COMP4702/COMP7703 - Machine Learning

Prac W6 – Performance contd. and Loss Functions

Aims:

- To gain some practical experience in evaluating supervised machine learning models.
- To produce some assessable work for this subject.

Procedure:

In Example 4.5 in the Lindholm et al. textbook, a thyroid dataset is used. This dataset is available on the course blackboard site (downloaded from: https://archive.ics.uci.edu/ml/datasets/thyroid+disease), in the files `ann-train.data` and `ann-test.data` (it seems the books has used these files for their training and hold-out validation sets). The files are (space delimited) "csv" files, with the last column being the class label.

Question 1: Train a *k*-NN model (choose some reasonable value for *k*) on the training set and calculate a confusion matrix for the hold-out validation set.

Question 2: Attempt to reproduce Example 4.5 from the Lindholm et al. textbook. You will need to:

- Convert the data into a binary classification problem.
- Train a logistic regression model on the training data.
- Evaluate the trained model to calculate a confusion matrix.
- Vary the decision threshold for the model as done in Example 4.5 and recalculate the confusion matrix.

Question 3: In Prac W4 we applied linear regression to a pokemon dataset, where the loss function was sum of squares (or mean squared) error. Revisit this task but add (a) L^2; (b) L^1 regularisation to the loss function, with some suitable value for the regularization hyperparameter (see Section 5.3 of the textbook). Compare the coefficient values from your different trained models.