# Assignment 1

## CS 750/850 Machine Learning

- **Due**: Thursday 9/04 at 11:59PM AOE
- **Submission**: Turn in as a **PDF** and the **source code** (qmd or ipynb) on MyCourses. You can directly submit an IPython ipynb notebook, or convert it to qmd (see https://quarto.org/docs/tools/jupyter-lab.html#converting-notebooks).

This document is created in Quarto which is an extension of the popular Markdown document format. Please see https://quarto.org/ for more details on how to create Quarto documents and on development tools that are available.

I recommend that you complete the theoretical questions using a pen and paper. You can scan the solutions or bring them on paper to the class. There is no need to typeset the solution, which takes extra time.

Please use email only as the last resort. Post your questions in the discussion forum, or please come to the office hours. The instructor would be delighted to see everyone during the office hours at least once during the semester.

### Problem 1 [30%]

To learn to use R (https://www.r-project.org/) or Python, read the labs in Chapter 2 of the textbook. We recommend that you use quarto notebooks with RStudio to typeset your solutions. Other options exist for Python.

1. Download the advertising dataset (`Advertising.csv`) from https://www.statlearning.com/s/Advertising.csv and load it into R/Python.
2. What are the minimum, maximum, and mean value of each feature?
3. Produce a scatterplot matrix of all variables
4. Produce a histogram of TV advertising

You are free to use Python, but then you are on your own when it comes to installing all the necessary tools. You can, however, use the Python version of the ISLR textbook. We cannot guarantee being able to help with the assignments or installation of Python development tools.

You can use Rstudio, VS Code, vim, emacs, jupyter, or any editor that you may be familiar with.

### Hints

1. An easy way to launch help for any function in R, such as `summary`, is to execute: `> ?summary`
2. See https://quarto.org/docs/tools/rstudio.html for how to generate a PDF from a quarto notebook in R-studio.
3. You may also need to install LaTeXwhich you can get from https://www.latex-project.org/get/, which can be tricky on Windows.
4. For more advanced and prettier plotting capabilities, see the package `ggplot`: http://ggplot2.tidyverse.org/ and the cheatsheets at https://www.rstudio.com/resources/cheatsheets/

## Problem 2 [30%]

Consider the following bivariate distribution $p(x, y)$ of two discrete random variables $X$ and $Y$:

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|
| $y_1$ | 0.01 | 0.02 | 0.03 | 0.1 | 0.1 |
| $y_2$ | 0.05 | 0.1 | 0.05 | 0.07 | 0.2 |
| $y_3$ | 0.1 | 0.05 | 0.03 | 0.05 | 0.04 |

1. Verify that this is a valid probability distribution.
2. Compute the marginal distributions $p(x)$ and $p(y)$ for all values $x$ and $y$.
3. The conditional distributions $p(x \mid Y = y_1)$ and $p(y \mid X = x_3)$.

### Hint

You can either solve the problem by hand or write a short script that work on the probability matrix:

```
P <- matrix(c(0.01,0.02,0.03,0.1,0.1,
              0.05,0.1,0.05,0.07,0.2,
              0.1,0.05,0.03,0.05,0.04),
            nrow = 3, ncol = 5, byrow=TRUE)
```

Note that you can use `dimnames` to give appropriate names to the dimensions of the matrix.

## Problem 3 [30%]

*Based on a true story, according to*: The Drunkard's Walk: How Randomness Rules Our Lives, Leonard Mlodinow. A diagnosis of a rare disease does not always mean what you think.

Suppose that you applied for life insurance and underwent a physical exam. The bad news is that your application was rejected because you tested positive for HIV. The test's *sensitivity* is 99.7% and *specificity* is 98.5% [https://en.wikipedia.org/wiki/Diagnosis_of_HIV/AIDS#Accuracy_of_HIV_testing]. However, after studying the CDC website, you find that in your ethnic group (age, gender, race, …) only one in 25,000 people is actually infected. What is the probability that you actually have HIV conditional on having a positive test?

In the statement above:

- Sensitivity refers to the probability of a positive test conditioned on truly having the condition.
- Specificity refers to the probability of a negative test conditioned on not having the condition.

Please complete the following steps:

1. Define the random variables that can be used to represent this problem
2. Express probabilities given in the statement of the problem
3. Give the formula that answers the question of having HIV when a test is positive?
4. Compute the probability

### Hint

You should use the Bayes theorem to solve this problem.

## Problem 4 [10%]

Give a simple example of a random variable which has a CDF but does not have a PDF.