INDENG 242A: Homework 2

Due October 17, 2025, 11:59pm

Submit your homework (typed or scanned) in a single pdf file on GradeScope. Justify your answers for full credit and state clearly any assumptions you make – some problems may be intentionally slightly vague (what assumptions are reasonable?). If you use AI tools for anything other than debugging code, you must declare this clearly and include as an appendix to your HW a clear description of what you used it for and how (you may just copy paste the chat if you are using an LLM).

Q1. Framingham Heart Study (80 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.

In the late 1940s, the U.S. government took steps to study cardiovascular disease. In order to 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. The dataset is in the file `framingham.csv`.

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 1. 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) (50 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 $955,000 associated with the disease; this cost includes both the costs of treatment for CHD, $330,000, as well as a cost intended to capture the decreased quality and length of life experienced by patients with CHD, which is $625,000. Also, your colleague has determined that patients who take the preventative medicine being studied will have their probability of developing CHD

Table 1: Variables in the dataset `framingham.csv`.

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

within the next 10 years reduced by 90%; in other words, if their current 10-year risk (probability) of developing CHD is p without taking the medication, then their 10-year risk (probability) with the medicine would instead be $(0.10 * p)$. Regardless of whether a patient eventually develops CHD, there is a $95,000 cost associated with taking this recently approved medication. A decision tree capturing your colleague's analysis is shown in Figure 1 (below).

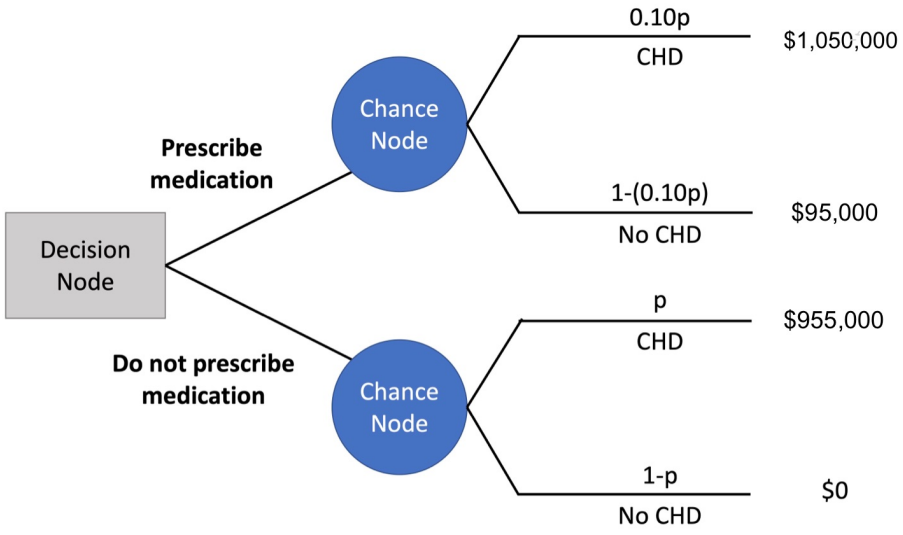
Using all of the provided independent variables, build a logistic regression model to predict the probability that a patient will experience CHD within the next 10 years. Use dataset **`framingham_train.csv`** to train your model. This training set has 2560 data points, which are randomly selected from 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. Please answer the following questions concerning your model:

Use this model as well as relevant details from your colleague's analysis to make recommendations to a policy maker who is familiar both with CHD and with your colleague's analysis, but who does not have a strong background in statistics. Write a brief memo that answers/addresses the following items:

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. **Use all features from Table 1 to build your model.**

ii) What are the most important risk factors for 10-year CHD risk identified by the model? Pick one

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



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 *iii*) to separate patients into those who are at high risk for CHD (risk exceeding the threshold $\bar{p}$) and those who are at low risk for CHD (risk below the threshold $\bar{p}$). 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:

```
Female, age 39, GED education, currently a smoker with an average of 6 cigarettes
per day.  Currently not on blood pressure medication, has not had stroke and
```

`is not hypertensive. Currently diagnosed with diabetes; total Cholesterol at 230. Systolic/diastolic blood pressure at 110/50, BMI at 28, heart rate at 72, glucose level at 80.`

What is the predicted probability that this patient will experience CHD in the next ten years? Based on your calculated $\bar{p}$ threshold from *iii*) 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 $625,000 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.

Q2. Support Vector Machines (20 points)

Recall the **MNIST** dataset of handwritten digits, which we have seen in Discussion 4. **Pick your favorite two digits (say 1 and 8)**, and classify them with the following models:

(a) SVM with linear boundary;

(b) Kernel SVM with polynomial kernels of degree 3, i.e, $K(x, z) = (1 + x^\top z)^3$;

(c) Kernel SVM with radial basis function (RBF) kernels.

To download the image data and transform them into vectors, please refer to discussion 4 material *Image Classification.ipynb* (you can find it on bCourses). For each method, you should first train the model on the training dataset, and then test its performance on the test dataset. Report the test accuracy and AUC.

*(Kernel) SVMs have been implemented in "sklearn.svm.SVC" in the "sklearn" library. Please refer to the Python notebook of Discussion 5 for more details.