}{FPR} = 1$, точка лежит на диагонали, площадь под кривой равна в среднем 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.

Решение.
Расписываем, используя условие, что принадлежность к классам 0 и 1 для объектов x и x_n независимые события.
$E_N = P(y \neq y_n) = P(y = 0, y_n = 1) + P(y = 1, y_n = 0) = P(0|x) \cdot P(1|x_n) + P(1|x) \cdot P(0|x_n)$
При $n \to \infty$ $x_n \to x$, и в силу непрерывности получаем:
$E_N = P(y \neq y_n) \approx P(0|x) \cdot P(1|x) + P(1|x) \cdot P(0|x) = 2P(0|x) \cdot P(1|x)$
Используя, что $E_B = min(P(1|x), P(0|x))$, получаем что:
$E_N \approx 2P(0|x) \cdot P(1|x) = 2(1 - E_B)(E_B) \leq 2E_B$.
Что и требовалось доказать.