Лабораторная работа №5  
**«Линейная регрессия. Градиентный спуск».**

**Работа на паре «Парная регрессия»:**

**Часть 1 (1 балл).**

Загрузите из интернета данные о росте и весе 25000 подростков в возрасте 18 лет (<http://socr.ucla.edu/docs/resources/SOCR_Data/SOCR_Data_Dinov_020108_HeightsWeights.html>).

Постройте гистограммы распределения обоих параметров, рассмотрев разбиение диапазона значений на 16 интервалов.

Постройте диаграмму рассеяния (X-Y диаграмму) этих параметров.

Вычислите коэффициент корреляции Пирсона между этими параметрами.

Используя МНК найдите коэффициенты линейной регрессии веса от роста.

Оцените качество регрессионной модели, используя оценки: среднеквадратическая ошибка Q, средняя абсолютная ошибка MAE, коэффициент детерминации R2.

**Часть 2 (1 балла).**

Реализуйте методы пакетного градиентного спуска с одним параметром и стохастического градиентного спуска. Величину для оценки точности выберите сами исходя из требования наличия не менее 100 итераций.

На одном поле изобразите графики изменения значения ошибки модели на каждой итерации.

Сравните время работы обоих методов (модуль time). Сделайте выводы.

**Часть 3 (1 балл).** (дополнительно, если останется время)

Рассматривая вес как результативный признак (Y), а рост как факторный признак (Х0), сформируйте ещё 6 штук факторных признаков типа (фантазируйте!):

* Х1 = (Х0)2;
* Х2 = (Х0)1/2 ;
* Х3 = (Х0 – 100)2 ;
* Х4 = (Х0)2 – ln(Х0).

Для каждого из «новых» факторных признаков повторите расчёты по аналогии с частью 1 (фактически меняется только содержание столбца Х). Сделайте выводы.

Для получения заявленных баллов необходимо выполнить на паре расчётную часть (показав результат преподавателю, засчитывается только реализованные на паре задания) и добавить соответствующую главу в отчёт по лабораторной, где по каждому выполненному заданию привести формулировку, расчётные формулы, скриншоты результатов (графики, диаграммы, числа с пояснениями).

Баллы за полностью выполненные задания на паре дублируются. Все результаты оформляются в общем отчёте.

**Работа на дом «множественная регрессия» (5 баллов):**

Загрузите из интернета данные о стоимости недвижимости в Бостоне (файл «boston\_.csv»).

Из представленных факторов:

* CRIM : уровень преступности на душу населения
* ZN : доля жилой земли, зонированной для участков площадью более 25 000 кв.футов.
* INDUS : доля земли, не связанных с розничной торговлей, (акров на один город).
* CHAS: : фиктивная переменная (равна 1, если рядом есть река; 0 в противном случае)
* NOX : концентрация оксидов азота
* РМ : среднее количество комнат в одном жилом помещении
* AGE : доля квартир, занимаемых владельцами, построенных до 1940 года.
* DIS : взвешенные расстояния до пяти бостонских центров занятости
* RAD : показатель доступности к радиальным магистралям
* TAX : полная ставка налога на имущество за 10 000 долларов США.
* PTRATIO: соотношение учеников и учителей по городам
* B : условная величина, вычисляемая как 1000(BK — 0,63)2, где Bk - доля негров по городам
* LSTAT : уровень снижения статуса населения в процентах
* MEDV : Это средняя стоимость домов, занимаемых владельцами, в 1000 долларов США.

в качестве результативного признака (Y) взять MEDV, а все остальные 13 рассматривать как факторные (Х1 ... Х13).

Построить гистограмму распределения результативного признака разбив диапазон значений на 30 интервалов.

Построить корреляционную матрицу (14х14) связей признаков друг с другом. Попробуйте оформить её в виде тепловой карты.

Для MEDV и двух факторов с наибольшей корреляцией постройте диаграммы рассеяния (Xi - Y) (Xk - Y).

Разбить совокупность данных на тестовую выборку (100 случайно отобранных объектов) и обучающую выборку (остальные).

Построить Модель1, используя пакетный градиентный спуск. Вывести коэффициенты модели. Оценить время работы метода, итоговую точность модели, построить график изменения точности по итерациям.

Построить Модель2, используя другой вид градиентного спуска. Вывести коэффициенты модели. Оценить время работы метода, итоговую точность модели, построить график изменения точности по итерациям. Вид метода определяется остатком от деления номера студента на 3: 1 – стохастический спуск, 2 – Mini-batch, 0 – адаптивный метод (правило изменения шага определите самостоятельно).

Для построения Модели3 выполнить подготовку данных. Добавьте к имевшимся 13 факторным признакам добавьте ещё 40:

* 13 квадратов существующих признаков
* 13 квадратных корней существующих признаков
* 14 попарных произведений разных признаков (на Ваш выбор)

Из полученных 53 признаков отберите 10 в соответствиис методикой отбора существенных факторов, описанной в прилагаемом файле «Многофакторная регрессия. Выбор существенных факторов.pdf».

Построение модели выполните, используя градиентный спуск любого вида.

Результаты работы оформите в виде отчёта, где по каждой модели привести формулировку задания, расчётные формулы и правила, скриншоты результатов (графики, диаграммы, таблицы, числа с пояснениями), выводы. В приложениях привести коды построения каждой из моделей.

Реализацию моделей выполнить «ручками». Отчёт оформите код в виде юпитер-блокнота, где пошагово выполняются задания и визуализируются промежуточные и окончательные результаты.

Работу можно выполнять вдвоем.