Regression Algorithms

Week 11 Day 02

DS 3000 – Foundations of Data Science

1

Reminders

HW 7

Thursday, November 14

FP3

Thursday, November 14

In-class activity

FP4

Available now

2

Outline

Multiple Regression

Regression Algorithms

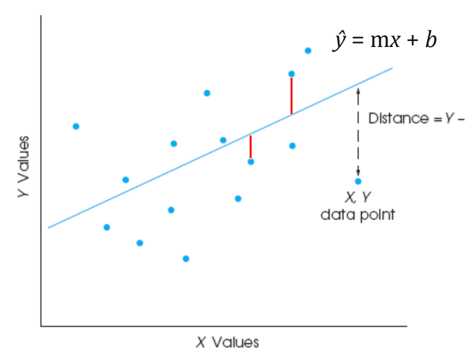
FP3 and FP4

3

Ordinary Least Squares

The distance between the actual data points (y) and the predicted point on the line ($\hat{y}$) is defined as

$y - \hat{y}$



The goal of regression is to find the equation for the line that minimizes these distances.

min sum of squared error (SSE)

$$\arg\min_{m,b} \sum_{i=1}^{N} e_i^2 = (y_i - (mx_i + b))^2$$

4

Multiple Linear Regression

$$\hat{y} = m_1x_1 + m_2x_2 + \ldots + m_nx_n + b$$

$\hat{y}$ is the predicted value of the outcome variable (y)

ms are the feature **coefficients**

xs are the values of the features

b is the **intercept**

5

6

Regression Metrics

R-squared:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Mean Squared Error:

$$\text{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2$$

7

Preprocessing: Scaling

Some ML algorithms require features to be on the same scale

MinMaxScaler

Calculates the min and max each feature in the training set

For each feature x_i: transforms a given feature value to a scaled version, x_i', using the formula

$$x_i' = (x_i - x_i^{MIN})/(x_i^{MAX} - x_i^{MIN})$$

8

9

Ridge Regression

Regularized linear regression

Restricts coefficients to reduce model complexity

Adds a penalty for large variations in coefficients

$$\underset{m,b}{\arg\min} \sum_{i=1}^{N} e_i^2 = (y_i - (mx_i + b))^2 + \alpha \sum_{j=1}^{p} m_j^2$$

Uses L2 regularization:

Minimize sum of squared coefficients

10

Ridge Regression: Tuning

L2 regularization can be controlled by the alpha, α, parameter

$$\arg\min_{m,b} \sum_{i=1}^{N} e_i^2 = (y_i - (mx_i + b))^2 + \alpha \sum_{j=1}^{p} m_j^2$$

Higher alpha values mean

- More regularization and more restricted coefficients
- Simpler model (reduces overfitting)

11

12

Lasso Regression

An alternative to Ridge for regularizing linear regression

Restricts coefficients to reduce model complexity

Adds a penalty for large variations in coefficients

$$\arg\min_{m,b} \sum_{i=1}^{N} e_i^2 = (y_i - (mx_i + b))^2 \quad + \quad \alpha \sum_{j=1}^{p} |m_j|$$

Uses L1 regularization:

Minimize the sum of the absolute values of the coefficient

13

Lasso Regression: Tuning

L1 regularization can be controlled by the alpha, α, parameter

$$\arg\min_{m,b} \sum_{i=1}^{N} e_i^2 = (y_i - (mx_i + b))^2 \quad + \quad \alpha \sum_{j=1}^{p} |m_j|$$

Higher alpha values mean

More regularization and more restricted coefficients

Some coefficients will be exactly zero (automatic feature selection)

Simpler model (reduces overfitting)

Easier to interpret

14

15

Linear, Ridge, and Lasso Regression

Linear regression has no parameters

Scales well to high-dimensional datasets

Use Ridge or Lasso when you need to tune a regression model

Use Ridge if you have many important features

Use Lasso if you have many features but only a few of them are expected to be important

16

FP3: Peer-review

If you are in	Review
Group 01	Group 17
Group 02	Group 16
Group 03	Group 15
Group 04	Group 14
Group 05	Group 13
Group 06	Group 12
Group 07	Group 11
Group 08	Group 10
Group 09	Group 08
Group 10	Group 07
Group 11	Group 06
Group 12	Group 05
Group 13	Group 04
Group 14	Group 03
Group 15	Group 02
Group 16	Group 01
Group 17	Group 09

17

FP4: DS Report

18