

Домашнее задание №1

Жуков Кирилл, Б13-201

25 сентября 2023 г.

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.

Решение: прогоним сначала 1NN на тестовой выборке, а потом на ней же 2NN. Допустим, точки были классифицированы 1NN с некоторой долей ошибок.

Рассмотрим новую точку, которая была правильно классифицирована 1NN. Это значит, что ближайший сосед новой точки имел такой же класс. Поскольку в случае 2NN найдется еще один ближайший сосед, то в случае, если его класс совпадает с классом новой точки, то 2NN правильно ее классифицирует, иначе будет отказ от классификации, т.е. ошибка.

Теперь рассмотрим новую точку, которую 1NN классифицировал неправильно. Это значит, что ее ближайший сосед имеет отличный от искомого класс. Тогда второй по дальности сосед, найденный в 2NN, либо предскажет так же неправильный класс, либо будет отказ от классификации.

Таким образом, получаем что результат 2NN всегда будет не лучше 1NN.

2 Критерий Gini

Обычно в конкретном листе решающее дерево отвечает тем классом, который 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?

Решение: пусть p_i - вероятность выбора точки в i -й из n классов (нумерация начинается с 0). Тогда коэффициент Джинни G записывается в следующем виде:

$$G = \sum_{i=0}^{n-1} p_i(1 - p_i) = \sum_{i=0}^{n-1} p_i - \sum_{i=0}^{n-1} p_i^2 = 1 - \sum_{i=0}^{n-1} p_i^2$$

Найдем вероятность верного ответа P для нашей стратегии. Неизвестный объект имеет i -й класс с вероятностью p_i . При этом мы ответим ему правильно с такой же вероятностью p_i . Поэтому вероятность правильного ответа для объекта i -го класса равна p_i^2 . Значит, для объекта неизвестного класса вероятность правильного ответа равна

$$P = \sum_{i=0}^{n-1} p_i^2$$

Таким образом,

$$G = 1 - P \quad \Rightarrow \quad P = 1 - G$$

3 Решающее правило в листе

Какая стратегия поведения в листьях решающего дерева приводит к меньшей вероятности ошибки при многоклассовой классификации: отвечать тот класс, который преобладает в листе, или отвечать случайно с тем же распределением классов, что и в листе?

Решение: пусть p_i - доля i -го из n классов в листе (она же есть вероятность выбора i -го класса).

Тогда в первой стратегии вероятность правильного ответа P_1 будет равна

$$P_1 = \max\{p_i\}_{i=1}^n = p_m$$

В то же время вероятность правильного ответа P_2 во второй стратегии из решения задачи №2 равна

$$P_2 = \sum_{i=1}^n p_i^2$$

Сравним вероятности двух стратегий P_1 и P_2 :

$$\begin{aligned} P_1 - P_2 &= p_m - \sum_{i=1}^n p_i^2 = p_m - p_m^2 - \sum_{\substack{i=1 \\ i \neq m}}^n p_i^2 = \\ &= p_m(1 - p_m) - \sum_{\substack{i=1 \\ i \neq m}}^n p_i^2 = p_m \sum_{\substack{i=1 \\ i \neq m}}^n p_i - \sum_{\substack{i=1 \\ i \neq m}}^n p_i^2 \geq 0 \end{aligned}$$

Таким образом, мы получили, что $P_1 \geq P_2$. Значит, первая стратегия приводит к меньшей вероятности ошибки.

4 *Наивный байес и ближайший центроид

В лекции о простых методах машинного обучения мы не упомянули метод ближайшего центроида. Устроен он очень просто: считаем среднее арифметическое векторов признаков объектов в каждом классе и получаем «центроиды» классов, а далее при классификации относим объект к тому классу, центроид которого оказался ближе.

Покажите, как наивный байесовский классификатор с гауссовским распределением связан с методом ближайшего центроида, если дисперсии всех признаков во всех классах одинаковые, а матожидание оценивается по выборке с помощью оценки максимального правдоподобия?

Решение:

5 Матричное дифференцирование

Покажите справедливость следующих выражений (x и b - столбцы, A - матрица):

1. $\frac{\partial}{\partial x} x^2 = 2x$
2. $\frac{\partial}{\partial x} (Ax + b)^T (Ax + b) = 2A^T (Ax + b)$
3. $\frac{\partial}{\partial A} \ln |A| = A^{-1}$

Решение:

1. Определим $x^2 := \langle x, x \rangle$.

$$\begin{aligned} [D_{x_0} \langle x, x \rangle](h) &= \langle [D_{x_0} x](h), x \rangle + \langle x, [D_{x_0} x](h) \rangle \\ &= 2\langle x, [D_{x_0} x](h) \rangle = \langle 2x, h \rangle \end{aligned}$$

Значит, $\frac{\partial}{\partial x} x^2 = 2x$.

2. Определим $(Ax + b)^T (Ax + b) := \langle Ax + b, Ax + b \rangle$.

$$\begin{aligned} [D_{x_0} \langle Ax + b, Ax + b \rangle](h) &= \langle [D_{x_0} (Ax + b)](h), Ax_0 + b \rangle + \\ &\quad + \langle Ax_0 + b, [D_{x_0} (Ax + b)](h) \rangle \\ &= 2\langle Ax_0 + b, Ah \rangle \\ &= 2\langle Ax_0 + b, Ah \rangle = 2\langle A^T (Ax_0 + b), h \rangle \\ &= \langle 2A^T (Ax_0 + b), h \rangle \end{aligned}$$

Значит, $\frac{\partial}{\partial x} (Ax + b)^T (Ax + b) = 2A^T (Ax + b)$.

3. Воспользуемся формулой производной сложной функции

$$[D_{x_0} u \circ v](H) = [D_{v(x_0)} u]([D_{x_0} v](H))$$

$$\begin{aligned} [D_{X_0} \ln \circ \det](H) &= \frac{1}{\det(X_0)} \cdot \langle \det(X_0) \cdot X_0^{-T}, H \rangle \\ &= \langle \frac{1}{\det(X_0)} \cdot \det(X_0) \cdot X_0^{-T}, H \rangle = \langle X_0^{-T}, H \rangle \end{aligned}$$

Значит, $\frac{\partial}{\partial A} \ln |A| = A^{-1}$.

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 .

Решение: