Statistical Machine Learning
Statistics GR 5241 — Spring 2022

Homework 2

Collaboration policy: Collaboration on solving the homework is allowed, after you have thought about the problems on your own. It is also OK to get clarification (but not solutions) from books or online resources, again after you have thought about the problems on your own. There are two requirements: first, cite your collaborators fully and completely (e.g., "Jane explained to me what is asked in Question 3 (a)"); second, write your solution *independently*: close the book and all of your notes, and send collaborators out of the room, so that the solution comes from you only.
The following problems are due on Friday, March 4th, 11:59pm.

1. **Decision Tree.**

   (a) (8 points) Suppose X and Y are discrete variables. Let IG be the information gain and H be the entropy. Show that $IG(X;Y) = H(X,Y) - H(X|Y) - H(Y|X)$.

   (b) The following is a small synthetic data set where we try to predict the usage of individual mobile phones based on their income, age, education, and marital status. In this section, you can assume that the decision tree is built using the ID3 algorithm, where each attribute is used only as an internal node.

Income	Age	Education	Martial Status	Usage
Low	Old	University	Married	Low
Medium	Young	College	Single	Medium
Low	Old	University	Married	Low
High	Young	University	Single	High
Low	Old	University	Married	Low
High	Young	College	Single	Medium
Medium	Young	College	Married	Medium
Medium	Old	High School	Single	Low
High	Old	University	Single	High
Low	Old	High School	Married	Low
Medium	Young	College	Married	Medium
Medium	Old	High School	Single	Low
High	Old	University	Single	High
Low	Old	High School	Married	Low
Medium	Young	College	Married	Medium

i. (2 points) What is the initial entropy of Usage?

ii. (5 points) Which attribute should be chosen at the root of the tree? Show your calculation for the information gains (IG) and explain your choice in a sentence.

iii. (5 points) Draw the full decision tree for the data.

(c) (5 points) Occam's Razor can be interpreted as simpler hypotheses are generally better than the complex ones. Does ID3 follow Occam's Razor? How about C4.5? Explain briefly.

2. **Nonlinear Decision Boundary**

We have seen that in $p = 2$ dimensions, a linear decision boundary takes the form

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0.$$

We now investigate a non-linear decision boundary.

(a) (3 points) Sketch the curve

$$(1 + X_1)^2 + (2 - X_2)^2 = 4.$$

(b) (3 points) On your sketch, indicate the set of points for which

$$(1 + X_1)^2 + (2 - X_2)^2 > 4,$$

as well as the set of points for which

$$(1 + X_1)^2 + (2 - X_2)^2 \leq 4.$$

(c) (3 points) Suppose that a classifier assigns an observation to the blue class if

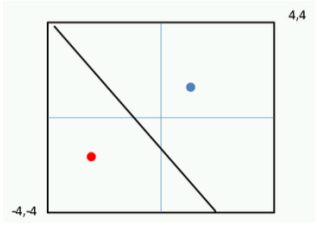
$$(1 + X_1)^2 + (2 - X_2)^2 > 4,$$

and to the red class otherwise. To what class is the observation $(0, 0)$ classified? What about $(-1, 1)$, $(2, 2)$ or $(3, 8)$?

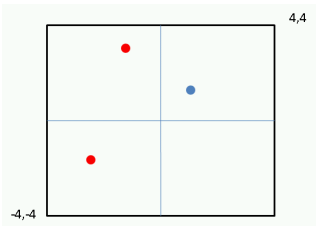
(d) (3 points) Argue that while the decision boundary in (c) is not linear in terms of X_1 and X_2, it is linear in terms of X_1, X_1^2, X_2 and X_2^2.

3. **KNN v.s. SVM.**

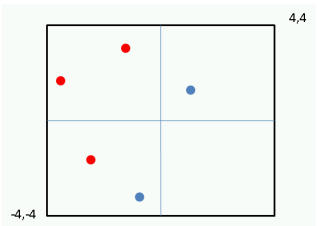
For each of the following figures, we are given a few data points in 2-d space, each of which is labeled as either positive (blue) or negative (red). Assuming that we are using L2 distance as a distance metric, draw the decision boundary for 1-NN for each case. In other words, with your decision boundary, the new test data can be classified into corresponding categories. As an example, we draw the decision boundary for you with figure (a).



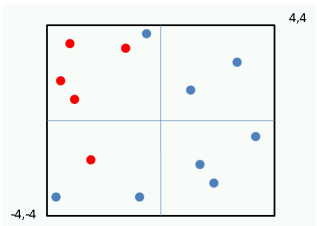
(a)



(b)



(c)



(d)

(a) (6 points) Assuming that we are using L_2 distance as a distance metric, draw the decision boundary for 1-NN for each case. In other words, with your decision boundary, the new test data can be classified into corresponding categories. As an example, we draw the decision boundary for you with figure (a).

(b) (6 points) Instead of 1-NN we now use SVM for each case, allowing only 1 miss-classification. Is the data linear separable? How will the decision boundary change from 1-NN to SVM.

(c) (6 points) What kind of kernel can we apply to Figure (d) such that the blue and red points are linearly separable?

4. **SVM for Regression.** Let $x \in \mathbb{R}^d$ be the feature vector and $y \in \mathbb{R}$ be the label. In this question, we use a linear predictor for the label, i.e. given the feature vector, we predict the label by

$$\hat{y}(x) = w^\top x$$

where $w \in \mathbb{R}^d$ is the linear coefficient vector. In this question, we consider the **epsilon insensitive loss function**, defined by

$$L_\epsilon(y, \hat{y}) = \begin{cases} 0 & \text{if } |y - \hat{y}| < \epsilon \\ |y - \hat{y}| - \epsilon & \text{otherwise.} \end{cases}$$

where ϵ is a tuning parameter. To obtain a good w, we would like to solve the following optimization:

$$J(w) = \frac{1}{n} \sum_{i=1}^{n} L_\epsilon(y, \hat{y}(x_i)) + \lambda \|w\|_2^2. \tag{1}$$

(a) (5 points) When $\epsilon = 0$, is the loss function the same as the absolute error loss? Show why. What role does ϵ play?

(b) (10 points) Notice that (1) is not a differentiable objective. Show how to convert (1) into a optimization problem whose objective is differentiable and constraints are linear by introducing slack variables.

5. **Comparison of Different Classifiers.**

In this part, we are going to explore the famous iris data set. There are features described in the first 4 columns as

$$X_1 = \text{sepal length in cm},$$
$$X_2 = \text{sepal width in cm},$$
$$X_3 = \text{petal length in cm},$$
$$X_4 = \text{pental width in cm}.$$

There are 3 different iris classes, Setosa, Versicolor as well as Virginica. We are going to implement decision tree, soft threshold SVM, Naive Bayes, as well as the AdaBoost algorithm on the iris dataset and compare the reults.

(a) (3 points) Based on the dataset, there are 3 different iris classes, Setosa, Versicolor as well as Virginica. Find a represent figure online for each different iris class.

(b) (4 points) For each feature, i.e. X_1, X_2, X_3, X_4, calculate the correlation between X_i and Y. Which feature you will discard the first and why?

(c) (3 points) Normalized X_i for $i = 1, 2, 3, 4$. Scatter plot (X_1, X_3), (X_1, X_4) and (X_3, X_4). Are the 3 different classes linearly separable?

(d) (20 points) Perform the Decision Tree classifier without constraint on the maximum depth, the KNN classifier with $K = 3$, the SVM classifier with $\gamma = 2$ and $C = 1$, the Naive Bayes classifier, and the AdaBoost with 30 Decision Tree classifiers with maximum depth 3. Plot the $3 \times 5 = 15$ decision surfaces of the 5 classifiers trained on these 3 pairs of features of the iris dataset and report $5 \times 3 = 15$ scores respectively. Compare and summarize the advantage and disadvantage of each classifier, together with the figures you find in part (a).