

## Homework Assignment #2

February 14, 2021

**Problem 1:** (20 points)

Consider a binary classification problem where there is a single feature  $X \in \mathbb{R}$  and the dependent variable  $Y \in \{0, 1\}$ . Let  $P_{X,Y}$  denote the joint distribution over pairs  $(X, Y)$ , and let  $h : \mathbb{R} \rightarrow \{0, 1\}$  denote a generic *classifier*. We define the error rate of  $h$  as  $R(h) := \Pr(Y \neq h(X))$ , where the probability is computed from the joint distribution  $P_{X,Y}$ . Suppose that we collect data  $(x_1, y_1), \dots, (x_n, y_n)$ , which are assumed to be independent and identically distributed from the distribution  $P_{X,Y}$ , and that we train a classifier  $\hat{h}$  using this data.

Below we consider two precise specifications of the joint distribution  $P_{X,Y}$ . In both cases, derive the largest numerical value  $\varepsilon^*$  that you can for which it holds that  $R(\hat{h}) \geq \varepsilon^*$ . Carefully explain how you arrived at your specific value of  $\varepsilon^*$  in both cases.

- a) (10 points)  $X$  is restricted to the set  $\{0, 1\}$  (i.e., is categorical) and  $P_{X,Y}$  is given by the distribution in Table 1.
- b) (10 points) The marginal distribution of  $Y$  is given by  $\Pr(Y = 0) = 0.4$  and  $\Pr(Y = 1) = 0.6$ . Given that  $Y = 0$ , the distribution of  $X$  is normal with mean 3 and variance 1. Given that  $Y = 1$ , the distribution of  $X$  is normal with mean  $-1$  and variance 1.

Table 1

Outcome	Probability
$(X = 0, Y = 0)$	0.15
$(X = 0, Y = 1)$	0.25
$(X = 1, Y = 0)$	0.4
$(X = 1, Y = 1)$	0.2

**Problem 2:** (10 points)

Consider a binary classification problem where  $X \in \mathbb{R}^p$  and  $Y \in \{0, 1\}$ . For a fixed  $x \in \mathbb{R}^p$ , suppose that  $\Pr(Y = 1 \mid X = x) = p$  for some  $p \in [0, 1]$ . Consider a prediction problem where there is a loss  $L_{FN} > 0$  associated with predicting  $Y = 0$  when the actual outcome is  $Y = 1$ , and another loss  $L_{FP} > 0$  associated with predicting  $Y = 1$  when the actual outcome is  $Y = 0$ . There is no loss associated with true positives or true negatives (i.e., a correct answer). Show that there is a threshold value  $\bar{p}$  such that following the expected loss criterion for making a prediction is equivalent to predicting  $Y = 1$  if  $p \geq \bar{p}$  and predicting  $Y = 0$  otherwise. What is the value of  $\bar{p}$ ?

**Problem 3: Framingham Heart Study (Adapted from Bertsimas Chapter 7)** (70 points)

Heart disease is one of the leading causes of death worldwide. Over 8 million people died from coronary heart disease (CHD) in 2019, which was the leading cause of death that year. (For 2020, it is estimated that COVID-19 has overtaken heart disease as the leading cause of death globally.)

In the late 1940s, the U.S. government took steps to study cardiovascular disease. In order develop high quality data for their study, they decided to track a large cohort of initially-healthy people over time. The town of Framingham, Massachusetts (a suburb of Boston) was selected as the site for the study, which commenced in 1948. The study enrolled 5,209 participants aged 30-62. Participants were given a questionnaire and a medical exam every two years. They also collected data on the participants' physical characteristics and behavioral characteristics, in addition to the medical test data. Over the years, the study has expanded to include multiple generations and has collected many more factors including genetic information. This data is now famously known and is simply called the Framingham Heart Study.

In this exercise, you are asked to build models using Framingham Heart Study data in order to predict CHD and to make recommendations to better prevent heart disease. There are 3,658 total observations in our data, with each observation representing the data from a particular study participant. There are 16 variables in the dataset, which are described in Table 2. You will be asked to predict **TenYearCHD** (whether the patient experiences coronary heart disease within 10 years of their first examination). As a consequence of your modeling efforts, you should be able to identify **risk factors**, which are the variables that increase the risk of CHD.

- a) (40 points) To lower the risk of CHD, physicians can prescribe preventive medication such as blood-pressure-lowering or cholesterol-lowering medications. Many policy makers, when recommending certain preventive medications to patients at risk of developing CHD, rely on evidence-based analysis that weighs the pros and cons of such interventions. Health economic evaluation is a commonly applied methodology for decision-making that takes both medical costs and health benefits (a monetized version of improved life longevity) into consideration. In fact, many countries establish clinical practice guidelines using such formalized health economic evaluation methodologies (the National Institute for Health and Clinical Excellence in England, for example).

As prior work, let us suppose that a colleague of yours has completed a health economics study analyzing the costs and benefits of a recently approved medication aimed at preventing CHD. The colleague determined that patients who experience CHD within the next 10 years are expected to incur a lifetime cost of \$450,000 associated with the disease; this cost includes

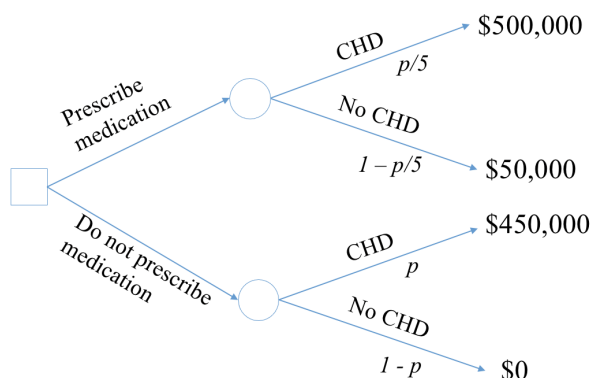
Table 2: Variables in the dataset `framingham.csv`.

Variable	Description
<code>male</code>	Is biological sex male
<code>age</code>	Age (in years) at first examination
<code>education</code>	Some high school, high school/GED, some college/vocational school, college
<code>currentSmoker</code>	Is a current smoker
<code>cigsPerDay</code>	Number of cigarettes per day
<code>BPMeds</code>	Is on blood pressure medication at time of first examination
<code>prevalentStroke</code>	Previously had a stroke
<code>prevalentHyp</code>	Currently hypertensive
<code>diabetes</code>	Currently has diabetes
<code>totChol</code>	Total cholesterol (mg/dL)
<code>sysBP</code>	Systolic blood pressure
<code>diaBP</code>	Diastolic blood pressure
<code>BMI</code>	Body Mass Index, weight (kg)/height (m) <sup>2</sup>
<code>heartRate</code>	Heart rate (beats/minute)
<code>glucose</code>	Blood glucose level (mg/dL)
<code>TenYearCHD</code>	Experienced coronary heart disease within 10 years of first examination

both the costs of treatment for CHD, \$150,000, as well as a cost intended to capture the decreased quality and length of life experienced by patients with CHD, which is \$300,000. Your colleague has determined that patients who take the preventative medicine being studied will have their probability of developing CHD within the next 10 years reduced by 80%; that is, if their current 10-year risk of developing CHD is  $p$  without taking the medication, then their 10-year risk with the medicine would instead be  $p/5$ . Regardless of whether a patient eventually develops CHD, there is a \$50,000 cost associated with taking this recently approved medication. A decision tree capturing your colleague's analysis is shown in Figure 1.

Using all of the provided independent variables, build a logistic regression model to predict the probability that a patient will experience CHD within the next 10 years. Use dataset `framingham_train.csv` to train your model. This training set has 2560 data points, which are randomly selected from the original `framingham.csv` dataset (around 70%). Use dataset `framingham_test.csv` to test your model. This test set has the remaining 1098 data points.

Figure 1: Decision tree for prescribing the approved medication to prevent CHD. The leaf nodes represent cost values.



Please answer the following questions concerning your model.

- i)* What is the fitted logistic regression model? Do not simply copy the results of your code, but instead state the equation used by the model to make predictions.
- ii)* What are the most important risk factors for 10-year CHD risk identified by the model? Pick one of these variables and describe its impact on a patient's predicted odds of developing CHD in the next 10 years.
- iii)* Suppose that you wish to determine the optimal strategy for assigning which new patients receive the medication. Given your colleague's analysis of the costs and benefits associated with the recently approved treatment, identify a threshold value of  $p$ , call it  $\bar{p}$ , such that it is optimal to prescribe the medication to a patient if and only if their 10-year CHD risk exceeds  $\bar{p}$ .
- iv)* Describe the test set performance of the logistic regression model, using the threshold identified in part (*iii*) to separate patients into those who are at high risk for CHD (risk exceeding the threshold  $p$ ) and those who are at low risk for CHD (risk below the threshold  $p$ ). State the model's accuracy, True Positive Rate (TPR), and False Positive Rate (FPR), and briefly describe these three metrics in a way that is accessible to a non-technical audience.
- v)* If patients are prescribed the medication using the strategy implied by the model, use the test set data to provide an estimate(s) for the expected economic cost per patient. You should first report your estimate assuming that the CHD outcomes in the test set are not affected by the treatment decision. Is this assumption reasonable? You should then adjust your estimate in a way that takes into account the fact that the treatment decision impacts a patient's risk of developing CHD. (Hint: keep in mind that this dataset was collected before the option of prescribing the medication was even considered.)
- vi)* Consider a simple baseline model that predicts none of the patients are at high risk for CHD and therefore does not recommend treatment for any of the patients. Describe the

test set performance of the baseline model in terms of accuracy, TPR, and FPR, as well as expected economic cost per patient.

- vii) Use an example to explain how to use the model in a real clinical setting. Suppose a new patient arrives, and the physician accesses the patient's electronic medical records and retrieves the following about the patient:

Male, age 49, Some high school, currently a smoker with an average of 10 cigarettes per day. Not on blood pressure medication, has not had stroke, but has hypertension. Not diagnosed with diabetes; total Cholesterol at 197. Systolic/diastolic blood pressure at 140/100, BMI at 28, heart rate at 62, glucose level at 76.

What is the predicted probability that this patient will experience CHD in the next ten years? Based on the threshold from the decision tree, should the physician prescribe the preventive medication for this patient?

- b) (15 points) Show the ROC curve for your logistic regression model on the test set and describe how this curve may be helpful to decision-makers looking to further study the medication you have considered so far in this homework as well as other possible medications for preventing CHD. Describe one interesting observation implied by examining the ROC curve. What is the area under the curve (AUC) for your model in the test set?
- c) (10 points) Rather than explicitly dictating which patients should receive the medication, let us consider letting patients decide for themselves. Suppose that if a patient has health insurance, the treatment costs for CHD (including the proposed medication) will be covered by their insurance company. However, a patient will still incur an equivalent cost of \$300,000 for decreased quality of life if they develop CHD. Disregarding other factors such as side effects of the medication, if there were no insurance co-payment then it should be clear that every patient would always choose to receive the medication because it would cost them nothing and it would lower their risk of CHD. Thus let us consider setting a co-payment value  $C$  – the amount that each patient would have to pay in order to receive the medication – in order to provide an incentive for some patients to forego the treatment while others would choose to receive the treatment. What value of  $C$  should the insurance company charge as a co-payment for the medication in order that the patients would “self select” in a manner that is consistent with the previously examined “optimal strategy” discussed in part (a) above?
- d) (5 points) Are there any aspects of the analysis performed thus far that raise ethical concerns? If so, suggest at least one way that this analysis could be changed to address such concerns.