Домашнее задание: Линейная регрессия, регуляризация и робастность (Jupyter playbook для Google Colab)

Курс: Нечеткие системы

Цель

Закрепить практикой ключевые идеи лекции: формулировку и решение линейной регрессии, поведение функций потерь (MSE/MAE), влияние мультиколлинеарности, роль регуляризации (L2/L1/Elastic Net), корректную валидацию, а также робастные методы и отличие регрессии от PCA.

Формат сдачи

- Один ноутбук **Colab** (`.ipynb`) + краткий **отчёт в самом ноутбуке** (текстовые блоки Markdown с объяснениями).
- Имя файла: `hw_linreg_SurnameName.ipynb`. В первой ячейке: ФИО, группа, дата, версия.
- Фиксируйте сид случайности (например, `random_state = 42`); указывайте версии библиотек.
- **Код в ДЗ не прикладывать сюда**. В этом документе сформулированы только задания.

Данные

Выберите **один** из вариантов (на Ваше усмотрение):

A. **Синтетика для контроля свойств.** Сгенерируйте выборку с линейной зависимостью и управляемым шумом; добавьте управляемую мультиколлинеарность (например, линейные комбинации признаков) и выбросы.

B. **Реальный регрессионный датасет.** Любой открытый набор (например, housing-признаки и цена/медианный доход) или собственные данные. Обоснуйте выбор и очистку.

Во всех разделах ниже используйте единый выбранный датасет; при необходимости создавайте копии с модификациями (коллинеарность, выбросы и т.п.).

Структура ноутбука (обязательные разделы)

1. Разведочный анализ (EDA) и подготовка

1.1 Кратко опишите задачу, целевую переменную y и признаки X; укажите масштабы и единицы измерения (если есть).

1.2 Проведите базовый EDA: распределения, корреляции, пропуски, выбросы. Сформируйте гипотезы о линейности и наличии мультиколлинеарности.

1.3 Разделите данные на **train/val/test**. Обоснуйте схему (учтите риски утечек и особенности структуры данных; для временных рядов *не* перемешивайте).

1.4 Нормализация/стандартизация признаков: покажите, как она влияет на оценку и регуляризацию.

2. Базовая линейная регрессия (OLS)

2.1 Запишите *в ноутбуке текстом* постановку задачи: $\hat{y} = X\omega$, функционал MSE и *нормальные уравнения* $X^\top X\,\omega = X^\top y$.

2.2 Получите оценку *аналитически* (через $(X^\top X)^{-1}X^\top y$ либо эквивалентные численные процедуры) и *численно* (градиентным спуском). Сравните решения и метрики на **val**.

2.3 Покажите, что при мультиколлинеарности решение нестабильно: проиллюстрируйте плохую обусловленность (например, числом обусловленности $\kappa(X^\top X)$) и варьируйте её, наблюдая рост дисперсии оценок.

2.4 Проверьте *предпосылки Гаусса–Маркова* в практическом ключе: визуализируйте остатки vs. предсказания, проверьте гомоскедастичность качественно, обсудите независимость ошибок и риски смещения.

3. MSE vs MAE и влияние выбросов

3.1 Добавьте $\sim$5–10% выбросов в целевую переменную y (или в признаки) с контролируемой амплитудой.

3.2 Обучите модели, минимизирующие MSE и MAE (допускается готовая реализация). Сравните устойчивость к выбросам: качество на **val**, поведение коэффициентов, диаграммы остатков.

3.3 Сформулируйте, когда предпочтительна MSE, когда MAE, и как это связано с распределением шума.

4. Регуляризация: Ridge, Lasso, Elastic Net

4.1 **Ridge (L2).** Подберите гиперпараметр λ по K-fold cross-validation на **train** (оценка на **val**). Сравните качество и нормы весов с OLS.

4.2 **Lasso (L1).** Постройте путь решений (зависимость коэффициентов от λ). Покажите зануление весов и сравните число ненулевых признаков с Ridge.

4.3 **Elastic Net.** Продемонстрируйте компромисс L1/L2: подберите (λ, α), сравните со стабильностью Ridge и разреженностью Lasso.

4.4 Обсудите смещение–разброс: как меняются variance/ bias при росте силы регуляризации?

5. Робастные подходы (по выбору *любой* два)

5.1 **Huber/Log-cosh** как сглаженные альтернативы: покажите, как они ведут себя при выбросах на Ваших данных.

5.2 **RANSAC-регрессия:** выделение консенсусного подмножества при загрязнениях; сравнение с OLS.

5.3 **Квантильная регрессия:** предсказание условных квантили y (например, медианной) и оценка робастности.

6. PCA vs Регрессия (мини-исследование)

6.1 Постройте PCA по признакам X и визуально объясните, чем оптимизация PCA (ортогональные расстояния до подпространства) отличается от оптимизации в регрессии (вертикальные отклонения y).

6.2 Реализуйте **PCR** (регрессия по главным компонентам) и сравните с Ridge на коллинеарных данных: качество, устойчивость, интерпретируемость.

7. Финальная модель и отчёт

7.1 Выберите финальную модель по результатам **валидации**, заморозьте гиперпараметры и дайте **одну** честную оценку на **test**.

7.2 Кратко опишите: (i) как выглядели данные и зачем выбранная предподготовка; (ii) какая модель оказалась лучшей и почему; (iii) ключевые графики (остатки, кривые валидации); (iv) риски утечек и как Вы их предотвращали; (v) *что не получилось* и почему.

Чек-лист перед сдачей

- Есть train/val/test и объяснение схемы; гиперпараметры подбирались *без* заглядывания в test.
- Показаны и обсуждены нормальные уравнения и их численная реализация.
- Продемонстрированы эффекты мультиколлинеарности и влияние стандартизации.
- Сравнены MSE vs MAE на данных с/без выбросов.
- Есть Ridge, Lasso, Elastic Net с осмысленным подбором λ/α.
- Есть два робастных метода из списка (Huber/Log-cosh/RANSAC/Квантильная).

- Есть мини-исследование PCA vs Регрессия (+PCR).
- Финальная оценка дана только на **test**; в ноутбуке есть выводы.

Критерии оценивания (100 баллов)

- EDA & разбиение, отсутствие утечек — 10
- OLS: аналитика + численные методы, анализ обусловленности — 15
- MSE vs MAE на данных с выбросами — 10
- Ridge: CV-подбор λ, анализ смещения/разброса — 10
- Lasso: путь решений, разреженность, сравнение с Ridge — 10
- Elastic Net: подбор и интерпретация — 5
- Робастные методы (2 из 4) — 15
- PCA vs Регрессия + PCR — 10
- Финальная модель на test + отчёт с выводами — 10
- Репродуцируемость и аккуратность (seed, версии, графики, подписи) — 5

Бонус (до +10)

- Качественная диагностика гомоскедастичности/автокорреляции (тесты/аргументация) — +5
- Акуратная иллюстрация компромисса интерпретируемости и качества между PCR/Ridge/Lasso на Вашем датасете — +5

Памятка: формулы (для справки)

$$\hat{\omega}_{\text{OLS}} = (X^\top X)^{-1} X^\top y, \quad \hat{\omega}_{\text{ridge}} = (X^\top X + \lambda I)^{-1} X^\top y.$$

При MAE аналитического решения нет; используйте подходящие численные методы.