Задание 1: теоретическая часть

DMIA

21 апреля 2021

1 2NN

Может ли в методе k ближайших соседей при $k = 2$ получиться лучший результат, чем при $k = 1$? Отказы от классификации тоже считать ошибками.

2 Критерий Gini

Обычно в конкретном листе решающее дерево отвечает тем классом, который в в этом листе преобладает. Рассмотрим другую стратегию: посчитаем доли классов p_0 и p_1 (для простоты рассмотрим бинарную классификацию) в листе и при попадании неизвестного объекта в этот лист будем отвечать 0 с вероятностью p_0 и 1 с вероятностью p_1. Как критерий Gini связан с вероятностью верного ответа на объекте при такой стратегии ответа в листе?

3 Наивный байес и ближайший центроид

Как наивный байесовский классификатор с гауссовским распределением связан с методом ближайшего центроида (см. лекцию), если дисперсии всех признаков во всех классах одинаковые, а матожидание оценивается по выборке с помощью оценки максимального правдоподобия?

4 *Несимметричный модуль

Вспомним выкладки с анализом потерь вида $|a(x) - y|$ с помощью байесовского взгляда на задачу регрессии (из приложения к первой лекции). Попробуйте провести те же выкладки для потерь

$$\alpha \cdot (a(x) - y)[a(x) > y] + (1 - \alpha) \cdot (y - a(x))[a(x) < y]$$

К оценке какого параметра распределения $P(y|x)$ приведет минимизация таких потерь?

5 *Неудачный выбор функции потерь

В одном проекте заказчик очень хотел, чтобы исследователь решал не задачу классификации на классы 0 и 1 (которая и стояла), а задачу регрессии на тех же метках с модулем отклонения в качестве функции потерь (т.е. оптимизировал на обучающей выборке средний модуль отклонения ответа от меток 0 и 1). Замысел заказчика был в том, что оцененные числа получатся в интервале (0, 1), и это будет приближением для вероятности класса 1. Покажите, что если алгоритм минимизирует матожидание потерь на объекте при условии известного объекта x ($R(a(x), x)$ в обозначениях из приложения к первой лекции), то затея приведет к тому, что в ответах будут только 0 и 1.

6 **Связь ошибки 1NN и оптимального байесовского классификатора

Утверждается, что метод одного ближайшего соседа асимптотически (при стремлении плотности точек из обучающей выборки к бесконечности) 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.