EE 131A | Class Project
Probability and Statistics | Wednesday, February 19, 2020
Instructor: Lara Dolecek | Due: Monday, March 16, 2019 by 11:59 pm PDT via CCLE
TAs: Lev Tauz | levtauz@g.ucla.edu
Debarnab Mitra | debarnabucla@g.ucla.edu

Day in the Life of a Data-Scientist

Reading: Chapters 2 through 8 of *Probability, Statistics, and Random Processes* by A. Leon-Garcia
100 points total

In this project, we will use the tools learned throughout the course to analyze and solve problems faced by data scientists today. Each part will have a combination of MATLAB programming, mathematical analysis and technical writing. You will be graded on all three components.

To complete this project, you will submit a project report which will contain:

(a) All derivations for every part that requests it.

(b) All plots which should be computer-generated from MATLAB.

(c) An appendix where you report all of your MATLAB code used for this project.

In addition to the report, you will also submit all of your MATLAB code in a single zip file for submission.

For derivations, **show all your work**. When producing your plots, **clearly indicate** the x-axis, the y-axis, and what is being plotted (using legends, title etc.). You may need to rescale the x-axis to ensure that your plot is showing the right quantity.

1. (30 pts) *Data Imputation.* Consider the data sequence $\{x_1, x_2, \ldots, x_n\}$ of length $n = 100000$ provided in the file '`data.txt`'. Some of the data points are missing and are denoted by 'NaN' in the file (Eg. x_2, x_{10} are NaN's). You are required to fill out those missing samples using methods learned in the course. Assume that the data points are i.i.d. random variables $\{X_1, X_2, X_3, \ldots X_n\}$ having mean μ and variance σ^2. Let K_{miss} and K_{avail} be the set of indices where the data is missing and available respectively. You replace the missing data point $x_i, i \in K_{miss}$, by a constant number a_i. These a_i's

are chosen in such a way that the expected mean squared error E_{MMSE} of the missing data samples is minimized. E_{MMSE} of the missing data samples is defined as

$$E_{MMSE} = \mathrm{E}\big[\sum_{i \in K_{miss}} (X_i - a_i)^2\big].$$

You minimize E_{MMSE} by setting $\frac{d}{da_i} E_{MMSE} = 0$ for all $i \in K_{miss}$.

(a) Prove that $a_i = \mu$ for all $i \in K_{miss}$ minimizes E_{MMSE}. This shows that you should fill out all the missing samples by the mean μ of the distribution from which the data is derived.

Unfortunately, μ is unknown and needs to be empirically estimated from the available data (non-missing data). Take a single pass over the data and find the sample mean of the first N non-missing data samples among the 50,000 samples. This is our estimate $\hat{\mu}_N$ since it uses N non-missing samples.

(b) Find $\hat{\mu}_N$ for $N = 10, 20, 50, 100, 200, 300, 500, 1000, 2000, 10000, 20000, 30000, 60000$. Plot $\hat{\mu}_N$ vs N. What can you say about this empirical estimator from this plot?

We measure the accuracy $\hat{A}_N$ of our estimate $\hat{\mu}_N$ by computing the mean squared error with respect to the entire available data i,e

$$\hat{A}_N = \frac{\sum_{i \in K_avail} (x_i - \hat{\mu}_N)^2}{|K_{avail}|}$$

where $|K_{avail}|$ is the number of available data points.

(c) Find $\hat{A}_N$ for $N = 10, 20, 50, 100, 200, 300, 500, 1000, 2000, 10000, 20000, 30000, 60000$. Plot $\hat{A}_N$ vs N. Explain your observation.

(d) What is the limiting value of $\hat{A}_N$ as N becomes large? Explain why this value is not zero.

(e) Estimate the value of σ^2 by looking at the limiting value of $\hat{A}_N$.

The data file can be read using MATLAB's `dlmread` (i.e., use `A = dlmread('data.txt')` to read the entire sequence into an array `A`). You can also use any other program to do this problem.

2. (30 pts) *Central Limit Theorem* Let X_1,X_2,... be a sequence of iid random variable with finite mean μ and finite variance σ^2, and let M_n be the mean of the first n random variables in the sequence:

$$M_n = \frac{X_1 + X_2 + ... + X_n}{n}.$$

(a) Let X_i be a uniform continuous random variable taking values in the interval $(2, 6)$. Write a MATLAB program to plot the pdf and cdf of M_n. Consider $n = 1, 2, 3, 4, 5, 10, 20, 30$ and compare your results.

(b) Calculate analytically the mean and the variance of X_i and of M_n in part (a).

(c) Write a MATLAB program to generate a multivariate Gaussian random variable with the same mean and variance as M_n. Superimpose this plot on the plots from part (a).

(d) Repeat parts (a), (b), and (c) with X_i representing a toss of an unfair 5-sided die with each even side ($\{2, 4\}$) twice as likely as each odd side ($\{1, 3, 5\}$).

Use $t = 10^4$ samples in the above. While plotting histogram for discrete data, use `‘BinWidth’` as $\frac{1}{n+1}$.

3. (40 pts) *GDA: Gaussian Discriminant Analysis*

Consider a data sequence of independent random variables $\{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_n, y_n)\}$ of length n where $\vec{x} \in \mathbb{R}^m$ and $y \in \{0, 1\}$. Each $\vec{x}_i$ is a sample of a random vector and y specifies the distribution from which $\vec{x}_i$ is sampled from. If $y_i = 0$, we say the sample is from class 0 otherwise it is from class 1. We assume that each $\vec{x}_i$ is a vector of jointly Gaussian random variables and $\vec{x}_i \sim \mathcal{N}(\mu_0, \Sigma_0)$ if $y_i = 0$ or $\vec{x}_i \sim \mathcal{N}(\mu_1, \Sigma_1)$ if $y_i = 1$. Additionally, $y_i = 1$ with probability p and $y_i = 0$ with probability $1 - p$.

If we assume that a dataset satisfies the previous assumptions, we can determine simple rules to classify samples with unknown labels. One field where such assumptions work fairly well is in the field of biology. As such, you are given a data sequence $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$ where $\vec{x} \in \mathbb{R}^2$ represent a flower petal's length and width. You will use each $\vec{x}$ to classify whether it came from flower 0 or flower 1 (from here on these are termed as class 0 and class 1). Since all the samples are independent, let us consider a single sample $\vec{x}_i$. You will classify $\vec{x}_i$ to be from class 0 if $P(y_i = 0|\vec{x}) \geq P(y_i = 1|\vec{x})$ or to class 1 otherwise.

(a) Assume that $\Sigma = \Sigma_0 = \Sigma_1$, Σ_0 is full rank, and that $p = \frac{1}{2}$. Rewrite the classification rule $P(y = 0|\vec{x}) \geq P(y = 1|\vec{x})$ as a linear inequality, i.e. $\vec{b}^T \cdot \vec{x} + a \geq 0$ where $\vec{b}$ is a vector of length m and a is a scalar. Your answer will be in terms of Σ, μ_0, μ_1.

(b) Consider the data set in `data_2.txt`. These samples were drawn from a distribution where the covariance is the same for both classes. Use the linear inequality from part (a) to classify all the samples in `data_2.txt`.The parameters for this data set are

$$\mu_0 = \begin{bmatrix} 9 \\ 10 \end{bmatrix}, \mu_1 = \begin{bmatrix} 6 \\ 7 \end{bmatrix}, \Sigma = \begin{bmatrix} 1.15 & 0.1 \\ 0.1 & 0.5 \end{bmatrix}, p = \frac{1}{2}.$$

Report the percentage of samples that you classified to be from class 0. Plot the samples on a scatter plot and color all the samples from the same class using the same color. Additionally, plot the contour line of your linear inequality where $\vec{b}^T \cdot \vec{x} + a = 0$. You may use the MATLAB function `fcontour` to plot the contour line.

(c) Now, we consider the case when $\Sigma = \Sigma_0 = \Sigma_1$ and Σ_0 is full rank. As such, each sample is not equally likely to come from each class. Fortunately, the classification rule $P(y = 0|\vec{x}) \geq P(y = 1|\vec{x})$ can still be written as a linear inequality. Find this new linear inequality. Your answer will be in terms of Σ, μ_0, μ_1, p.

(d) Again, we consider the data set in `data_2.txt`. Using $p = 0.05$, repeat all the steps in part (b). What affect did changing p have on the linear boundary?

(e) Finally, we consider the case where we do not assume that all distributions have the same covariance, i.e. $\Sigma_0 \neq \Sigma_1$. The only thing we assume is that Σ_0 and Σ_1 are full rank. Rewrite the classification rule $P(y = 0|\vec{x}) \geq P(y = 1|\vec{x})$ as a quadratic inequality, i.e. $\vec{x}^T C\vec{x} + \vec{b}^T \cdot \vec{x} + a \geq 0$ where C is an m by m matrix, $\vec{b}$ is a vector of length m and a is a scalar. Your answer will be in terms of $\Sigma_0, \Sigma_1, \mu_0, \mu_1, p$.

(f) Consider a new dataset in `data_3.txt`. For this dataset, the parameters are

$$\mu_0 = \begin{bmatrix} 9 \\ 10 \end{bmatrix}, \mu_1 = \begin{bmatrix} 6 \\ 7 \end{bmatrix}, \Sigma_0 = \begin{bmatrix} 1.15 & 0.1 \\ 0.1 & 0.5 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 0.2 & 0.3 \\ 0.3 & 2 \end{bmatrix}, p = \frac{1}{2}.$$

Use the quadratic rule in part (e) to classify all the samples in `data_3.txt`. Similarly to part (b), report the percentage of samples that you classified to be from class 0. Plot the samples on a scatter plot and color all the samples from the same class using the same color. Additionally, plot the contour line of your quadratic inequality where $\vec{x}^T C\vec{x} + \vec{b}^T \cdot \vec{x} + a = 0$.