

Теоретическое задание 1

Математика в машинном обучении и метрические методы

курс «Машинное обучение 1», 2020

Базовая стоимость задания — 10 баллов, максимальная — 12 баллов. Для выполнения базовой части достаточно решить любое число задач, дающих в сумме 10 баллов.

Задачи

1. (2 балла) Дано n объектов распределённых равномерно по d -мерному единичному шару с центром в нуле. Найдите выражение для медианы расстояния от начала координат до ближайшего объекта. Метрика — евклидова. Проинтерпретируйте полученный результат в терминах применимости метода ближайшего соседа в различных ситуациях.

2. (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 распределено 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.

3. (3 балла) В двумерном пространстве рассматриваются два вектора с углом Θ между ними. Случайно равномерно выбирается прямая, проходящая через начало координат.

- С какой вероятностью оба вектора окажутся в одном полупространстве относительно прямой?
- С какой вероятностью оба вектора окажутся в разных полупространствах относительно прямой?
- Проинтерпретируйте результат с точки зрения использования семейства $F = \{f_\omega(x) = \text{sign}(\langle \omega, x \rangle)\}$ в качестве LSH для косинусного расстояния.

4. (2 балла) Пусть A произвольная не нулевая матрица, докажите, что если вектор v является собственным для матрицы $A^T A$ с с.з равным σ^2 , то есть $A^T A v = \sigma^2 v$, то вектор $\frac{1}{\sigma} A v$ является собственным для матрицы $A A^T$.

5. (2 балла) Есть ряд из вещественных чисел (числа могут быть отрицательными), каждое число означает выручку маркетплейса Ozon в определенный день. При условии, что известна выручка в каждый из предыдущих N дней:

- предложите формулу для оценки выручки на следующий день, используя метод максимального правдоподобия (ММП)
- предложите формулу для оценки дисперсии выручки, используя ММП