Xinwei Sun
Pattern recognition and machine learning
November 19, 2024

Homework: Probability (Deadline: 2024-12-05)

Problem 1. Please state the definition of the linear function, linear model. Which basis functions can we choose for a linear model?

Problem 2. Suppose we have data $\{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)\}$, and we would like to use linear regression.

- Please give the empirical risk. Is this risk function convex?
- Please introduce the gradient descent algorithm.

Problem 3. Suppose we have a data set with five predictors, $X_1 = \text{GPA}$, $X_2 = \text{IQ}$, $X_3 = \text{Gender}$ (1 for Females and 0 for Male), $X_4 = \text{Interaction between GPA and IQ}$, and $X_5 = \text{Interaction between GPA and Gender}$. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, $\hat{\beta}_3 = 35$, $\hat{\beta} = 0.01$, $\hat{\beta}_5 = -10$.

- Which answer is correct, and why?
  1) For a fixed value of IQ and GPA, males earn more on average than females.
  2) For a fixed value of IQ and GPA, females earn more on average than males.
  3) For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.
  4) For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.
- Predict the salary of a female with IQ of 110 and a GPA of 4.0.

Problem 4. Suppose we have data $\{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)\}$, please introduce the algorithm of KNN classifier, the perceptron, and the logistic regression.

Problem 5. Consider the Gini index, classification error, and cross-entropy in a simple classification setting with two classes. Create a single plot that displays each of these quantities as a function of $\widehat{p}_{m1}$. The x-axis should display $\widehat{p}_{m1}$, ranging from 0 to 1, and the y-axis should display the value of the Gini index, classification error, and entropy. Here, $\widehat{p}_{mk}$ represents the proportion of training observations in the mth region that are from the kth class.

Problem 6. Here we explore the maximal margin classifier on a toy data set. Suppose we are given $n = 7$ observations in $p = 2$ dimensions. For each observation, there is an associated class label, as illustrated in Table 1.

Obs.	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue

FIGURE 1. Table of problem 6.

- Sketch the optimal separating hyperplane, and provide the equation for this hyperplane. A hyperplane is defined by the equation:

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

- Describe the classification rule for the maximal margin classifier. It should be something along the lines of "Classify to Red if $\beta_0 + \beta_1 X_1 + \beta_2 X_2 > 0$, and classify to Blue otherwise." Provide the values for β_0, β_1, and β_2.
- Indicate the support vectors for the maximal margin classifier.

Problem 7. In this problem, you will generate simulated data, and then perform PCA and K-means clustering on the data.

(a) Generate a simulated data set with 20 observations in each of three classes (i.e. 60 observations total) and 50 variables.

1) Perform PCA on the 60 observations and plot the first two principal component score vectors. Use a different color to indicate the observations in each of the three classes. If the three classes appear separated in this plot, then continue to part (c). If not, then return to part (a) and modify the simulation so that there is greater separation between the three classes. Do not continue to part (c) until the three classes show at least some separation in the first two principal component score vectors.
2) Perform K-means clustering of the observations with K = 3. How well do the clusters that you obtained in K-means clustering compare to the true class labels?