COSC 311 - Lab 4

Dr. Joe Anderson

Due: 29 October

1 Objectives

1. Practice efficiently manipulating data with Python
2. Use the matplotlib, pandas libraries
3. Gain familiarity with statistical tools

2 Tasks

1. You may submit this lab in groups of one or two.

2. Download the “Adult” data set from the UCI Machine Learning data repository: [https://archive.ics.uci.edu/ml/datasets/Adult](https://archive.ics.uci.edu/ml/datasets/Adult). This dataset is record of adults, along with various occupational and lifestyle attributes. Each adult is “labeled” as to whether or not they make more or less than $50k per year. Using this as a driving label, one would typically want to design a process to determine what combinations of factors enable a person to make more than $50k per year.

   (a) Read the data into a pandas DataFrame object.
   
   (b) Use the data and the numpy library to compute the following:
      
      i. What are the 25th, 50th, and 75th percentiles of the “education-num” field?
      
      ii. What is the probability that an adult makes more than $50k given that their education-num is within the ranges defined by the above quantiles (from 0 to the 25th percentile, from the 25th to the 50th etc)?
      
      iii. Plot the change in probability that a person makes more and less than $50k given their years of education.
      
      iv. What is the covariance between the number of hours worked per week and education-num?
      
      v. Use the pandas.DataFrame.boxplot functionality to create a box-and-whisker plot which illustrates the spread of hours worked among adults who make both more and less than $50k.
      
      vi. Use the pandas.DataFrame.boxplot functionality to create a box-and-whisker plot which illustrates the spread of hours worked among adults from each native country and who make more and less than $50k.
      
      vii. Create a table where entry \((x, y)\) contains the conditional probability

      \[
P(A \text{ random adult has level of education } x | \text{their education-num is } y).
      \]

      viii. Create a table where entry \((x, y)\) contains the conditional probability of having marital status \(x\) given that they have occupation \(y\).
      
      ix. What is the conditional probability of making more or less than $50k given that a person works in each different occupation?
3. Answer the following questions using the fundamentals of probability.

(a) If $A$ and $B$ are independent, show that $\bar{A}$ and $B$, $\bar{A}$ and $B$, $\bar{A}$ and $\bar{B}$ are independent.

(b) Suppose we send 30% of our products to company $A$ and 70% of our products to company $B$. Company $A$ reports that 5% of our products are defective and company $B$ reports that 4% of our products are defective. For each probability below, compute the precise value by hand, and also write a short Python script to simulate the above scenario and estimate each probability by empirically examining the rates of each event.

(i) Find the probability that a product is sent to company $A$ and it is defective.

(ii) Find the probability that a product is sent to company $A$ and it is not defective.

(iii) Find the probability that a product is sent to company $B$ and it is defective.

(iv) Find the probability that a product is sent to company $B$ and it is not defective.

(c) Show that for events $A$ and $B$ that $P(A|B) > P(A)$ implies $P(B|A) > P(B)$.

3 Submission

Zip your source files and upload them to the assignment page on MyClasses. Be sure to include all source files, properly documented, a README file to describe the program and how it works, along with answers to any above discussion questions.