Statistical Data Mining and Machine Learning (full question)

7. (a) Explain the terms *training set*, *validation set*, *test set*, and *cross-validation*.

   (b) What is the difference between generative and discriminative modelling?

   (c) Consider a binary classification problem with dataset $(x_i, y_i)_{i=1}^n$ with $y_i \in \{+1, -1\}$. We use logistic regression to model the conditional distribution of the labels (y_i) given the data vectors (x_i). The objective function is

$$J(a, b) = \frac{C}{2}\|b\|^2 + \sum_{i=1}^{n} \log(1 + \exp(-y_i(a + b^\top x_i))).$$

   (i) Explain what each term in the objective function is and what it is for.

   (ii) Show that the objective function is convex. Why is convexity a useful property of the objective function?

   (iii) Suppose that the loss associated with incorrectly predicting $+1$ when the true label is -1 is 10 times larger than incorrectly predicting -1 when the true label is $+1$. How might you formulate an objective function for your logistic regression model that better reflects this unbalanced loss?

   (d) Consider a K-class classification problem with dataset $(x_i, y_i)_{i=1}^n$ with $y_i \in \{1, \ldots, K\}$. Suppose we model the conditional distribution of the labels as follows:

$$p(y_i = k|x_i) = \frac{\exp(a_k + b_k^\top x_i)}{\sum_{j=1}^K \exp(a_j + b_j^\top x_i)}.$$

   Write down the L_1-regularized empirical risk associated with the log loss for this model, and derive a gradient descent learning algorithm.