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.

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.

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.

Задача 4. (1 балл)

Докажите, что ROC-AUC случайного классификатора равен 0.5.

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?

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.