4: Classification

```
$ echo "Data Science Institute"
```

What is Classification?

Classification involves predicating a qualitative response by a assigning it to a category. The methods that are used to classify observations are called **classifiers** and most of them work by following two steps:

- Compute the probability that an observation belongs to a category.
- Classify the observation based on some probability threshold (i.e. if the probability that an observation belongs to some category is greater than 0.5 then assign the observation to that category)

Breakout Room

What are some classification methods?

Why use Classification?

We need to predict a qualitiative response.

Example

On the basis of DNA sequence data for a number of patients with and without a given disease, a biologist would like to figure out which DNA mutations are deleterious (disease-causing) and which are not.

Let's categorize this as a class!

Why not use linear regression?

Suppose we are trying to diagnose a patient with either a *stroke*, *drug overdose*, or *epileptic seizure* based on their symptoms. We can code this response as follows

> $$Y = \begin{cases} 1 & \text{if stroke;} \\ 2 & \text{if drug overdose;} \\ 3 & \text{if epileptic seizure.} \end{cases}$$

At this point we could use linear regression to predict Y based on a set of predictors. However there are several problems with this coding. One of them is we cannot use linear regression.

Breakout Room: Why can't we use linear regression?

> $$Y = \begin{cases} 1 & \text{if stroke;} \\ 2 & \text{if drug overdose;} \\ 3 & \text{if epileptic seizure.} \end{cases}$$

Why not use linear regression?

Other problems:

- Implies an ordering of the outcomes.
- The difference between epileptic seizure and stroke versus stroke and drug overdose is assumed to be the same.

Why not use linear regression?

Suppose we are trying to diagnose a patient with either a *stroke*, *drug overdose*, or *epileptic seizure* based on their symptoms. We can code this response as follows

> $$Y = \begin{cases} 1 & \text{if stroke;} \\ 2 & \text{if drug overdose;} \\ 3 & \text{if epileptic seizure.} \end{cases}$$

At this point we could use linear regression to predict Y based on a set of predictors. However there are several problems with this coding

A different ordering would give completely different results for the linear regression. ♦
There is no convenient way to code a qualitative response with more than two levels so that linear regression can be used. ♦

Why not use linear regression?

The 0/1 coding for a binary qualitative response variable does not suffer the same problems. However the probabilities we obtain will be difficult to interpret

- negative probabilities
- probabilities above 1

So, linear regression only able to give ♦ ***crude estimates of the probabilities for a binary response.*** ♦

In summary, we don't use linear regression for classification since:

- It does not work for a qualitative response variable with more than 2 classes.
- The probability estimates are not meaningful.

K-Nearest Neighbours

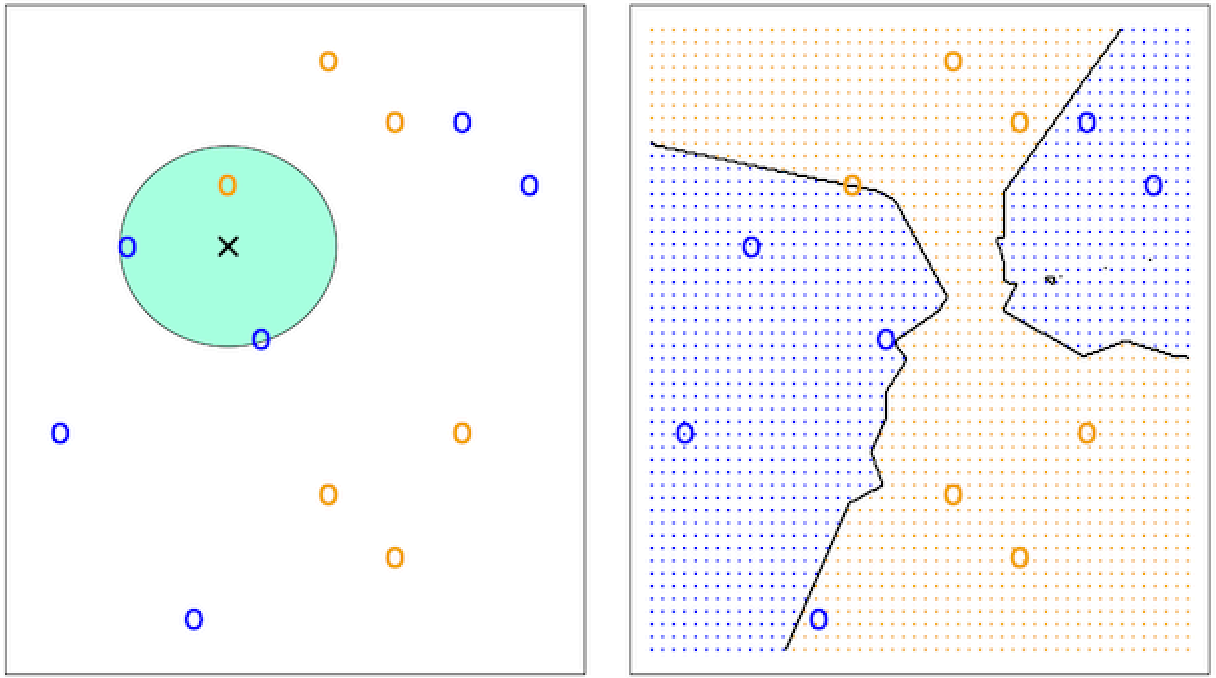
The K-nearest neighbors (KNN) classifier works very differently than any of the previous classification methods. For a test observation x_0, it identifies K training data points that are closest to x_0 (represented by $\mathcal{N}_0$) and estimates the conditional probability for class j as

> $$\Pr\left(Y = j \mid X = x_0\right) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I\left(y_i = j\right)$$

where $I(y_i = j)$ if an **indicator variable** that equals 1 is $y_i = j$ and 0 otherwise. The KNN classifier classifies the test observation x_0 to the class for which the above probability is the largest.

K-Nearest Neighbours

These figures illustrate the KNN approach with $K = 3$. To the left we see the 3 closest points to x are 1 orange and 2 blue so this observation will be classified as blue. The right figure shows the decision boundaries where an observation will be classified as blue or orange.



Exercise: K-Nearest Neighbours

Open the Classification Exercises Jupyter Notebook file.

- Go over the "K-Nearest Neighbours" section together as a class.
- 5 minutes for students to complete the questions from "K-Nearest Neighbours".
- Questions should be completed at home if time does not allow.

References

Chapter 4 and section 2.2.3 of the ISLP book:

James, Gareth, et al. "Classification." An Introduction to Statistical Learning: with Applications in Python, Springer, 2023.