Homework 2

October 8, 2019

- Chapter 3
  1) Give it some thought: #3, #4
     Remark: In problem 4, we define irrelevant attributes as follows. Suppose that the domain has m attributes. Denote
     $$\mathcal{S} = \{1, 2, \ldots, m\}.$$
     We represent an example by an m-dimensional vector of the form
     $$\boldsymbol{x} = (x_1, x_2, \ldots, x_m).$$
     Let s be a subset of $\mathcal{S}$. We define the distance between attribute vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ using the attributes in set s. Particularly, define
     $$d_s(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i \in s} d_i(x_i, y_i), \tag{1}$$
     where $d_i(x_i, y_i)$ is the distance for the i-th attribute between x_i and y_i. Let $h(\boldsymbol{x}, s)$ be the class of example $\boldsymbol{x}$ determined by a 1-NN classifier using distance measure d_s in (2). Let $c(\boldsymbol{x})$ be the class of example $\boldsymbol{x}$. We say that attributes in subset s are relevant, if the number of examples in the training set that are correctly classified using distance measure $d_{\mathcal{S}}$ is **less than or equal to** that correctly classified using distance measure d_s. We say that attributes in $\mathcal{S} - s$ are irrelevant attributes.
  2) Computer assignment: #1, #2
     Remark: Use the data set in iLMS. For each class, use the first 30 examples for training and the last 20 examples for testing. Your program should output the correct number of classified testing examples, the number of incorrectly classified testing examples, and the classification accuracy. You should upload your source code and a readme file that contains some explanations of your code.

Due date: Thursday, Oct. 17, 2019 (Note: You can submit the homework in class on the due date. Alternatively, you can submit your homework to Room 845 EECS building before 5 pm on the due date. No late homework is accepted.)