Задача 1

Реализовать генератор матрциц, который должен поддерживать функции:

- Генерация абсолютно случайной матрицы $n \times m$
- Генерация случайной диагональной матрицы $n \times n$
- Генерация случайной верхнетреугольной матрицы
- Генерация случайной нижнетреугольной матрицы
- Генерация симметричной матрицы
- Генерация вырожденной матрицы
- Генерация матрицы ступенчатого вида $n \times n$ ранга m
- Генерация возмущения матрицы $n \times m$, каждый элемент которой не превосходит по модулю заданный ε

Оценить вероятность того, что созданная матрица будет вырожденной.

Оценить величину нормы матрицы возмущений в зависимости от параметра ε (оценить верхнюю границу).

Задача 2

Используя ряд Маклорена, реализовать вычисление основных элементарных функций:

- Экспонента
- Натуральный логарифм
- Синус
- Косинус
- Тангенс
- Котангенс
- Арксинус
- Арккосинус
- Арктангенс
- Гиперболический синус
- Гиперболический косинус
- Гиперболический тангенс
- Гиперболический арктангенс

Оценить величину машинного эпсилон. Предложить модификации для некоторых функций и сравнить полученные результаты.

Задача 3

Реализовать вычисление трех основных норм векторов (L1, L2 и кубическую) и подчиненных им матричных норм. Реализовать вычисление числа обусловленности.

Примечание: для вычисления собственных значений можно использовать linalg.eigvals из модуля scipy.

Задача 4

Реализовать метод Гаусса приведения матрицы к ступенчатому виду. Реализовать функцию вычисления ранга матрицы. Сгенерировать вырожденные матрицы различных рангов и размеров и проверить алгоритм.

Задача 5

Реализовать метод Гаусса решения СЛАУ. Использовать данный метод для решения систем различных размеров. Оценить скорость работы метода Гаусса (необходимое количество операций) в зависимости от размера системы.

Задача 6

Сгенерировать СЛАУ (размер матрицы должен быть не менее 50×50). Решить СЛАУ методом Гаусса для различных возмущений столбца свободных членов. Оценить число обусловленности, используя полученные результаты. Вычислить число обусловленности и сравнить с численными оценками.

Дополнительные задачи

Задача 7

В этой задаче требуется найти аналитическое решение и проверить его с

помощью вычислений на Python. Решить только один пример (на выбор).

Примеры решения подобных задач есть в документе "Визуализация данных" к занятию А1.

1.1. Чему равна погрешность в определении действительного корня $x = 1$ уравнения $ax^4 + bx^3 + dx + e = 0$, если $a = 1 \pm 10^{-3}$, $b = 1 \pm 10^{-3}$, $d = -1 \pm 10^{-3}$, $e = -1 \pm 10^{-3}$?

1.2. Чему равна погрешность в определении корней уравнения $ax^3 + bx^2 = 0$, если $a = 1 \pm 10^{-3}$, $b = -4 \pm 10^{-3}$?

1.3. С каким числом верных знаков (или относительной погрешностью) должен быть известен свободный член в уравнении $x^2 - 2x + 0.999993751 = 0$, чтобы корни имели четыре верных знака?

1.4. С каким числом верных знаков (или относительной погрешностью) должен быть известен свободный член в уравнении $x^2 - 4x + 3.999901 = 0$, чтобы корни имели четыре верных знака?

1.5. Определить оптимальный шаг $h = \text{const}$ формулы численного дифференцирования $f'(x-h) \approx \left(f(x) - f(x-h)\right)/h$, $\max\limits_{[x-h,x]} |f''(x)| \le 100$, если абсолютная погрешность при задании $f(x)$, $f(x-h)$ не превосходит $\Delta = 0.1$.

1.6. Определить оптимальный шаг $h = \text{const}$ формулы численного дифференцирования $f'(x) \approx \left(f(x+h) - f(x-h)\right)/2h$, $\max\limits_{[x-h,x+h]} |f'''(x)| \le 100$, если абсолютная погрешность при задании, $f(x \pm h)$ не превосходит $\Delta = 0.1$.

1.7. Определить оптимальный шаг $h = \text{const}$ формулы численного дифференцирования $f'(x) \approx \left(3f(x) - 4f(x-h) + f(x-2h)\right)/2h$, $\max\limits_{[x-2h,x]} \left|f'''(x)\right| \le 100$, если абсолютная погрешность при задании $f(x)$, $f(x-h)$, $f(x-2h)$ не превосходит $\Delta = 0.1$.

1.8. Пусть приближенное значение первой производной функции $f(x)$ определяется при $h \ll 1$ по формуле $f'(x) \approx \left(3f(x) - 4f(x-h) + f(x-2h)\right)/2h$, а сами значения $f(x), f(x-h), f(x-2h)$ вычисляются с абсолютной погрешностью Δ. Какую погрешность можно ожидать при вычислении производной, если $\left|f^{(k)}(x)\right| \le M_k$, $k = 1,2,...$?

1.9. Пусть задана последовательность чисел x_n, $n = 0,\ 1,\ 2,\ ...,$ причем $x_{n+1} - 5x_n = 4$, а x_0 известно с относительной погрешностью 10^{-6}. При каких значениях x_0 относительная погрешность при вычислении x_n будет быстро возрастать с ростом n?

1.10. Пусть задана последовательность чисел x_n, $n = 0,\ 1,\ 2,\ ...,$ причем $5x_{n+1} - x_n = 4$, а x_0 известно с относительной погрешностью 10^{-6}. При каких значениях x_0 относительная погрешность при вычислении x_n будет быстро возрастать с ростом n?

Задача 8

Выбор метрики (нормы разницы между любыми двумя векторами, или функции расстояния между любой парой точек) очень важен для многих алгоритмов машинного обучения. Рассмотрим на примере задачи кластеризации.

Кластеризация — это разделение множества входных векторов на группы (кластеры) по степени «схожести» друг с другом.

Кластеризация в Data Mining приобретает ценность тогда, когда она выступает одним из этапов анализа данных, построения законченного аналитического решения. Аналитику часто легче выделить группы схожих объектов, изучить их особенности и построить для каждой группы отдельную модель, чем создавать одну общую модель для всех данных. Таким приемом постоянно пользуются в маркетинге, выделяя группы клиентов, покупателей, товаров и разрабатывая для каждой из них отдельную стратегию.

Евклидова метрика

— наиболее распространенная. Она является геометрическим расстоянием в многомерном пространстве.

Квадрат евклидовой метрики.

Иногда может возникнуть желание возвести в квадрат стандартное евклидово расстояние, чтобы придать большие веса более отдаленным друг от друга объектам.

Метрика городских кварталов (манхэттенская).

Это расстояние является суммой модулей разностей координат. В большинстве случаев эта метрика приводит к таким же результатам, как и для обычного расстояния Евклида. Однако отметим, что для этой меры влияние отдельных больших разностей (выбросов) уменьшается (так как они не возводятся в квадрат).

Расстояние Чебышева.

Это метрика шахматной доски (Расстоянием Чебышёва между n-мерными числовыми векторами называется максимум модуля разности компонент этих векторов). Это расстояние может оказаться полезным, когда желают определить два объекта как «различные», если они различаются по какой-либо одной координате (каким-либо одним измерением).

Расстояние Чебышёва называют также метрикой Чебышёва, равномерной метрикой, sup-метрикой и бокс-метрикой; также иногда она называется метрикой решётки, метрикой шахматной доски, метрикой хода короля и 8-метрикой.

Степенная метрика.

Иногда желают прогрессивно увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются. Это может быть достигнуто с использованием степенного расстояния.

Выбор метрики (критерия схожести) лежит полностью на исследователе. При выборе различных мер результаты кластеризации могут существенно отличаться.

Алгоритм k-means (k-средних)

Наиболее простой, но в то же время достаточно неточный метод кластеризации в классической реализации. Он разбивает множество элементов векторного пространства на заранее известное число кластеров k. Действие алгоритма таково, что он стремится минимизировать среднеквадратичное отклонение на точках каждого кластера. Основная идея заключается в том, что на каждой итерации перевычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике. Алгоритм завершается, когда на какой-то итерации не происходит изменения кластеров.

Проблемы алгоритма k-means:

- необходимо заранее знать количество кластеров. Мной было предложено метод определения количества кластеров, который основывался на нахождении кластеров, распределенных по некоему закону (в моем случае все сводилось к нормальному закону). После этого выполнялся классический алгоритм k-means, который давал более точные результаты.
- алгоритм очень чувствителен к выбору начальных центров кластеров. Классический вариант подразумевает случайный выбор класторов, что очень часто являлось источником погрешности. Как вариант решения, необходимо проводить исследования объекта для более точного определения центров начальных кластеров. В моем случае на начальном этапе предлагается принимать в качестве центов самые отдаленные точки кластеров.
- не справляется с задачей, когда объект принадлежит к разным кластерам в равной степени или не принадлежит ни одному.

Нечеткий алгоритм кластеризации c-means

С последней проблемой k-means успешно справляется алгоритм c-means. Вместо однозначного ответа на вопрос к какому кластеру относится объект, он определяет вероятность того, что объект принадлежит к тому или иному кластеру. Таким образом, утверждение «объект А принадлежит к кластеру 1 с вероятностью 90%, к кластеру 2 — 10% » верно и более удобно.

Остальные проблемы у c-means такие же, как у k-means, но они нивелируются благодаря нечеткости разбиения.

Метод нечеткой кластеризации C-средних имеет ограниченное применение из-за существенного недостатка — невозможность корректного разбиения на кластеры, в случае когда кластеры имеют различную дисперсию по различным размерностям (осям) элементов (например, кластер имеет форму эллипса). Данный недостаток устранен в алгоритмах Mixture models и GMM (Gaussian mixture models).

Документация методов кластеризации для sklearn есть здесь https://scikit-learn.org/stable/modules/clustering.html#k-means .

Используя библиотеку scikit-learn, реализуйте Gaussian mixture models и обычный k-means. Подберите такой набор данных, на котором первый метод справляется хорошо, а второй метод даёт плохие результаты, и продемонстрируйте это. Сделайте это для нескольких разных метрик и сравните результаты между собой.

https://scikit-learn.ru/example/ примеры подобного.

https://neurohive.io/ru/osnovy-data-science/vvedenie-v-scikit-learn/ введение в sklearn. На этом сайте много полезных статей и ссылок на курсы.