

Linear and Logistic Regression

!!! Разглеждаме dataset-а като едно векторно пространство

n (реда) \times m (колони)

n - observation (една точка в m -мерното пространство)

m - feature (едно измерение в данните)

! По конвенция \rightarrow матрично умножение - ред по стълб

Linear Regression

Определение

Линейният регресионен анализ се използва за прогнозиране на стойността на дадена променлива въз основа на стойността на друга променлива.

Линейната регресия е основен статистически метод за анализ на зависимостта между променливи. Тя се използва за предсказване на стойността на зависима променлива (наречена **целева променлива** или **отговор**) въз основа на една или повече независими променливи (наречени **признаци** или **предиктори**). В основната си форма, линейната регресия предполага, че съществува линейна зависимост между целевата променлива и независимите променливи.

Исходната променлива (зависима променлива, $target$) е число (става въпрос за регресия) и линейна комбинация на входните променливи с неизвестни коефициенти (в даден момент трябва да ни станат известни)

Пример:

Представи си, че искаш да предвидиш цената на жилище (y) в зависимост от неговата площ (x). Линейната регресия ще се опита да намери права линия, която най-добре описва тази зависимост, така че можеш да използваш модела за прогнозиране на цената на ново жилище въз основа на неговата площ.

Предимства на линейната регресия:

- **Простота:** Лесна за разбиране и интерпретация.
- **Ефективност:** Бързо се изпълнява дори при големи данни.
- **Интерпретируеми резултати:** Можеш лесно да разбереш как всяка независима променлива влияе на зависимата променлива.

Ограничения:

- **Линейност:** Предполага, че връзката между променливите е линейна, което не винаги е вярно в реалния живот.
- **Чувствителност към шум:** Моделът е чувствителен към аномалии и неточни данни.
- **Корелация на признаците:** Ако независимите променливи са силно корелирани помежду си (мултиколинеарност), това може да доведе до нестабилни коефициенти и трудности при интерпретацията.

Зависима и Независима променлива

В статистиката и анализа на данни, **зависима променлива** и **независима променлива** са термини, използвани за описание на връзката между различни променливи в даден модел.

Зависима променлива

Зависимата променлива (наричана още **отговор** или **целева променлива**) е тази, която се опитваме да предскажем или обясним. Тя зависи от стойностите на независимите променливи. С други думи, тя променя стойността си в зависимост от промените в независимите променливи.

Независима променлива

Независимата променлива (наричана още **фактор** или **предиктор**) е променливата, която оказва влияние върху зависимата променлива. Тя е независима в смисъл, че не зависи от другите променливи в модела, а влияе върху тях.

Пример:

Представи си, че правиш проучване за това как **времето за учене** влияе на **оценките на учениците** на изпит.

- **Зависима променлива:** Оценката на ученика на изпита (това е променливата, която искаме да предскажем).
- **Независима променлива:** Времето, което ученикът прекарва в учене (това е факторът, който оказва влияние върху оценката).

Gradient descent

Един от най-популярните алгоритми за оптимизация, използван в машинното обучение и други области за минимизиране на функцията на загубата. Той помага да се намери минималната стойност на дадена функция, която представлява разликата между предсказаните и реалните стойности (грешката на модела).

Представи си, че се намираш на върха на хълм и искаш да слезеш до най-ниската точка (минимум). За да постигнеш това, можеш да следваш наклона на хълма (градиента) в посоката на най-стръмното слизание. Градиентният спуск прави точно това – той изчислява наклона на функцията (градиента) и се движи в посоката, противоположна на него, за да намали стойността на функцията.

MinMaxScaler - взима всяка една колона и я трансформира в стойности между 0 и 1 (by default)