Homework 8

Due Tuesday, November 17

Please either give the assignment to Loraine at the CDS or send it via email to the graders **before noon**.

1. *Probability of error vs MSE (10 points).* We are interested in estimating a signal

$$X = \begin{cases} +1 & \text{with probability } \frac{1}{2} \\ -1 & \text{with probability } \frac{1}{2}. \end{cases}$$

The measurement that we actually observe is $Y = X + Z$, where Z is a uniform random variable between -2 and 2.

a. Find the minimum MSE estimate g_{MSE} of X given Y and the corresponding mean square error.

b. What is the probability of error of g_{MSE}?

c. Suppose that we want to minimize the probability of error. Find the optimal decoder g_{error} and its probability of error.

d. Compare the MSE of g_{error} to the minimum MSE.

2. *Halloween parade (20 points)* The city of New York hires you to estimate whether it will rain during the Halloween parade. Checking past data you determine that the chance of rain is 20%. You model this with the random variable R with pmf

$$p_R(1) = 0.2, \qquad p_R(0) = 0.8,$$

where $R = 1$ means that it rains and $R = 0$ that it doesn't. Your first idea is to be lazy and just use the forecast of a certain website. Analyzing data from previous forecasts, you determine that this website is right 70% of the time. You model this with a random variable W that satisfies

$$\mathrm{P}(W = 1|R = 1) = 0.7, \qquad \mathrm{P}(W = 0|R = 0) = 0.7.$$

a. What is your prediction given the forecast of the website (use the MAP estimate of R given W)? What is the probability of error under your model?

Unsatisfied with the accuracy of the website, you look at the data used for the forecast (they are available online). Surprisingly the relative humidity of the air is not used, so you decide to incorporate it in your prediction in the form of a random variable H.

b. Is it more reasonable to assume that H and W are independent, or that they are conditionally independent given R? Explain why.

You assume that H and W are conditionally independent given R. More research establishes that conditioned on $R = 1$, H is uniformly distributed between 0.5 and 0.7, whereas conditioned on $R = 0$, H is uniformly distributed between 0.1 and 0.6. Use the MAP estimate of R given W and H as your forecast.

c. What is your forecast if $H = 0.65$ and the website predicts no rain?

d. What is your forecast if $H = 0.55$ and the website predicts rain?

e. What is the probability of error under this new model?

3. *Heart-disease detection (20 points).* A hospital is interested in developing a system for automatic heart-disease detection[1]. Your task is to use the data in the *heart_disease_data.npz*[2] to detect heart disease in patients. You decide to model the problem as an estimation problem, in which a random variable H that indicates whether the patient suffers from heart disease or not:

$$H = \begin{cases} 0 & \text{if patient does not suffer from heart disease,} \\ 1 & \text{if patient suffers from heart disease.} \end{cases}$$

The available data contain the patient's sex, the type of chest pain experienced by the patient and the cholesterol of the patient. We model these quantities as the random variables S, C and X respectively, where

$$S = \begin{cases} 0 & \text{if patient is female,} \\ 1 & \text{if patient is male,} \end{cases}$$

$$C = \begin{cases} 0 & \text{if the pain is typical angina,} \\ 1 & \text{if the pain is atypical angina,} \\ 2 & \text{for other types of chest pain,} \\ 3 & \text{if there is no chest pain,} \end{cases}$$

and X is a continuous random variable.

a. Derive a MAP estimate of H given S and C that only depends on the pmf of H (p_H) and the conditional pmfs $p_{S|H}$ and $p_{C|H}$. Assume that if we know whether a patient is suffering from heart disease, the sex of the patient and the type of chest pain experienced by the patient are independent.

b. Complete the corresponding part of the script *hw8pb3.py* to estimate the necessary probability mass functions from the data in the arrays *data["heart_disease"]*, *data["sex"]* and *data["chest_pain"]* which were compiled from 218 patients. Apply your detection rule to predict whether a group of 50 other patients, whose information is stored in the vectors *data["sex_test"]* and *data["chest_pain_test"]*, suffer from heart disease. Calculate the error rate of your prediction algorithm by comparing your results to *data["heart_disease_test"]*, which indicates whether the patients suffer from heart disease or not.

c. Derive a MAP estimate of H given S, C and X that only depends on the pmf of H p_H, the conditional pmfs $p_{S|H}(s|h)$ and $p_{C|H}$ and the conditional pdf $f_{X|H}$, assuming that if we know whether a patient is suffering from heart disease, the sex, type of chest pain and cholesterol level of the patient are all independent.

[1] A patient is deemed to suffer from heart disease if at least one of his or her major vessels is 50% narrower than it should be.

[2] The data in this problem, which was compiled from five hospitals in Hungary, Switzerland and the United States, is available at `http://archive.ics.uci.edu/ml/datasets/Heart+Disease`.

d. You decide to model the cholesterol level of a patient conditioned on whether he or she suffers from heart disease as a Gaussian random variable. For both cases, complete the corresponding part of the script *hw8pb3.py* to estimate the mean and variance by computing the sample mean and variance of the vector *cholesterol* and compare the estimated pdf to the histogram of the data.

e. Complete the corresponding part of the script *hw8pb3.py* to apply your MAP rule incorporating the cholesterol data and compute the new error rate (using the cholesterol rates of the 50 new patients, stored in *data["cholesterol_test"]*). Do you trust this result?

f. We have made some conditional independence assumptions that do not necessarily hold. Another option would have been to estimate the joint distribution of all the random variables from the data. Is this a good idea? Why? (Hint: The answer is not necessarily yes or no. Reason about what happens if we have a lot of data available or not.)

Note: In machine learning applying MAP detection with conditional-independence assumptions is known as Naïve Bayes.