CS 6140: Machine Learning — Fall 2021— Paul Hand

HW 4 Revised (corrected hyperlinks)
Due: Wednesday October 13, 2021 at 2:30 PM Eastern time via Gradescope.

Names: [Put Your Name(s) Here]

You can submit this homework either by yourself or in a group of 2. You may consult any and all resources. You may submit your answers to this homework by directly editing this tex file (available on the course website) or by submitting a PDF of a Jupyter or Colab notebook. When you upload your solutions to Gradescope, make sure to tag each problem with the correct page.

Question 1. *In this problem, you will use logistic regression for heart attack prediction.*

Download the dataset at this Kaggle site. It is a CSV file of 14 attributes of 294 people. The meaning of the attributes is as follows:

- age: age in years
- sex: sex (1 = male; 0 = female)
- cp: chest pain type – Value 1: typical angina – Value 2: atypical angina – Value 3: non-anginal pain – Value 4: asymptomatic
- trestbps: resting blood pressure (in mm Hg on admission to the hospital)
- chol: serum cholestoral in mg/dl
- fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- restecg: resting electrocardiographic results – Value 0: normal – Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) – Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
- thalach: maximum heart rate achieved
- exang: exercise induced angina (1 = yes; 0 = no)
- oldpeak = ST depression induced by exercise relative to rest
- slope: ignore
- ca: ignore
- thal: ignore
- num: diagnosis of heart disease (angiographic disease status) – Value 0: < 50% diameter narrowing – Value 1: > 50% diameter narrowing

You will train binary classifiers to predict the diagnosis of heart disease using some or all of the following features: age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak. You may find helpful the following step-by-step guide.

(a) Randomly select a test dataset consisting of 20% of the examples with num= 0 and 20% of the examples with num= 1. The remaining examples will constitute the training data. Plot histograms of each of the features for the training data and the test data.

Response:

(b) Use logistic regression to learn a binary classifier that predicts the diagnosis of heart disease using only the features: age, sex, cp, chol. Plot the ROC curve and the precision-recall curve for your classifier. Remove from your training set any example for which any of these features is missing. You may use an existing computer package that computes the logistic regression, such as scikit-learn. You may choose to follow other data cleaning steps in the step-by-step guide

Response:

(c) Same as part (b), but use the following features: age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak.

Response:

(d) Compare your two classifiers. Which would you argue is better for deployment in practice?

Response:

(e) Solve the logistic regression in part (b) using gradient descent and a cross-entropy loss. Plot the ROC curve and the precision-recall curve.

Response:

(f) Solve the logistic regression in part (b) using gradient descent and a square loss. Plot the ROC curve and the precision-recall curve. How does your classifier compare to that from part (e)?

Response: