Spring 2020 318 711 Homework 4

Use programming tool (e.g., Matlab or Python Scikit-Learn) as needed.

Problem 1

Redo Problem 3 of Homework 3 with a random forest classifier. Compare your results with those from Homework 3.

Problem 2

Redo Problem 4 of Homework 3 with a random forest classifier.

Problem 3

Suppose $x_1, x_2, \ldots, x_n$ are i.i.d. samples of a Gaussian random variable X with $N(a, \sigma^2)$ and assume a is known. Find the maximum likelihood estimate of σ^2.

Problem 4

Same as Problem 3, except now we assume σ^2 is known and a is unknown with a prior

$$p(a) = \begin{cases} \frac{a}{\delta^2} e^{-\frac{a^2}{2\delta^2}} & \text{if } a \geq 0 \\ 0 & \text{if } a < 0 \end{cases} \tag{1}$$

Find the MAP estimate of a. The derivation might more complex here than Problem 3 so just go as far as you can without spending a whole day on it.

Problem 5

X and Y are random variables and we want to estimate Y from X. As discussed in class, the optimal solution to this problem is $\widehat{Y} = E[Y|X]$. Suppose the joint pdf of X and Y is

$$p(x, y) = \begin{cases} cx & \text{if } (x, y) \in A \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where A is a region given by $x \geq 0, y \geq 0, x + y \leq 1$. Sketch A and find $E[Y|X]$.

Problem 6

For this problem, we provided you with two data sets, one for training and one for testing. In both cases, the data comes as (x, y) pairs (x is augmented).

a) Apply linear regression to the training data and this will produce a weight vector w (augmented). Using w, we can define a linear function $y = f(x) = w^T x$ (x is augmented). Is this a reasonably good model for the training data? (You can calculate the mean square error of this model on the training data and/or plot this function and the training data together).

b) See if the $f(x)$ (i.e., w) you find in part a) performs well on testing data.

Problem 7

Repeat Problem 6 on the data set we provided for this problem (Problem 6). Does the linear regression still work well? If yes, why? If not, why not?

Problem 8

Repeat Problem 7 using a neural network. Is your result better than that of Problem 7?

Problem 9

In this problem, we provided you with a 2-dimensional data set that contains three clusters. The "true" cluster centers are at $(0,0),(1,\sqrt{3}),(-1,\sqrt{3})$ (this is how we generated the data).

a) Perform clustering using the K-means algorithm and display the resulting clustering centers along with the data.

b) Repeat with the EM algorithm and compare your results with part a. What are your observations?