Practical/Tutorial Week 4 – Basic Parametric models and Statistical Learning

DATA7703 - Machine Learning for Data Scientists

Aims:

- To gain some experience in performing regression with linear and polynomial models and classification with parametric models.

Linear and Polynomial Regression

Question 1:

a) Plot the function (f(x) = x^3 + 1) over the domain [-1,1]. This is the "true" function for our regression problem. However, the data that we observe (i.e. what we train the model with) will have Gaussian noise added to the output/target values.
b) Create a "sample training set" of 30 points by generating a random set of x/input values, and then corresponding outputs by evaluating the true function at each input value and adding Gaussian random noise. Plot the data on the same axes as the function.
c) Perform linear regression using your training set and calculate the training set (sum of squares) error.
d) Perform polynomial regression using your training set and calculate the training set (sum of squares) error. Experiment with different order polynomials and observe the effect on the training set error.

Question 2:

a) Download the dataset pokemonregr.csv from blackboard. This contains 7 of the columns from the pokemon dataset. Open the file in a spreadsheet or preview the first few lines to see what it "looks like".
b) Fit a linear regression model to the data, using the final column (weight in kg) as the output/target variable.
c) What are the coefficients/parameters of your regression model?
d) Recall that the coefficients can be interpreted as a weight of the importance of each weight to the predicted output. Which feature is most important for predicting weight?
e) One potential issue is that the data you have used is not normalized (read the short section at the end of Sec.2.2 in the textbook. Normalise your input features and refit the linear regression model. Compare the coefficients for the resulting model to your answer from (d).

Logistic Regression

Question 3:

Note: implementations of logistic regression can be a part of functions that do additional things and return lots of additional information. Don't worry about this extra stuff for now. Feel free to use ChatGPT to help you with this question!

a) Fit a logistic regression model to the data in w3classif.csv. What are the model parameter values?
b) Given a test data point x' = (1.1, 1.1), what does your model predict as p(y'=1|x')?
c) Plot the data with discriminant function and the decision regions for your model.