PH240C Homework 1

Due: Septerm 22nd in class

September 9, 2021

1. Suppose we fit logistic regression to predict the probability a PH240C student gets an A in the class, from two variables. The variable are average hours of study per week (X_1) and GPA in other statistics courses taken (X_2). The model estimates $\beta_0 = -4, \beta_1 = 0.05, \beta_2 = 1$. (Note: these are made up numbers! Do not try to predict your grades with them. 10 points per question.)

   (a) Predict the probability of getting an A for a student who studies 5 hours a week and has a GPA of 3.5 in other statistics courses.

   (b) What are the odds that this student will get an A?

   (c) How many hours a week does this student need to study for the model to predict a 50% chance of getting an A?

2. For this question, please use graph paper; you can print it for free from many websites, for example, www.printfreegraphpaper.com. (15 points per question)

   (a) Draw the hyperplane defined by $2X_1 - 2X_2 - 1 = 0$. Indicate the set of points satisfying $2X_1 - 2X_2 - 1 > 0$ with a "+" sign, and the set of points satisfying $2X_1 - 2X_2 - 1 < 0$ with a "−" sign.

   (b) Suppose your hyperplane is the optimal separating hyperplane for an SVM classifier fitted to some data, with the margin $m = \sqrt{2}$. Draw the margin lines.

   (c) What class label (+ or −, as defined above) does this SVM predict for the following points: (1,4); (1,1); (2, -5); (2, -1); (4,2)?

   (d) Suppose these five points were part of the training data, and their true labels, given in the same order, are −, −, +, +, −. Calculate the corresponding slack values (ξ_i's) for each of the five points.

3. Using the dataset "heart_disease.csv" which contains a dataset of heart disease patients from the Cleveland Clinic:

   - Chest-pain type. Type of chest-pain experienced by the individual: 1 = typical angina, 2 = atypical angina, 3 = non-angina pain, 4 = asymptomatic angina
   - Fasting Blood Sugar. Fasting blood sugar level relative to 120 mg/dl: 0 = fasting blood sugar $\leq$120 mg/dl, 1 = fasting blood sugar >120 mg/dl
   - Resting ECG. Resting electrocardiographic results: 0 = normal, 1 = ST-T wave abnormality, 2 = left ventricle hyperthrophy
   - Exercise Induced Angina: 0 = no, 1 = yes

- ST Depression Induced by Exercise Relative to Rest: ST Depression of subject
- Peak Exercise ST Segment: 1 = Up-sloaping, 2 = Flat, 3 = Down-sloaping
- Thal: Form of thalassemia: 3 = normal, 6 = fixed defect, 7 = reversible defect
- Diagnosis of Heart Disease: Indicates whether subject is suffering from heart disease or not: 0 = absence, 1 = heart disease present

The goal is to learn a binary classifier for hear disease using support vector machines. Please consider at least three different choices of kernel function (linear, Gaussian, polynomial) to produce the corresponding classifiers. What are important predictors in the model? How would you quantify them?