Homework 3: Applied Probability and Statistics

University of Arizona CSC 380: Principles of Data Science

Homework due at 11:59pm on September 23

This assignment will strengthen your understanding of the fundamental concepts in applied probability and statistics. The questions in this assignment will build on lecture material as well as the assigned readings from (Wasserman, L. 2004. "All of Statistics").

Deliverables Submit your responses as a PDF along with code to D2L by the stated deadline. **Show all work along with answers.** This is for your benefit as incorrect answers may receive partial credit if the work demonstrates understanding.

Problem 1: Conditional Independence (2 points)

Let $A, B, C \in \{0, 1\}$ be three binary random variables with the following joint probability distribution:

A	B	C	$P(A, B, C)$
0	0	0	0.192
0	0	1	0.144
0	1	0	0.048
0	1	1	0.216
1	0	0	0.192
1	0	1	0.064
1	1	0	0.048
1	1	1	0.096

a) By direct calculation, compute the marginal $P(A, B)$.

b) By direct calculation compute the marginals $P(A)$ *and* $P(B)$.

c) Are the random variables A *and* B *dependent? Why or why not?*

d) Compute the conditional $P(A, B \mid C)$

e) Show that A *and* B *are conditionally independent, given* C.

Problem 2: Diagnostic Tests and Bayes' Rule (1 point)

I have decided to get myself tested for COVID-19 antibodies. However, being comfortable with statistics, I am curious about what the test means for my actual status. Let's investigate these questions, showing all your work:

a) According to the FDA, the UA COVID-19 antibody test (known as ELISA pan-Ig) has a sensitivity *(a.k.a. true positive rate) of 97.5% and a* specificity *(a.k.a. true negative rate) of 99.1%. Assume that 5% of the population actually have COVID-19 antibodies. Write down the joint probability (prior/likelihood) with events for disease state $S \in \{true, false\}$ and test result $R \in \{true, false\}$.*

b) Assuming I receive a positive *test result, use Bayes' rule to calculate the probability that I actually have COVID-19 antibodies?*

c) Assuming I receive a negative *test result, what is the probability that I* do not *have COVID-19 antibodies?*

d) Assume I take the test twice, and receive a positive *result in both tests. What is the probability that I have COVID-19 antibodies according to Bayes' rule?*

e) Now assume that only 1% of the population has COVID-19 antibodies. Repeat parts (b) and (c) with this revised prior belief.

Problem 3: Bootstrap Confidence Intervals (3 points)

This question will walk you through the process of estimating the correlation of two random variables, along with confidence intervals using the bootstrap method. Lecture slides and Chapter 8 of the textbook (Wasserman) will be helpful if you need a refresher. For our chosen model, we will use a bivariate Gaussian distribution: $P(X, Y; \mu, \Sigma) = \mathcal{N}(\mu, \Sigma)$. For simplicity, let us assume that the distribution is zero-mean $\mu = (0, 0)^T$. Let us assume that the true (unknown) covariance matrix is,

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

We have written the covariance matrix in terms of the correlation coefficient,

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sigma_X\sigma_Y} = \frac{0.5}{1 \cdot 1} = 0.5.$$

Using **numpy.random.seed** set your random number generator seed to 0 and answer the following:

*a) Create a dataset by drawing $N = 100$ samples from our model using **numpy.random** function **multivariate_normal**. Create a scatterplot of your data using **matplotlib.pyplot** functions **scatter** or **plot**. Label your axes X and Y.*

b) Compute and report the plug-in estimator of correlation, given by:

$$\hat{\rho}(X_1, \ldots, X_N) = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2 \sum_j (Y_j - \bar{Y})^2}}$$

Where $\bar{X} = 1/N \sum_i X_i$ is the sample mean (and similarly for $\bar{Y}$).

*c) Using **np.random.choice** sample $M = 50$ points $X_1^*, \ldots, X_M^*$ from your generated data **with replacement**. Generate a new estimate from this data from your data $\hat{\rho}(X_1^*, \ldots, X_M^*)$ and report.*

*d) Repeat the above process $B = 100$ times to generate bootstrap estimates $\hat{\rho}_{M,1}, \ldots, \hat{\rho}_{M,B}$. Each estimator should be based on a different sample of $M = 100$ points drawn, with replacement, from the original N data. Display a histogram of your bootstrap estimates using **matplotlib.pyplot.hist** with 30 bins. Label your axes.*

e) Compute and report your bootstrap estimate of the correlation as the sample mean $\bar{\rho} = \frac{1}{B} \sum_i \hat{\rho}_{M,i}$.

f) Compute and report the 95% confidence interval as $\bar{\rho} \pm 2\bar{\sigma}_\rho$ where $\bar{\sigma}_\rho$ is the sample standard deviation. Does the interval contain the true correlation?