CEE 690: Health and Environmental Data Science, Spring 2020

Homework 1: 100 Total Points

Due: February 5th by 11:59 PM

Submissions required electronically through Sakai

There are two major goals of this assignment:

1. Refresh ideas from probability and statistics to make the mathematical concepts discussed in the next set of lectures more familiar and solidify the math introduced in lectures 2-3 (problems 1-2).
2. Beginning building up the ability to code using the types of algorithms we talked about. This assignment will be using synthetic datasets so that we can focus on the algorithmic concepts, and we will begin moving to more complex problems (problems 3-4).

Problem 1: Refreshing Probability (25 points)

An important equation from probability is Bayes' theorem:

$$p(b|a) = \frac{p(a|b)p(b)}{p(a)}$$

Suppose that we have a test for lung cancer with the following properties:

- 95% sensitive (95% of true cases are predicted as positive), i.e. p(test is positive | patient is positive)=.95
- 95% specific (95% of negative cases are predicted as negative), i.e. p(test is negative | patient is negative)=.95

The overall rate of lung cancer is 0.07%, i.e. p(patient is positive)=.0007

(a) Suppose that give this test to a random individual. If the test is positive, what is probability that the individual is actually a positive case?

(b) We generally don't actually test random individuals; we typically test patients that are "at risk." Suppose we give the test to a patient in a clinic, where the probability of cancer is much higher than in a random individual. If p(patient is positive)=.1 instead, what is the probability that a patient is positive given that the test is positive?

(c) What is the accuracy of this test using p(patient is positive)=.0007?

Consider a second test with the following properties:

- 90% sensitive (90% of true cases are predicted as positive), i.e. p(test is positive | patient is positive)=.90
- 99.9% specific (99.9% of negative cases are predicted as negative), i.e. p(test is negative | patient is negative)=.999

(d) Repeat (a-c) with these properties.

(e) When does this test seem preferable to our first test?

Consider a third test that is:

- 10% sensitive
- 99.99% specific

(f) Under p(patient is positive)=.0007, what is the accuracy of this test?
(g) Under p(patient is positive)=.05, what is the accuracy of this test?
(h) Is accuracy a good metric for these tests? Why or why not?

This same type of Bayes' theorem calculations can be used to calculate what the number or reported false results in the literature are. For those interested, Ioannidis's 2005 article is highly recommended. Such calculations will get more critical as we get into statistical testing.

Problem 2: Understanding Properties of Model Losses (20 points)

Mean squared error is one of the most fundamental approaches to model fitting, so we want to understand some of its key properties.

(a) First, let's understand how this works in a single variable. Prove what value of a (as a function of data) will minimize the mean squared error (should be a review from undergraduate statistics, but may have been a while…):

$$\mathrm{a}^* = \arg\min \sum_{i=1}^{N} (x_i - a)^2$$

As a reminder, the minimum will occur where the derivative with respect to a is 0.

(b) Suppose that we have a statistical model with $p(x_i|a) = N(a, \sigma^2)$. Show that for any value of σ^2 that maximizing the log-likelihood will be equivalent to the minimizing the mean squared error above. Remember that monotonic transformations (including constants) do not change the locations of minima/maxima.

The logistic loss introduced in lecture 2 often seems an arbitrary loss as first glance, so we want to understand how this fits into a statistical model.

(c) As a reminder, the logistic or cross-entropy loss is:

$$\ell\big(y, \sigma(z)\big) = -y \log \sigma(z) - (1 - y) \log\big(1 - \sigma(z)\big)$$

Show how minimizing the logistic loss is the same as the negative log-likelihood for a Bernoulli random variable (a weighted coin flip):

$$\ell\big(y, \sigma(z)\big) = -\log Bernoulli\big(y; \sigma(z)\big),$$
$$Bernoulli(y; p) = p^y (1 - p)^{1-y}$$

(d) Show that minimizing $\sum_{i=1}^{N} \ell\big(y_i, \sigma(b \odot x_i)\big)$ is the same as the maximum likelihood solution for $Bernoulli\big(y_i; \sigma(b \odot x_i)\big)$ over all the data.

Problem 3: Basics of Model Selection in Classification (40 points)

In the Gitlab
(https://gitlab.oit.duke.edu/dec18/cee690_05_sp2020/blob/master/HW1_Template.ipynb

), there is a template file for homework 1 that produces 3 synthetic datasets. There are also .csv files if you would like to load these datasets into a non-Python format.

For the purposes of this assignment, you should split the dataset into a single training and validation dataset. Assume that I have a test set hidden away somewhere, and we won't worry about cross-validation until the next homework assignment.

Use *accuracy* as the metric to determine the best model for this assignment. We will be quickly talking about better approaches for binary classification.

(a) 1-d dataset on whether students pass a test at not.
   i. Visualize (plot) the dataset to get a sense of what's going on.
   ii. Fit a logistic regression to the dataset. What is the validation error?
   iii. Fit a *k*-Nearest Neighbors classifier to the dataset. Try different values of *k* to determine the optimal number of neighbors. What is the best validation error?
   iv. Visualize the predictions from logistic regression and kNN on the same plot
   v. What model do you prefer? Why?

(b) 2-d dataset that forms circles.
   i. Visualize (plot) the dataset to get a sense of what's going on.
   ii. Fit a logistic regression to the dataset. What is the validation error?
   iii. Fit a *k*-Nearest Neighbors classifier to the dataset. Try different values of *k* to determine the optimal number of neighbors. What is the best validation error?
   iv. What model do you prefer? Why?

(c) 2-d dataset that forms moons.
   i. Visualize (plot) the dataset to get a sense of what's going on.
   ii. Fit a logistic regression to the dataset. What is the validation error?
   iii. Fit a *k*-Nearest Neighbors classifier to the dataset. Try different values of *k* to determine the optimal number of neighbors. What is the best validation error?
   iv. What model do you prefer? Why?

Problem 4: How robust are your results to different validation splits? (10 points)

Looking forward, we want to understand some of the randomness that is inherent in our performance estimation, so we want to start building an intuition for that.

Using the three datasets from problem 3, create 5 different random splits.

(a) Fit a *k*-Nearest Neighbors classifier to the dataset. Try different values of *k* to determine the optimal number of neighbors. What is the best validation error on each different random split? What is the optimal number of neighbors in each split?
(b) How does this compare to what you expect?

Note: here you won't be able to predict very well; this is largely setting the scene for what happens as overfitting gets worse with higher dimensions and more complex modeling. We're going to repeat this as we go through the course.

Pseudo-Problem 5: Administrative (5 Points):

(a) (5) How many hours did this assignment take you? (There is **NO** correct answer here, this is just an information gathering exercise)

(b) (5) Verify that you adhered to the Duke Community Standard in this assignment (https://studentaffairs.duke.edu/conduct/about-us/duke-community-standard). (I.E. write "I adhered to the Duke Community Standard in the completion of this assignment"