ECE 492-45 Homework 2

Material Covered: Probability Basics, Statistical Learning Introduction

Study Guide:

- Make sure you have a reasonably good grasp of the probability materials we covered in class. If unclear, read relevant sections in an undergraduate probability and/or statistics book such as Devore (3.3, 4.2, 2.4) and Leon-Garcia (3.2, 5.7). Doing a couple of exercises will help.
- The ISLR book has online solutions. For problems that are not assigned, you may choose to quickly go over them with the help from the solutions to improve your understandings in course materials.

Problem 1 (20 points) [Conditional Expectation, Variance Operator]

a) Given the joint PMF for random variables X and Y in the table, compute the following quantities and tabulate your results: $p_X(x)$, $p_Y(y)$, $p_{X|Y}(x|y)$, $p_{Y|X}(y|x)$, $\mathbb{E}[X]$, $\mathbb{E}[Y]$, $\mathbb{E}[X|Y=y]$, $\mathbb{E}[Y|X=x]$. (Intermediate steps must be shown to receive full points.) Explain the difference between $\mathbb{E}[X|Y=y]$ and $\mathbb{E}[X|Y]$.

$Y \setminus X$	1	2	3
0	0.3	0.1	0.3
1	0.1	0.1	0.1

Table 1: Joint PMF, $p_{XY}(x,y)$

b) Prove the following formulas:

$$\operatorname{Var}(aX+b) = a^2 \operatorname{Var}(X), \tag{1a}$$

$$\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X,Y), \tag{1b}$$

$$\operatorname{Var}(X-Y) = \operatorname{Var}(X) + \operatorname{Var}(Y), \text{ when } X \text{ and } Y \text{ are uncorrelated.} \tag{1c}$$

$$\operatorname{Var}\left(\sum_i a_i X_i\right) = \sum_i a_i^2 \operatorname{Var}(X_i),\ X_i\text{'s uncorrelated. Useful for Problem 5d.} \tag{1d}$$

You may find the following equations useful: i) the shortcut formula for variance, $\operatorname{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}X)^2$; and ii) the definition of covariance, $\operatorname{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$. Answer the following questions:

- Why does b not appear on the right-hand side of (1a)?
- How does the variance of sum of two random variable compare to the sum of the variance of individual variables when the variables are negatively or anti-correlated? Can you give an extreme example?
- Why is it a plus sign rather than a minus sign on the right-hand side of (1c)?

Problem 2 (20 points) [Real-Life Statistical Learning Examples] Complete *ISLR-2.4.4*. Information from *ISLR-2.4.2* may be helpful.

Problem 3 (20 points) [`Auto` Data Analysis] Complete *ISLR-2.4.9*.

Problem 4 (20 points) [`Boston` Data Analysis] Complete *ISLR-2.4.10*.

Problem 5 (20 points) [Curse of Dimensionality] Read the first paragraph of the problem statement of *ESLII-2.4*. Note that we may also write $\mathbf{X} = (X_1, X_2, \ldots, X_p)$, where $X_k \sim \mathcal{N}(0, 1)$ for $k = 1, \ldots, p$. Use a programming language of your choice. To get started, set $p = 10$. Note that in this problem, all vectors are column vectors.

a) Write a computer program to randomly draw/generate $N = 100$ vectors from the template random vector $\mathbf{X}$, namely, $\{\mathbf{x}^{(i)},\ i = 1, \ldots, N\}$. Note that each vector should contain p normality distributed random numbers. Plot all vectors as points in a 3-D space consisting of the first, second, the last coordinates.

b) Calculate the coordinate value of each point after being projected on to a fixed direction specified by $\mathbf{a} = \mathbf{x}_0/||\mathbf{x}_0||$, namely, $z^{(i)} = \mathbf{a}^T\mathbf{x}^{(i)}$. Here, $\mathbf{x}_0$ is an arbitrary nonzero vector of length p, "T" is the transpose operation, and $z^{(i)} \in \mathbb{R}$. What are the sample mean and sample variance of the projected coordinates $\{z^{(i)},\ i = 1, \ldots, N\}$?

c) Repeat a) and b) for $p \in [1, 80]$. You may want to use a `for` loop to achieve this. Optionally, put your code for parts a) and b) into a function to make your code easier to read. Plot the sample variance of the projected coordinates as a function of p.

d) Calculate the squared distance of each point to the origin, namely, $d_i^2 = ||\mathbf{x}^{(i)}||^2$. What is the sample mean of $\{d_i^2,\ i = 1, \ldots, N\}$? Plot the sample mean of the squared distance as a function of p in the same plot of c). Limit the range of y-axis between 0 and 80. For $p = 5$, inspect the values of any five d_i^2's. Do the results in b) and c) match with conclusion drawn in the third paragraph of *ESLII-2.4*?

e) Use the formulas from Problem 1b, prove that $\mathrm{Var}(Z) = 1$ where $Z = \mathbf{a}^T\mathbf{X}$, and $\mathbb{E}[D^2] = p$ where $D = ||\mathbf{X}||$. Are the theoretical results in this part consistent with the simulated results obtained in c) and d)? (Hint: The sum of p squared normal random variables is a chi-square random variable χ_p^2. The mean of χ_p^2 is p.)