CS6923 Machine Learning, Spring 2019
Prof. Sellie, NYU School of Engineering

Homework 1

Submit on NYU Classes by Fri. Sept. 13 at 8:00 p.m. Submit two files: (1) a **pdf** file with your written answers for Part I. (2) a **zip** file including code and answers for Part II. You do not have to typeset your written answers: as long as your handwriting is readable, you can write out the answers by hand and scan your answers. If you do that, use a utility like camscanner to make sure that the scan is clear.

You may work together with one other person on this homework. If you do that, *hand in JUST ONE homework for the two of you*, with both of your names on it. You may *discuss* this homework with other students but YOU MAY NOT SHARE WRITTEN ANSWERS OR CODE WITH ANYONE BUT YOUR PARTNER.

Part I: Written Exercises

1. A scientist studying crabs has 5 different species (labeled 1 through 5), with the following probabilities of finding them in a cave:

Label	**Probability**
1	25%
2	5%
3	20%
4	35%
5	15%

   (a) What is the probability of incorrectly labeling a crab if the scientist predicts the species uniformly at random?

   (b) Can the scientist improve his error by picking the same label every time? If so, what label should the scientist choose and what would the probability of being incorrect be?

2. You are a scientist studying two species of rock crabs. One type is *orange*, the other type is *blue*. You have a beautiful black and white photo you want to use in an article you are writing on the crabs you study, but you unfortunately don't remember if the crab was blue or orange.

   Suppose you know the measurements for two of the unknown crab's features: 'RearWidth' = 11.1, and 'CarapaceLength'= 27.8 (you had placed a ruler in the picture). From your research, you've determined that these features are approximately distributed according to a Gaussian. Your estimates for the mean and covariance of these features for both types of crab are as follows:

   - $\mu_{\text{blue}} = \begin{bmatrix} 29.42 \\ 33.98 \end{bmatrix}$ and $\Sigma_{\text{blue}} = \begin{bmatrix} 50.56 & 57.49 \\ 57.49 & 65.54 \end{bmatrix}$
   - $\mu_{\text{orange}} = \begin{bmatrix} 33.88 \\ 37.80 \end{bmatrix}$ and $\Sigma_{\text{orange}} = \begin{bmatrix} 46.79 & 52.19 \\ 52.19 & 58.59 \end{bmatrix}$

   Given the information above, which color crab is most likely? Justify your answer.

3. Consider a coin with unknown bias θ, where θ is the probability of Heads ($p(Heads) = \theta$). Suppose you flip a coin 5 times and get the following: HHTHH. (H is for Heads and T is for Tails.)

   (a) As a function of θ, what is $p(HHTHH)$?

   (b) As a function of θ, what is $\log p(HHTHH)$?
   Express your answer as a function of θ, in the form $a \log \theta + b \log(1 - \theta)$, for some constants a and b.

   (c) Find the *maximum likelihood estimate of* θ. That is, find $\arg\max_\theta p(HHTHH \mid \theta)$.
   Use the log trick! You must show your work.

Part II: Programming and Questions

4. Complete the lab in the python notebook `salmon-classification.ipynb` to be able to predict if a fish is from Alaska or Canada based on two features:

   - x_1 = diameter of rings for first-year freshwater growth (in hundredths of an inch)
   - x_2 = diameter of rings for first-year marine growth (in hundredths of an inch)

This question should be answered using the salmon dataset that was given in the file `Salmon_Dataset.csv`. The first line in the file contains the header information. Each subsequent line in the file gives the information for one example. Each example in the file has three features (and an index), but for this assignment you will only use two of the features.

You will model the data from the Alaskan salmon as a bipartite Gaussian where you will estimate the parameters μ_0 and $\mathbf{\Sigma}_0$ using maximum likelihood estimation. Similarly, you will do the same for the Canadian salmon.

You will then use this information to predict if the following fish are from Canada or from Alaska:

Freshwater	Marine
144	403
76	442
100	470
155	349
99	403
124	341
136	438
152	301
99	481
80	398

Outputs Your program should output the following values and write them to a separate text file (or the standard output) which you will NOT hand in.

- The estimated μ_0 and $\mathbf{\Sigma}_0$ of the Gaussian for the Alaskan salmon
- The estimated μ_1 and $\mathbf{\Sigma}_1$ of the Gaussian for the Canadian salmon
- The predicted class (Alaskan or Canadian) of the fish in the table

Questions Answers the following questions and put your answers in a pdf file called `proganswers.pdf`

- The estimated μ_0 and $\mathbf{\Sigma}_0$ of the Gaussian for the Alaskan salmon
- The estimated μ_1 and $\mathbf{\Sigma}_1$ of the Gaussian for the Canadian salmon
- The predicted class (Alaskan or Canadian) of the fish in the table

The dataset is from Johnson, R.A. and Wichern, D. W. Applied Multivariate Statistical Analysis (Prentice Hall, International Editions, 2002, fifth edition)