COM 525000 Statistical Learning
Homework #4

(Due January 11, 2020 before noon to the TA at EECS 613.)

Note: Detailed derivations are required to obtain a full score for each problem. (Total 100%)

1. (6%+8%+8%+12%) Let us consider the dataset $\mathcal{D} = \{(x_{i1}, x_{i2}, y_i)\}_{i=1}^{8} = \{(\frac{-1}{2}, \frac{-1}{2}, 4), (\frac{-3}{4}, 0, 2), (\frac{1}{2}, \frac{1}{2}, 0), (\frac{1}{2}, \frac{-1}{2}, 10), (\frac{3}{2}, 0, 6), (\frac{3}{2}, \frac{3}{2}, 4), (\frac{5}{2}, 2, 5), (0, 2, 15)\}$, each data point has two predictors as input and a single output label.

(a) Determine the regression tree with 4 splits.

(b) Following (a), find the sequence of subtrees obtained by the weakest link pruning algorithm.

(c) Suppose that the solution in (a) is the tree obtained in the first iteration of the boosting algorithm in Algorithm 8.2 of the text book, where $\lambda = 0.5$. Find the tree (with 3 splits) in the second iteration of the boosting algorithm.

(d) Suppose that bagging is performed on two bootstrapped datasets $\mathcal{D}^{*1} = \{(\frac{-1}{2}, \frac{-1}{2}, 4), (\frac{1}{2}, \frac{1}{2}, 0), (\frac{3}{2}, 0, 6), (\frac{5}{2}, 2, 5), (\frac{-1}{2}, \frac{-1}{2}, 4), (\frac{1}{2}, \frac{1}{2}, 0), (\frac{3}{2}, 0, 6), (\frac{5}{2}, 2, 5)\}$ and $\mathcal{D}^{*2} = \{(\frac{-3}{4}, 0, 2), (\frac{1}{2}, \frac{-1}{2}, 10), (\frac{3}{2}, 0, 6), (\frac{3}{2}, \frac{3}{2}, 4), (\frac{5}{2}, 2, 5), (0, 2, 15), (\frac{-3}{4}, 0, 2), (\frac{3}{2}, 0, 6)\}$. Find the estimated output label of input $(X_1, X_2) = (1, 1)$, and compute the out-of-bag error estimate.

2. (4%+12%+6%+8%) Suppose you are given 6 one-dimensional points: 3 with negative labels $x_1 = -1$, $x_2 = 0$, $x_3 = 1$ and 3 with positive labels $x_4 = -3$, $x_5 = -2$, $x_6 = 3$. Note that the data cannot be perfectly separated in $\mathbb{R}$, but, by applying the following feature map $\phi(x) = (x, |x|)$ (which transforms points in $\mathbb{R}$ to points in $\mathbb{R}^2$), a linear separator can perfectly separate the points in the new $\mathbb{R}^2$ feature space induced by ϕ.

(a) Find the kernel $K(x, x')$ that corresponds to the feature map ϕ.

(b) Construct a maximal margin separating hyperplane in the new feature space. This hyperplane will be a line in $\mathbb{R}^2$, which can be parameterized by its normal equation, i.e. $w_1Y_1 + w_2Y_2 + c = 0$ for appropriate choices of w_1, w_2, and c. Here, $(Y_1, Y_2) = \phi(X)$ is the result of applying the feature map ϕ to the original feature X. Plot the transformed points, plot the separating hyperplane and the margin, and circle the support vectors. Also, compute the values for w_1, w_2, and c, and the margin for your hyperplane. (Hint: The solution is evident by observing the points in the $\mathbb{R}^2$-plane. You do not need to solve any optimization problem explicitly.)

(c) Draw the decision boundary of the separating hyperplane in the original $\mathbb{R}$ feature space.

(d) Suppose that the margin expands by 50% without changing the hyperplane. Find the total proportion of violations. Is it possible to find a new hyperplane that yields less total proportion of violations? If so, please give an example.

3. (8%+8%+8%) Suppose that there are $n = 6$ observations, each with $p = 2$ features, given by $(X_1, X_2) = (1,1), (2,0), (2,3), (2,1), (3,1), (3,2)$.

(a) Cluster the observations into $K = 2$ clusters by performing K-means clustering. Initialize by taking the first 3 observations as the first cluster and the other 3 observations as the second cluster. Plot the observations and their cluster labels at the initialization and after each iteration until convergence.

(b) Perform hierarchical clustering using complete linkage and the squared difference distance measure $d_{i,i'} \triangleq \|x_i - x_{i'}\|^2$. Sketch the resulting dendrogram and indicate on the plot the height at which each fusion occurs, as well as the observations corresponding to each leaf in the dendrogram.

(c) Find the first principal component and the proportion of variance explained (PVE) by this component.

4. (12%) Consider a special case of the Gaussian mixture model in which the covariance matrices $\mathbf{\Sigma}_k$ of the components are all constrained to have a common value $\mathbf{\Sigma}$. Derive the EM equations for maximizing the likelihood function under such a model.