Задачи 1,3,4

Задача 1 | Дано: n объектов, d-мерный единичный шар с центром в нуле.

Найти: выражение для медианы расстояния от начала координат до ближайшего объекта.

Решение: Медиана – точка, в которой вероятность принимает значение $\frac{1}{2}$.
Посчитаем $1/2$ квантиль распределения $x_1, \dots, x_n$, где x_i – расстояние до n-ого объекта.
Зафиксируем расстояние t.
Вероятность, что объект попадёт в шар радиуса t, равна t^d (*).
$P(x<t) = t^d$, а вероятность, что не попадёт: $P(x>t) = 1-t^d$
Расстояние до ближайшего объекта – это первая порядковая статистика $x_{(1)}$.
$P(x_{(1)}<t) = 1 - P(x_{(1)}>t) = 1 - (P(x>t))^n$ (т.е. если $x_{(1)} > t$, то и все значения $x > x_{(1)}$ тоже больше t, поэтому вероятность в степени n. $(x_{(1)} \le x_{(2)} \le \dots \le x_{(n)})$)
Тогда
$P(x_{(1)}<t) = 1-(1-t^d)^n$
Приравниваем к $1/2$, чтобы определить выражение для медианы.

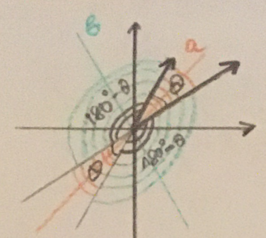
$$1-(1-t^d)^n = \frac{1}{2}$$
$$(1-t^d)^n = \frac{1}{2}$$
$$(1-t^d) = \left(\frac{1}{2}\right)^{\frac{1}{n}}$$
$$t^d = 1-\left(\frac{1}{2}\right)^{\frac{1}{n}}$$
$$\boxed{t = \left(1-(1/2)^{1/n}\right)^{1/d}}$$

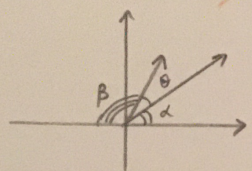
(*) Объём d-мерного шара радиуса $r=1$ равен $V_{dr} = \frac{\pi^{d/2} r^d}{\Gamma(1+d/2)}$, для радиуса t: $V_{dt} = \frac{\pi^{d/2} t^d}{\Gamma(1+d/2)}$, тогда вероятность, что объект попадёт в шар радиуса t при возможном максимальном радиусе r находим как отношение подходящих нам случаев ко всем случаям:
$\frac{V_{dt}}{V_{dr}}$ = [одинаковые множители сокращаются] $= \frac{t^d}{r^d} = \left(\frac{t}{r}\right)^d = \left(\frac{t}{1}\right)^d = t^d$

Рассмотрим полученный результат в терминах применимости метода ближайшего соседа.
При фиксированном d с ростом n расстояние до ближайшего объекта от начала координат уменьшается.
При фиксированном n с ростом d расстояние до ближайшего объекта увеличивается, причём уже для $n=100$ и $d=8$ это расстояние превышает 0,5 (около 0,54), то есть большинство объектов (т.к. мы рассматривали медиану) находятся ближе к границе нашего шара, чем к любому другому объекту. Это создаёт проблема для прогнозирования в случае knn, по сути, нам приходится экстраполировать предсказание для новой удалённой точки из соседних точек, а не интерполировать между ними. Выборка больше разреженная, чем плотная, так как с ростом d необходимо увеличивать

n экспоненциально. Так, если для $d=1$ и $n=100$ плотность пропорциональна $n^{1/d}=100$, то для случая с 10 переменными такой плотности можно достичь с уже $\sqrt[10]{n}=100 \Rightarrow n=100^{10}$ объектов. Это феномен проклятья размерности, и для метода ближайшего соседа он играет большую роль.

Задача 3





$\alpha+\beta+\theta = 180°$

$\alpha+\beta = 180° - \theta$

1) Вероятность того, что оба вектора окажутся в одном полупространстве относительно прямой

Чтобы такое произошло, необходимо, чтобы прямая попала в сектор бирюзового цвета. Так как прямая выбирается случайно равномерно, то определим вероятность как отношение всех исходов к подходящим исходам:

π – число всех исходов, всех возможных углов между выбранной прямой и осью координат.

$180° - \theta = \pi - \theta$ – число всех исходов, для которых оба вектора окажутся по одну сторону от прямой (в одном полупространстве относительно прямой)

Тогда вероятность будет равна

$$\boxed{\frac{\pi-\theta}{\pi}, \text{ или } 1-\frac{\theta}{\pi}}$$

Стоит отметить, что для угла $\theta > 180°$, можно перейти к углу $360° - \theta$. Математически это будет верно, т.к. $\cos(360° - \theta) = \cos\theta$.
Практически это означает, что мы просто перенумеровали векторы, сделав 1й вектор 2м, а 2й – 1м согласно симметричности меры.

2) Вероятность, что оба вектора окажутся в разных полупространствах относительно прямой, это сектор оранжевого цвета.

Аналогично с предыдущим заданием определяем вероятность как отношение всех исходов к подходящим:

Число всех исходов остаётся тем же и равно π, а число подходящих исходов (таких углов между выбранной прямой и осью координат, что прямая лежит между векторами) равно θ.

Тогда вероятность будет равна $\boxed{\frac{\theta}{\pi}, \text{ или } 1-\left(1-\frac{\theta}{\pi}\right)}$

3) При использовании семейства $F = \{f_\omega(x) = \text{sign}(\langle \omega, x \rangle)\}$ в качестве LSH для косинусного расстояния, где ω – нормальный вектор плоскости, проходящей через начало координат, а f_ω показывает, в каком полупространстве (вместе с ω и напротив) оказался рассматриваемый вектор x.

В таком случае, используя вероятности из п.1 и п.2, получаем, что чем меньше угол θ между векторами (исходными, из условия задачи), тем больше вероятность, что векторы окажутся в одном полупространстве, т.е. похожие векторы вероятнее будут иметь одинаковый хэш. (Это также хорошо иллюстрируется примером с $\theta = 0°$, когда f_ω будет одинаковой для двух векторов). Для противоположных векторов с $\theta = 180°$ вероятность, что f_ω будет иметь разные знаки, равна $\frac{\theta}{\pi} = 1$.

Задача 4 | Доказать, что $AA^T \cdot \left(\frac{1}{\sigma} Av\right) = \lambda \cdot \left(\frac{1}{\sigma} \cdot Av\right)$, где λ - с.з.

Дано: $A^T A v = \sigma^2 v$

Решение: $AA^T \cdot \left(\frac{1}{\sigma} Av\right) =$ [вынесли константу] $= \frac{1}{\sigma} AA^T A v =$ [замена из Дано] $=$

$= \frac{1}{\sigma} A \cdot \sigma^2 v =$ [выносим σ^2] $= \sigma^2 \cdot \left(\frac{1}{\sigma} Av\right)$

Т.о. $\underline{AA^T \cdot \left(\frac{1}{\sigma} Av\right) = \sigma^2 \cdot \left(\frac{1}{\sigma} Av\right)}$, $\lambda = \sigma^2$