GENERAL ASSEMBLY

INTRODUCTION TO CLASSIFICATION

Angelo Klin
Katra Analytics

LEARNING OBJECTIVES

- Define Class Label and Classification
- Build a K-Nearest Neighbours using the scikit-learn library
- Evaluate and tune a model by using metrics such as classification accuracy and error

PRE-WORK

PRE-WORK REVIEW

- Understand how to optimise for error in a model
- Understand the concept of iteration to solve problems
- Measure basic probability

OPENING

INTRODUCTION TO CLASSIFICATION

OPENING

INTRODUCTION TO CLASSIFICATION

- So far we have worked primarily with regression problems
- We have focused on predicting a continuous set of values
- That means we have been able to use distance to measure how accurate our prediction is
- However, for other problems, we need to predict binary responses
  - Will a loan will default or not?
  - Is an email a spam or not?

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

DIRECTIONS: (5 MINUTES)

1. What if we want to build a model to predict a set of values, like a photo colour or the gender of a baby?
2. Can we use regression for binary values?
3. Do the same principles apply?

WHAT IS CLASSIFICATION?

WHAT IS CLASSIFICATION?

- **Classification** is a Machine Learning problem for solving a set value given the knowledge we have about that value
- Many classification problems are trying to predict **Binary** values
- For example, we may be using patient data (medical history) to predict whether the patient is a smoker or not

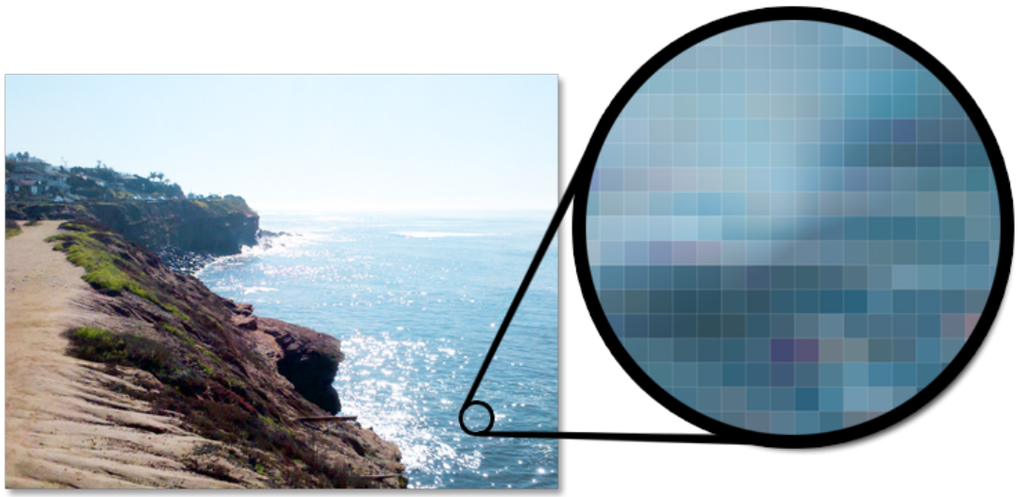
WHAT IS CLASSIFICATION?

- Some problems do not appear to be binary at first glance, however, you can boil down the response to a **Boolean** (True/False) value

- What if you are predicting whether an image pixel will be **red** or **blue**?

- We do not need to predict that a pixel is **blue**, just that it is **not red**

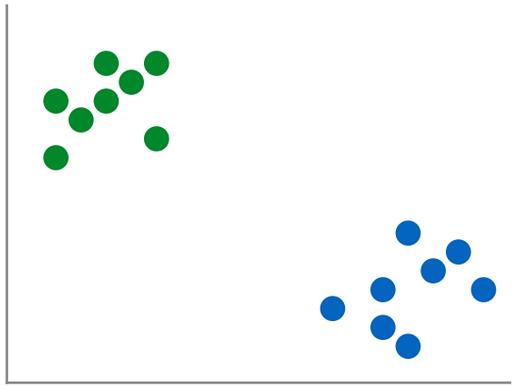
- This is similar to the concept of dummy variables



WHAT IS CLASSIFICATION?

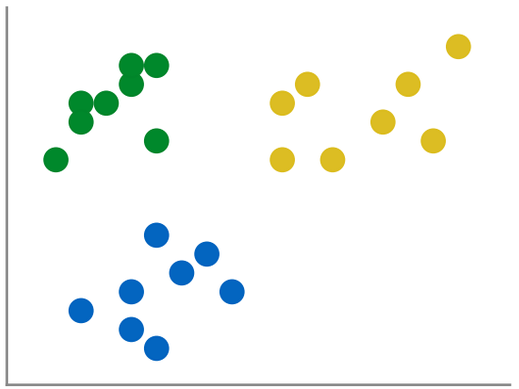
- Binary classification is the simplest form of classification

Binary Classification



- However, classification problems can have multiple class labels
  - Instead of predicting whether the pixel is green or blue, you could predict whether the pixel is yellow, green or blue

Multi-class Classification



WHAT IS A CLASS LABEL?

- A Class Label is a representation of what we are trying to predict
  - The target

- Examples of Class Labels

Data Problem	Class Label
Patient data problem	Smoker, Not Smoker
Pixel colour	Red, Green, Blue

DETERMINING BETWEEN REGRESSION AND CLASSIFICATION

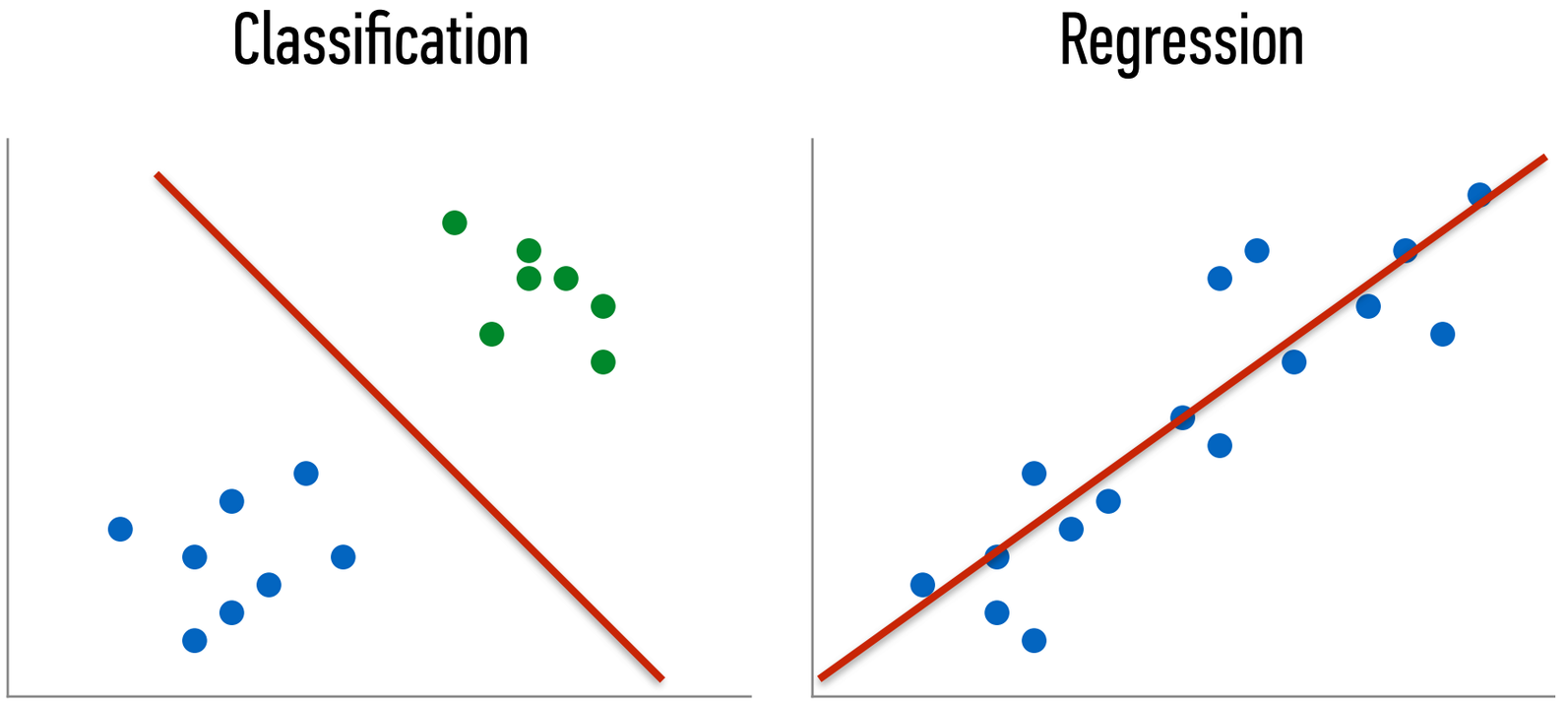
- One of the easiest ways to determine if a problem is regression or classification is to determine if our target variable can be ordered mathematically

- For example, if predicting company revenue, $100 is greater than $90
  - This is a regression problem because the target can be ordered

- However, if predicting pixel colour, red is not inherently greater than blue
  - Therefore, this is a classification problem

DETERMINING BETWEEN REGRESSION AND CLASSIFICATION

- Classification and Regression differ on what is to be predicted



GUIDED PRACTICE

REGRESSION OR CLASSIFICATION?

ACTIVITY: REGRESSION OR CLASSIFICATION?

DIRECTIONS: (5 MINUTES)

1. Review the following situations and decide if each one is a regression problem, a classification problem or neither:
   a. Using the total number of explosions in a movie, predict if the movie is by JJ Abrams or Michael Bay
   b. Determine how many tickets will be sold to a concert given who is performing, where and the date and time
   c. Given the temperature over the last year by day, predict tomorrow's temperature outside
   d. Using data from four cell phone microphones, reduce the noisy sounds so the voice is crystal clear to the receiving phone
   e. With customer data, determine if a user will return or not to an e-commerce website in the next 7 days

INDEPENDENT PRACTICE

BUILD A CLASSIFIER

ACTIVITY: BUILD A CLASSIFIER

EXERCISE

DIRECTIONS: (20 MINUTES)

1. Re-explore the iris dataset and build a program that classifies each data point
   a. Use `if-else` statements and some Pandas functions
2. Measure the **Accuracy** of your classifier using the math of "total correct" over "total samples"
3. Your classifier should be able to:
   a. Get one class label 100% correct (one type of iris is easily distinguishable from the other two)
   b. Accurately predict the majority of the other two classes with some error (hint: make sure you generalise)

ACTIVITY: BUILD A CLASSIFIER

DIRECTIONS: (20 MINUTES)

1. How simple could the if-else classifier be while remaining relatively accurate?
2. How complicated could our if-else classifier be and remain completely accurate? How many if-else statements would you need or nested if-else statements, in order to get the classifier 100% accurate? (The above uses a count of 2)
3. Which if-else classifier would work better against iris data that it has not seen? Why is that the case?

WHAT IS K NEAREST NEIGHBOURS?

WHAT IS K NEAREST NEIGHBOURS?

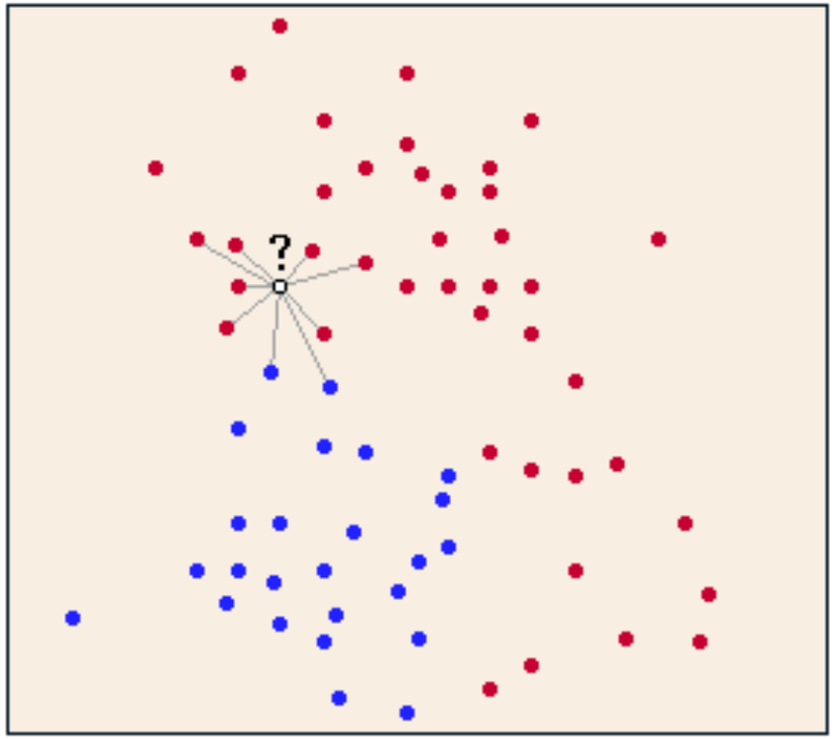
- Suppose we want to determine your favourite type of music
  - How might we determine this without directly asking you?

- Generally, friends share similar traits and interests (e.g. music, sports teams, hobbies, etc)
  - We could ask your five closest friends what their favourite type of music is and take the majority vote

- This is the idea behind KNN: we look for things similar or close to our new observation and identify shared traits
  - We can use this information to make an educated guess about a trait of our new observation

WHAT IS K NEAREST NEIGHBOURS?

- KNN uses distance to predict a Class Label
- This application of distance is used as a measure of similarity between classifications
- We are using shared traits to identify the most likely Class Label



WHAT IS K NEAREST NEIGHBOURS?

- K Nearest Neighbours (KNN) is a classification algorithm that makes a prediction based upon the closest data points

- The KNN algorithm:
  - For a given point, calculate the distance to all other points
  - Given those distances, pick the k closest points
  - Calculate the probability of each class label given those points
  - The original point is classified as the class label with the largest probability (“votes”)

ACTIVITY: KNOWLEDGE CHECK

DIRECTIONS: (5 MINUTES)

1. In what other tasks do we use a heuristic similar to K Nearest Neighbours?

DEMONSTRATION

KNN IN ACTION

WHAT HAPPENS IN TIES?

- What happens if two classes get the same number of votes?
  - This could happen in binary classification if we use an even number for k
  - This could also happen if there are multiple class labels
- In scikit-learn, it will choose the class that it first saw in the training set

WHAT HAPPENS IN TIES?

- We could implement a weight, taking into account the distance between the point and its neighbours
- This can be done in scikit-learn by changing the weights parameter to “distance”
- Try changing the weights parameter
  - How does this affect accuracy?

WHAT HAPPENS IN HIGH DIMENSIONALITY?

- Since KNN works with distance, higher dimensionality of data (i.e. more features) requires **significantly** more samples in order to have the same predictive power

- Consider this: with more dimensions, all points slowly start averaging out to be equally distant
  - This causes significant issues for KNN

- Keep the feature space limited and KNN will do well
  - Exclude extraneous features when using KNN

WHAT HAPPENS IN HIGH DIMENSIONALITY?

- Consider two different examples: classifying users of a newspaper and users of a particular toothpaste
  - The features of the newspapers are very broad and there are many: sections, topics, types of stories, writers, online vs print, etc
  - However, the features of a toothpaste are more narrow: has fluoride, controls tartar, etc
- For which problem would KNN work better?
  - KNN would work better on classifying users of a particular toothpaste since the feature set is more narrow and distinct

CLASSIFICATION METRICS

CLASSIFICATION METRICS

- Metrics for regression do **not** apply to classification
- We could measure the distance between the probability of a given class and an item being in that class
  - Guessing 0.6 for a 1 is a 0.5 error
- But this overcomplicates our goal: understanding binary classification, whether something is black or white, right or wrong
- To do this we will measure “correctness” or “incorrectness”

CLASSIFICATION METRICS

- We will use two primary metrics: accuracy and misclassification rate
  - **Accuracy** is the number of correct predictions out of all predictions in the sample
    - This is a value we want to **maximise**
  - **Misclassification rate** is the number of incorrect predictions out of all predictions in the sample
    - This is a value we want to **minimise**

- These two metrics are directly opposite of each other

$$accuracy = 1 - misclassification\ rate$$

CLASSIFICATION METRICS

- **WARNING**: You cannot use regression evaluation metrics for a classification problem or vice versa
  - This is a common mistake

- scikit-learn will not intuitively understand if you are doing regression or classification, so make sure to manually review your metrics

INDEPENDENT PRACTICE

SOLVING FOR K

SOLVING FOR K

- One of the primary challenges of KNN is solving for k
  - How many neighbours do we use?

- The smallest k we can use is *1*, however, using only one neighbour will probably perform poorly

- The largest k we can use is *n - 1* (every other point in the data set), however, this would result in always choosing the largest class in the sample; This would also perform poorly

ACTIVITY: SOLVING FOR K

DIRECTIONS: (35 MINUTES)

1. Use the lesson 8 starter code
   a. ~/lessons/lesson-08/code/starter/starter-8.ipynb
2. Use the iris data set to answer the following questions:
   b. What is the accuracy for k = 1?
   c. What is the accuracy for k = n - 1?
   d. Using Cross Validation, what value of k optimises model accuracy. Create a plot with k as the x-axis and accuracy as the y-axis (called a “fit chart”) to help find the answer

ACTIVITY: SOLVING FOR K (BONUS)

DIRECTIONS: (35 MINUTES)

1. By default, the KNN classifier in scikit-learn uses the Minkowski metric for distance
   a. What type of data does this metric work best for?
   b. What type of data does this distance metric not work for?
   c. You can read about distance metrics in the scikit-learn documentation
2. It is possible to use KNN as a regression estimator. Determine the following:
   a. Steps that KNN Regression would follow
   b. How it predicts a regression value

TOPIC REVIEW

TOPIC REVIEW

- What are Class Labels?
  - What does it mean to classification?
- How is a classification problem different from a regression problem?
  - How are they similar?
- How does the KNN algorithm work?
- What primary parameters are available for tuning a KNN estimator?
- How do you define: accuracy, misclassification?

BEFORE NEXT CLASS

BEFORE NEXT CLASS

DUE DATE

INTRODUCTION TO CLASSIFICATION

Q & A

EXIT TICKETS

DON'T FORGET TO FILL OUT YOUR EXIT TICKET

Exit Ticket Link

What's the lesson number?	08
What was the topic of the lesson?	Introduction to Classification

CREDITS AND REFERENCES

INTRODUCTION TO CLASSIFICATION

- Machine Learning Methods - Computerphile
  - Uwe Aickelin
    - Professor of Computer Science, Faculty of Science, The University of Nottingham
      - Website
      - youTube