Regression

Week 11 Day 01

DS 3000 – Foundations of Data Science

1

Reminders

HW 7

Thursday, November 14

FP3

Thursday, November 14

Remember to post it under the discussion forum as well

2

Outline

Outfitting/Underfitting

Model Optimization

Correlation

Simple Linear Regression

Multiple Linear Regression

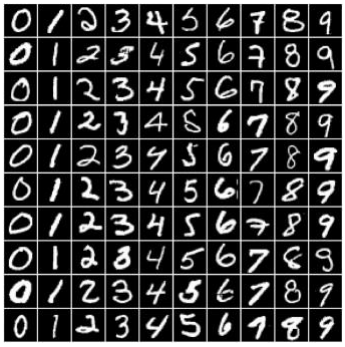
3

Overfitting/Underfitting

4

Machine Learning (ML)

ML algorithms can generalize from existing examples of a task



digit = 2

digit = 0

digit = 9

5

Overfitting/Underfitting

Generalizability

Refers to an algorithm's ability to give accurate predictions for new, previously unseen data

Assumptions:

Future unseen data (test set) will have the same properties as the current training sets

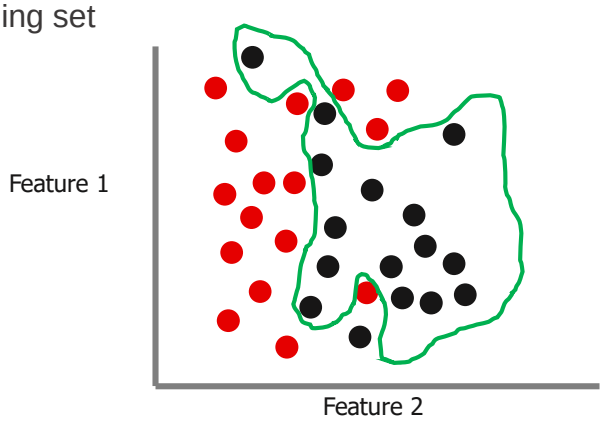
Models that are accurate on the training set are expected to be accurate on the test set

6

Overfitting

Building a model that is too complex for the amount of information we have

Occurs when you fit a model too closely to the particularities of the training set

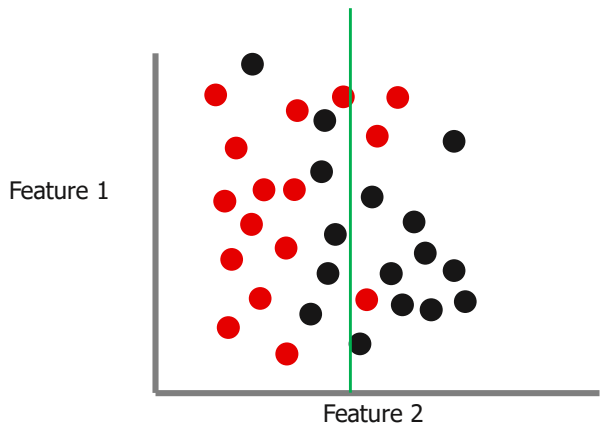


7

Underfitting

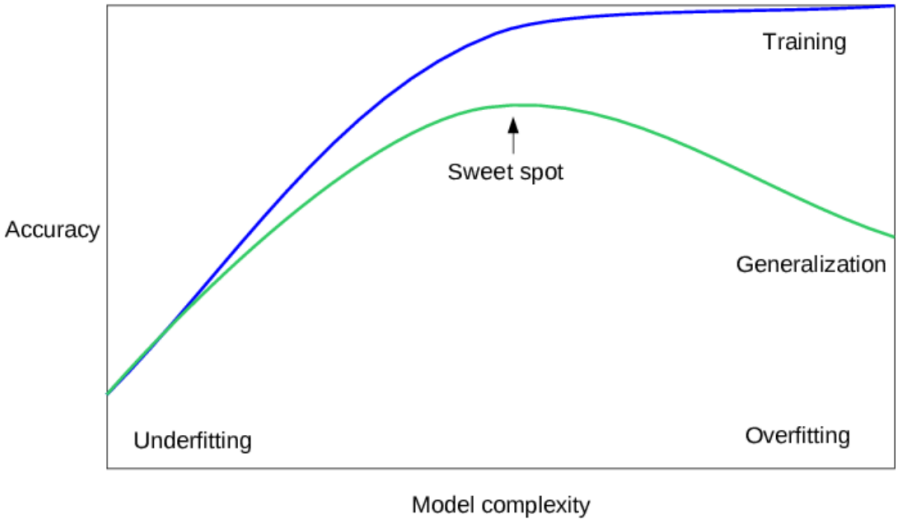
Choosing too simple a model

When the model is too simple, it will fail to capture all aspects of and variability in the data



8

Trade-off between Model Complexity and Accuracy

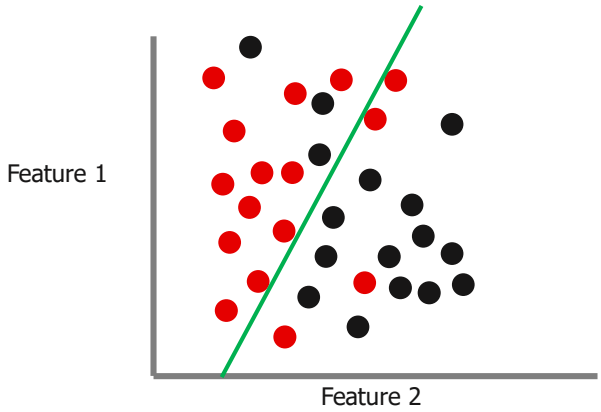


9

Optimization

Goal: Improve the accuracy while avoiding compromising the complexity of the model and thus generalizability of the algorithm

This is the model we want to find



10

Hyperparameter Tuning

Aka optimization or regularization

Choosing parameter values that produce the best possible predictions

The complexity of models constructed using an algorithm can be changed by tuning the specific hyperparameters of that algorithm

11

k-Nearest Neighbors: Tuning

Model complexity

n_neighbors : number of nearest neighbors (k) to consider

By default n_neighbors = 5

As you decrease k, you increase the risk of overfitting

Decision boundaries are more varied

The model becomes more complex (k=1 most complex)

12

13

Support Vector Machines: Tuning

The strength of regularization, or tuning, is determined by C

- By default C=1

Larger values of C: less regularization

- Fit the training data as well as possible
- Each individual data point is important to classify correctly
- Increased model complexity

Smaller values of C: more regularization

- More tolerant of errors on individual data points

14

Decision Trees: Tuning

max_depth:

Controls maximum depth (number of split points)

Most common way to reduce tree complexity and overfitting

Increasing max_depth leads to increased model complexity

- More likely to overfit

15

ICA6: Fruits Dataset

Download the dataset

Use cross-validation to identify the algorithm that generalizes the best

Then change the corresponding parameter

- Try a default value, if any, and then small and large values
- Look at the results on the training data and testing data
- What can you conclude?

16

Correlation

17

Correlational Data

Evaluating relationships for numerical scores (two variables at a time)

- Scores in each pair are identified as *X* and *Y*

Data can be presented in a list showing the two scores for each individual

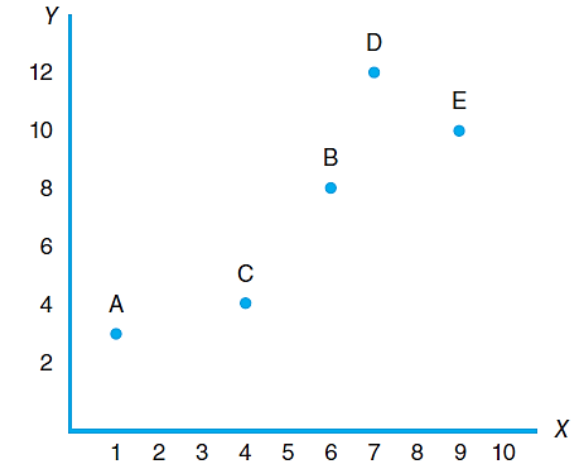
Scores can be shown in a scatter plot graph

- Each individual's score is shown as a single dot with a horizontal coordinate (*X*) and a vertical coordinate (*Y*)

18

Scatter Plot Data from a Correlational Study

Person	X	Y
A	1	3
B	6	8
C	4	4
D	7	12
E	9	10



19

Measuring Relationships

A **correlation coefficient** measures and describes the relationship between two variables

Example: Pearson $r = +0.8$ or $r = -0.8$

It describes three characteristics of a relationship:

- Direction
- Form
- Strength or degree

20

The Direction of the Relationship

The tendency of the change, as indicated by the sign

Positive relationship

Two variables change in the same direction

As one variable increases, the other variable increases

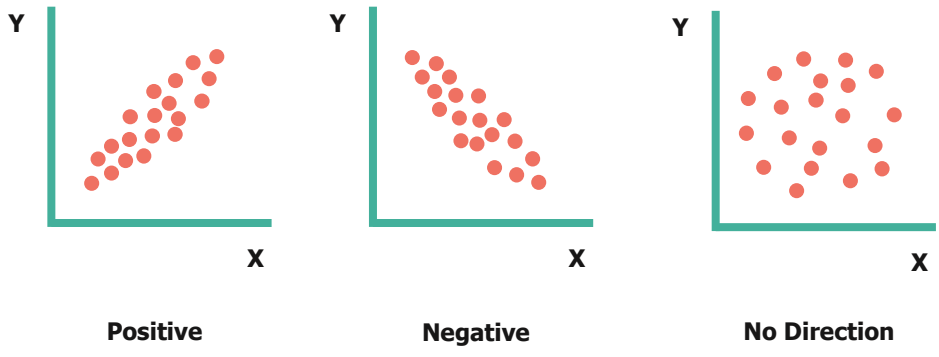
Negative relationship

Two variables change in opposite directions

Increases in one variable matches with decreases in the other variable

21

The Direction of the Relationship



22

The Form of the Relationship

Linear relationship

The data points in the scatter plot tend to cluster around a straight line

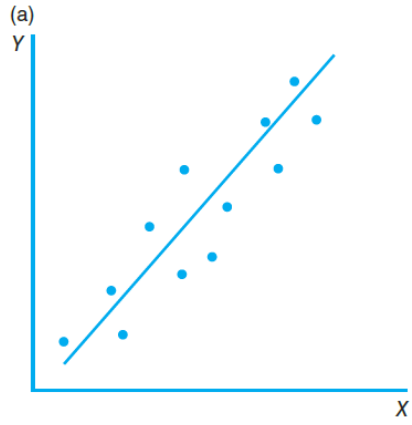
Positive linear relationship

Each time the X variable increases by 1 point, the Y variable increases in a consistently predictable amount

A Pearson correlation describes and measures linear relationships when both variables are numerical scores from interval or ratio scales

23

Linear Relationships



24

The Strength of the Relationship

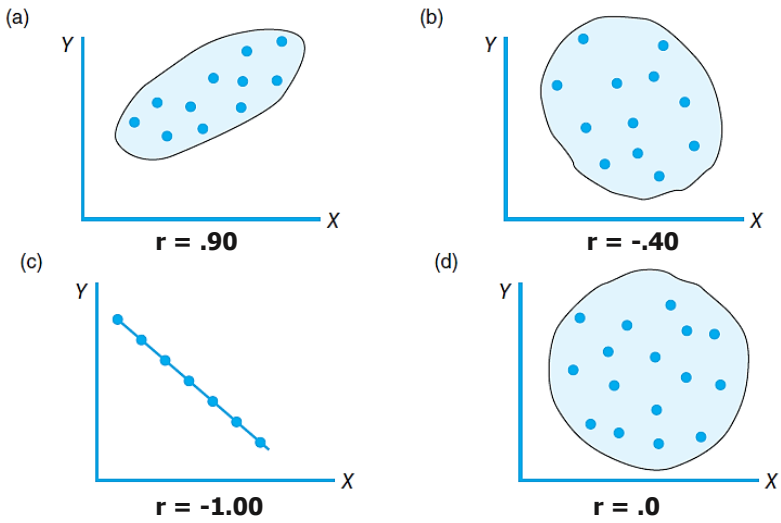
The numerical value indicates the strength or consistency of the relationship

Ranges from 0.0 to 1.0

- A correlation coefficient of 1.0 (or -1.0) indicates a perfectly consistent relationship
- A correlation coefficient of 0 indicates no consistency between the two variables whatsoever
- Intermediate values indicate different degrees of consistency

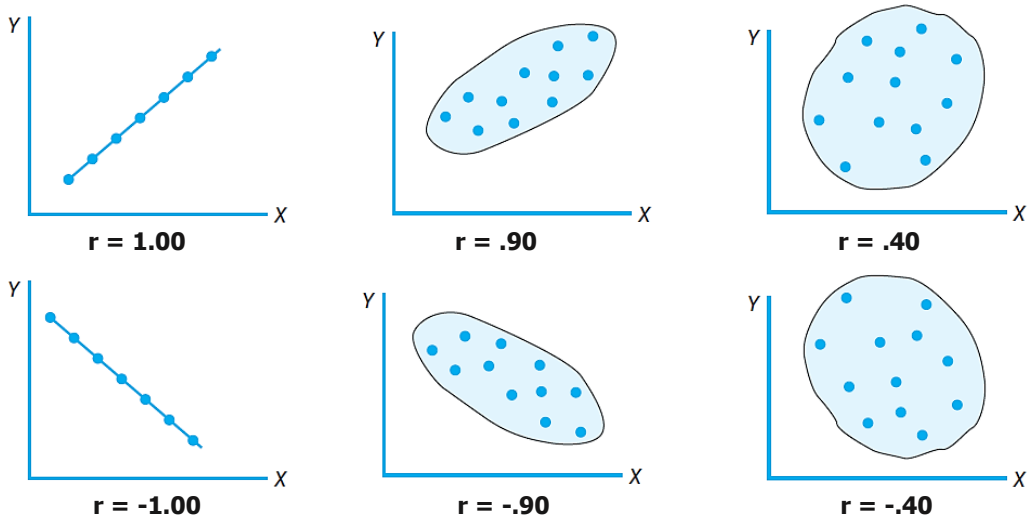
25

The Strength of the Relationship



26

The Strength of the Relationship



27

Interpreting the Strength of a Correlation

Guidelines for interpreting different degrees of consistency

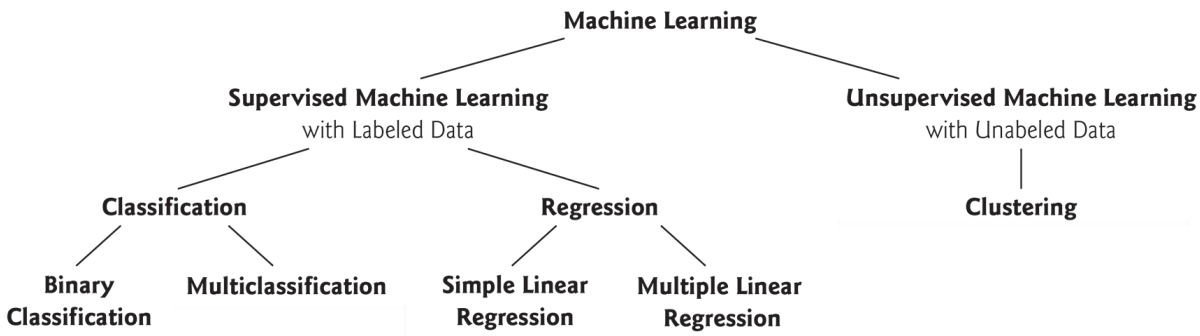
Magnitude of *r*	Degree of Relationship
r = 0.1	Small
r = 0.3	Medium
r = 0.5	Large

28

Supervised Learning: Regression

29

Two Types of ML Tasks



30

Regression

Regression algorithms predict the values of continuous outcome variables

Simple regression is based on one single feature variable

Multiple regression is based on multiple feature variables

31

Simple Linear Regression

Describes the relationship between a feature and target with a straight line, known as the regression line

$$\hat{y} = mx + b$$

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

m is the slope of the line (the unstandardized regression coefficient)

x is the value of the predictor variable

b is the intercept, or the point where the regression line intercepts the y-axis (the value of y when x = 0)

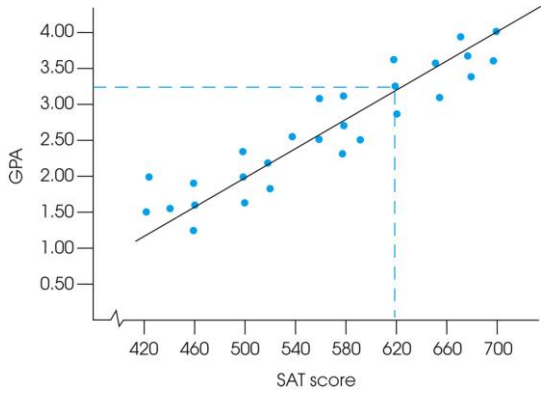
32

Regression Line

The line through the data

- Makes the relationship easier to see
- Shows the central tendency of the relationship
- Can be used for prediction

Regression analysis precisely defines the line



33

Simple Linear Regression

Regression is a method of finding an equation describing the best-fitting line for a set of data

How to define a “best fitting” straight line when there are many possible straight lines?

The answer:

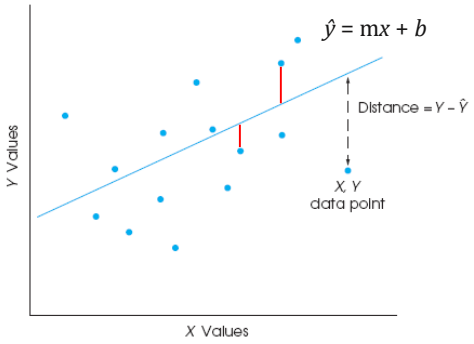
a line that is the best fit for the actual data is the one that minimizes prediction errors

34

Simple Linear Regression

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)

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

35

Time Series Analysis

Time series

Sequences of values (**observations**) associated with points in time

- daily closing stock prices
- hourly temperature readings
- changing positions of a plane in flight
- annual crop yields
- quarterly company profits
- time-stamped tweets from Twitter users worldwide

Simple linear regression is commonly used to make predictions from time series data

36

Time Series Tasks

Time series analysis

Looks at existing time series data for patterns (like **seasonality**)

Time series forecasting

Uses past data to predict the future

Univariate time series: one observation per time

Multivariate time series: two or more observations per time

37

38