AI Coursework Lecture Notes

Coursework may change (marks may move around)

Penguins Dataset

Idea is to take the dataset, shown below:

A table of numbers and letters

Description automatically generated

1. Has three different species, you know 4 figures (bill\_length, bill\_depth, flipper\_length, body\_mass) sex also import, year less so. May want to consider males and females separately.
2. The obvious task is to classify the penguins according to these numbers. It’s not going to be super accurate. **This is coursework to show I have learnt the things you can do with datasets and there are parameters you can choose.** Not a competition about who can classify the penguins the best, haven’t specified how to classify. Most marks are for having a go.

Start

Start by plotting a few graphs of what one characteristic looks like against another characteristic, initial exploration. Think a bit what happens to missing lines, get rid of NA values, see if something has gone wrong. Look at three ranges, mean standard deviation.

Do three things

Two algorithms based on classification/supervised learning one on unsupervised learning.

Supervised

Use the data and try to classify the penguin. This will work a little bit but not terribly well, this is kind of the point. One of the algorithms can be knn, not expecting something really sophisticated, start with the easiest thing.

If doing knn and doing distances, body mass in grams shows range of few hundred kilos, bill length is a few millimetres, if you put data in and did knn, the distance will be wholly dominated by body mass. Should you scale it? One thing you can do is normalise by taking away the mean and divide by the standard deviation, then they become comparable. Might also compare with different metrics, should you just take Euclidian distance between two penguins in body description space or should you take distance between penguins as something more complicated where you use the L1 distance. There is opportunity to do some experimentation, doesn’t need to work to get marks. A few sensible experiments, what would happen if I changed k? How do is it change with k, do the graph, that rarely produced an obvious good value of k, have a look and have a discussion.

Unsupervised learning

One other thing, might be unsupervised learning how clustered or maybe regression? For regression could regress bill length vs bill depth, when you do that you could do that for all the penguins or you could it for each penguin by type or each penguin by gender. See if regression coefficients different in each case.

Formatting

Lecturer emphasis subject is about communication, use data to tell a story, at some point you want to say ‘this is something’,’ this is something else’ when we’re doing this we have to say how clear are we when we say this is a chinstrap penguin. What are the assumptions that go into it? Presentation is important. We should use default font and section heads.

Make graphs nice, don’t try to do too much on one graph, include units. Make sure they can read graph, when you export a png in python, it will be 10inches by 7inches. Before you export them, set them to smaller (4 by 3 inches) then export this way, fonts will be the right size.

Don’t have a huge range on x axis so all data is crushed into one little line. If you have two colours don’t have points on top of the other

Trying to communicate not just ‘this penguin is this species’ we want to make it clear what we have done and how sure we are.

Exceptional

If you do obvious thing and do it nicely you get all marks, extra marks for something surprising.

Captions in the graphs should be self-contained, people don’t read things all the way through. Should be able to just read the caption. Shouldn’t just be ‘this is a graph of bill length vs bill depth’. Need to explain what you’ve done, may repeat the text but that’s okay. The figure tone is completely factual ‘this is what we did, this is what you see’, text ‘this is why we did it’ and ‘this is what we conclude from it’.

Second part

The idea is making sure we’ve thought a bit about it, some examples, maybe evidence of having read a paper or two and blogposts. Wants us to think about some examples and read around them.