Statistics 745

Group Assignment 1

1. The cora AI/ML text data are used. The data consist of a series of computer science papers. The three components of the data are:

   - The first variable is the papers topic, it is either about ML or AI.
   - The second set is a subset of the words in the title of the paper. Each word is a partial match, i.e. the word learning is learn. Certain words, like the, on, etc. are removed.
   - The third set is an observed graph. Each edge is the count of the number of papers that cite the same document.

   Answer the following questions:

   (a) Identify $\mathcal{G}$, $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{A}$. Describe each of them.

   (b) Explain how this problem could be construed as a supervised problem. Now, explain how this problem could be construed as an unsupervised problem.

   (c) What words are in the title of the first document? From this, can you guess what the title is? Lastly, list the document indices that are similar to the first document.

2. Consider estimation theory with absolute error loss (referred to as L_1 loss)

$$L(Y, f(X)) =\mid Y - f(X) \mid$$

   Denote $Pr(Y > c \mid X = x) = \int_c^\infty f_x(t)dt$ with conditional density $f_x(t)$. We define c as the conditional median of Y given $X = x$ (i.e. $c = \text{median}(Y \mid X = x)$) whenever c satisfies $Pr(Y \leq c \mid X = x) = \frac{1}{2}$. For this, we are interested in the pointwise minimizer $\hat{f}$ of $EPE(f)$, in other words $\hat{f}$ satisfies:

$$\hat{f}(x) = \arg\min_c E_{Y|X}[\mid Y - c \mid\mid X = x].$$

   Derive the function f(x) that satisfies this.