Problem Set 1

1. [100 marks]

   Consider a data set (X, Y), where X is the feature representing "Years-of-Education" and Y is the class/label representing "Salary". Assume the following set of possible values: $X \in \{0, 1, \ldots, 20\}$ and $Y \in \{0, 1, \ldots, 10000\}$ (a value $Y = y$ means a salary of y thousand pounds). Assume further that we are given the true distribution $\mathbb{P}(x, y)$ of the data (X, Y).

   (a) Which of the above assumptions about the data (X, Y) is the least realistic? [15]

   (b) Say that we want to test the performance of some classification algorithm using the following loss model: $L(\widehat{y}, y) = \widehat{y} - y$, where $\widehat{y}$ is the predicted value and y is the true value. Is this a reasonable loss model? Justify your answer. [25]

   (c) Consider some feature x. What is the best corresponding label $\widehat{y}$ for the loss model $L_1(\widehat{y}, y) = \widehat{y} - y$? What about for the loss model $L_2(\widehat{y}, y) = y - \widehat{y}$? [25]

   (d) Under which loss model would $E[Y]$ be the best corresponding label for some feature x? [35]
