Метрические методы. Метрики качества
10 баллов. +2 бонусных балла

Задача 1. (1 балл)

Оценим время работы алгоритма ближайших соседей по количеству операций. Пусть X — обучающая выборка размера n, Y — тестовая выборка размера m. Размерность признакового пространства d.
Таким образом, $X \in \mathbb{R}^{n \times d}$, а $Y \in \mathbb{R}^{m \times d}$.
Квадрат евклидова расстояния между объектами x_i и y_j записывается как:

$$\rho(x_i, y_j) = \sum_{k=1}^{d} (x_i^k - y_j^k)^2.$$

- Определите количество операций, необходимое для подсчета всех попарных расстояний в наивном случае.
- Предложите способ, с помощью которого можно уменьшить количество операций. Оцените количество операций для предложенного метода.

Решение

Наивный метод

Вычитание - 1 опреация, возведение в квадрат - ещё одна операция, всего d слагаемых. То есть на вычисление $\rho(x_i, y_j)$ требуется $2d$ операций. Так как всего mn пар, то количествоо операций будет равными:

$$N = 2dnm.$$

Ускоренный вариант

Раскроем квадраты:

$$\rho(x_i, y_j)^2 = \sum_{k=1}^{d} (x_i^k)^2 + \sum_{k=1}^{d} (y_j^k)^2 - 2\sum_{k=1}^{d} x_i^k y_j^k$$

То есть:

$$\rho(x_i, y_j)^2 = \|x_i\|^2 + \|y_j\|^2 - 2(x_i \cdot y_j)$$

Для каждого x_i и y_j квадрат нормы можно вычислить предварительно квадраты норм, это займёт суммарно $(m+n)d$ операций, вычисление скалярного произведения также, как и в наивном варианте, занимает $O(mnd)$ операций, но его можно вычислить как $X^T Y$, что на пркатике позволяет использовать векторизованные вычисления и ускроить процесс.

Задача 2. (2 балла)

Дано n объектов, распределённых равномерно внутри d-мерного единичного шара с центром в нуле.

- Найдите выражение для медианы расстояния от начала координат до ближайшего объекта.
- Проинтерпретируйте полученный результат в терминах применимости метода ближайшего соседа в различных ситуациях.

Считайте, что метрика в задаче евклидова.
Указание: попробуйте смоделировать событие и посчитать его вероятность в терминах функций распределения.

Решение

Нахождения медианы

Объём шара в R^d:

$$V(r) = \frac{\pi^{n/2}}{\Gamma(n/2+1)} r^d$$

Так как объекты распределены в шаре радиуса 1 равномерно, то вероятность нахождения объекста внутри шара радиуса $r \in [0;1]$, являющаяся функцией распределения случайной величины $\rho = \|X\|$ - расстояния от объекта X до центра, равна:

$$\mathbb{P}(X \in B_0(r)) = \frac{V(r)}{V(1)} = r^d = \mathbb{P}(\rho < r) = F_\rho(r)$$

Упорядочим выборку $X_1, ..., X_n \to X_{(1)}, ..., X_{(n)}$ так, чтобы $\rho_{(1)} \leq \rho_{(2)} \leq ... \leq \rho_{(n)}$.

$F_{\rho_{(1)}}(r) = \mathbb{P}(\rho_{(1)} \leq r) = 1 - \mathbb{P}$(все точки на расстоянии больше r) $= 1 - (1-r^d)^n$

Медиана удовлетворяет уравнению $F_{\rho_{(1)}}(m) = 1 - (1-m^d)^n = \frac{1}{2}$
Отсюда

$$m = (1 - (\frac{1}{2})^{1/n})^{1/d}$$

Интерпретация

Найдём пределы при фиксированных n и d.

$$\lim_{d \to \infty} m = 1$$

$$\lim_{n \to \infty} m = 0$$

Получаем, что при большой размерности, все точки будут сконцентрированны у поверхности и будет почти невозможно найти действительно близкого соседа, что демонстрирует проклятье размерности. При этом можно его избежать, но для этого потребуется количество данных, растущее экспоненциально по отношению с d.

Задача 3. (3 балла)

Решается задача классификации с помощью алгоритма ближайших соседей (метрика евклидова). Для тестового объекта z ближайшим соседом с расстоянием ρ_x является x, вторым ближайшим соседом с расстоянием ρ_y является объект y. Остальные объекты обучающей выборки находятся от z на достаточно большом расстоянии.

Ко всем объектам добавляется новый признак: для z и y значение признака распределёно равномерно на отрезке $[-1, 1]$, для всех остальных объектов значение признака равно нулю.

- Посчитайте вероятность того, что теперь ближайшим соседом для z будет не x, а y.
- Проинтерпретируйте полученный результат в терминах применимости метода ближайших соседей.

Указание: возможно в этой задаче пригодится знание криволинейных интегралов.

Решение

Нахождение вероятности

Обозначим за ρ'_x и ρ'_y новые расстояния от z до x и y, а через p и q новый признак для z и y соответственно. Тогда

$$(\rho')^2_x = \rho^2_x + p^2$$

$$(\rho')^2_y = \rho^2_y + (p-q)^2$$

Условие того, что новое расстояние до y меньше:

$$\rho^2_y + (p-q)^2 < \rho^2_x + p^2$$

$$q^2 - 2pq < \rho^2_x - \rho^2_y$$

$$2pq - q^2 > \rho^2_y - \rho^2_x = D > 0$$

Найдём искомую вероятность $\mathbb{P}(2pq - q^2 > D)$. Знаем, что

$$f_p(x) = f_q(x) = \frac{1}{2}\mathbb{I}\{x \in [-1; 1]\}$$

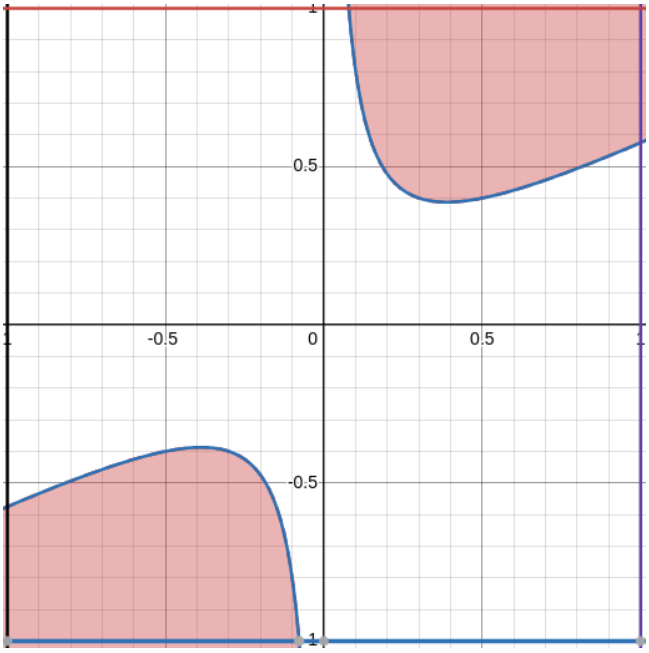
тогда

$$f_{p,q}(x) = f_p(x)f_q(x) = \frac{1}{4}$$

Найдём $F_\xi(D)$:

$$\mathbb{P}(2pq - q^2 > D) = \iint\limits_{\{(p,a):2pq-q^2>D\}} \frac{1}{4}\mathbb{I}\{-1 \le x \le 1\}dpdq$$

Задача сводится к тому, чтобы найти площадь пересечения множества (на



картинке закрашено красным, оси в виде (q, p)), по которому происходит интегрирование, и единичного квадрата. Заметим, что если $D > 1$, то пересечение пустое и искомая вероятность равна 0. Найдём точки пересечения графика с квадратом в первой координатной четверти - это $(1 - \sqrt{1-D}; 1)$ и $(1; \frac{D+1}{2})$. Тогда

$$\iint\limits_{\{(p,a):2pq-q^2>D\}} \frac{1}{4}\mathbb{I}\{-1 \le x \le 1\}dpdq = 2\cdot\frac{1}{4}\int\limits_{1-\sqrt{1-D}}^{1} dq \int\limits_{\frac{q}{2}+\frac{D}{2q}}^{1} dp = \int\limits_{1-\sqrt{1-D}}^{1} (1-\frac{q}{2}-\frac{D}{2q})dq$$

$$= \frac{1}{4}(\sqrt{1-D} + 1 - d + D\ln(1-\sqrt{1-D})$$

Подытожим:

$$\mathbb{P}(y \text{ стал ближе } x) = \begin{cases} \frac{1}{4}(\sqrt{1-(\rho_y^2-\rho_x^2)} + 1 - (\rho_y^2-\rho_x^2)(1-\ln(1-\sqrt{1-(\rho_y^2-\rho_x^2)}))), \rho_y^2-\rho_x^2 \leq 1 \\ 0, (\rho_y^2-\rho_x^2) > 1 \end{cases}$$

Интерпретация

Порядок ближайших соседей может быть изменён даже из-за случайного признака, поэтому для данного метоа важен отбор наиболее релевантных признаков.

Задача 4. (1 балл)

Докажите, что ROC-AUC случайного классификатора равен 0.5.

Решение

Ясно, что при случайной классификации, вероятности TP, TN, FP и FN одинаковы и равны (TPR и FPR также будут равны) , так как ROC-AUC - это площадь под графиком $TPR(FPR)$, который в данном случае представляет собой диагональ единичного квадрата, то значение этой метрики будет равно $\frac{1}{2}$.

Задача 5. (2 балла)

Пусть, $a = a(x)$ ответ алгоритма. На сколько может уменьшиться ROC-AUC при использовании функции $\min(a, 0.5)$ над оценками алгоритма?

Решение

Все объекты с оценками $> 0,5$ получат оценку $0,5$, а оценка объектов с исходной $\leq 0,5$ не изменится. ROC-AUC можно интерпретировать как вероятность того, что случайный положительный объект будет оценён выше, чем случайный отрицательный.

Приведём пример для изменения на $0,5$:
Пусть всех положительные объекты имеют оценки, больше 0,5, а отрицательные - 0,5, тогда изначально ROC-AUC = 1, после преобразования все объекты имеют одинаковые метки и ROC-AUC = 0,5.
Заметим, что если изначально ROC-AUC $\leq$ 0,5, то он точно не сможет вырасти, так как количество пар, в которых у положительного объекта оценка будет выше, чем у отрицательного не увеличится. Если же 0,5 $<$ ROC-AUC $<$ 1, то после преобразования ROC-AUC не может стать меньше 0,5, так как соотншения между оценками TP и TN, а так же FP и FN не изменятся, а у всех TP и FP будут одинаковые метки. Значит в таком случае ROC-AUC изменится меньше, чем на 0,5.

Ответ: 0,5.

Задача 6. (3 балла)

Подробнее ознакомьтесь с материалом по ROC-AUC по ссылке и решите следующую задачу:
Пусть на ответах алгоритма m (принимающих значения от 0 до 1) задано распределение объектов класса 1 (доля объектов класса 1 в зависимости от ответа алгоритма) следующим образом:

$$\mathbb{P}(m \in [a, b] \mid y = 1) = \int_a^b p(z)dz.$$

Распределение объектов класса 0 задаётся так:

$$\mathbb{P}(m \in [a, b] \mid y = 0) = \int_a^b (2 - p(z))dz,$$

где $p(z) = -1.5z^2 + 3z$.
Найдите вероятностные оценки на величины TPR, FPR и ROC-AUC.

Решение

$$TPR(x) = \mathbb{P}(m > x \mid y = 1) = \int_x^1 (-1,5^2 + 3z)dz = 0,5x^3 - 1,5x^2 + 1.$$

$$FPR(x) = \mathbb{P}(m > x \mid y = 0) = \int_x^1 (2{+}1,5z^3{-}3z))dz = -0,5x^3{+}1,5x^2{-}2x{+}1,$$

$$ROC{-}AUC = \int_1^0 TPR(x)FPR'(x)dx = \int_0^1 (0,5x^3{-}1,5x^2{+}1)(1,5x^2{-}3x{+}2)dx = 0,75$$