

# Data Science Toolbox Question Sheet

## 05.1 Introduction to Classification

Daniel Lawson

### Block 5

1. The baseline classifier is often chosen to be *logistic regression*. From a computational standpoint, is logistic regression any harder than regular regression?
2. Describe how K-nearest neighbours can be used as a classifier for a sample point that is not in the training data set.
3. In Linear Discriminant Analysis (LDA):
  - a. You are given the equation for a scatter matrix as:

$$M = \sum_{i \in D_k} (\vec{x} - \vec{\mu}_k) (\vec{x} - \vec{\mu}_k)^T .$$

- Is this the within-class or between-class scatter matrix, and why?
- b. How could you choose the correct number of dimensions  $k$ ?
  - c. You are provided with a test datapoint  $x$ . Interpret the following equations for prediction: A:  $Pr(x|y = c)$ , B:  $Pr(x|y = c)p(y = c)$ , C:  $argmax_c(Pr(x|y = c))$ .
4. For a Support Vector Machine (SVM):
    - a. If we define the SVM for classifying a point  $x$  via the equation  $w \cdot (x - w_0) = w \cdot x + b = 0$ , what do the quantities  $w_0$ ,  $w$ , and  $b$  mean geometrically?
    - b. The SVM finds the ‘maximum margin hyperplane’. What is being maximised, in terms of the above quantities?
    - c. Quadratic Programming is used to solve for the optimal margins. In what sense is Quadratic Programming quadratic, and in what sense is it not?