PH227: AI and Data Science

Tutorial Sheet 6: Python/Data-Science Part

Note: This tutorial sheet contains programming problems related to: (a) chi-square tests, and (b) k-means clustering algorithm. You have to submit the codes corresponding to the starred (*) problems to the TAs. The marks of the starred problems are indicated next to them.

1. * Consider the Titanic passenger dataset file Titanic-Dataset.csv with total 892 rows (including the header), and 12 columns about which you can learn from the internet. This dataset has some missing values, however, the columns on whose data we will perform the chi-sqaure test, contain no missing values; therefore, you can use the file without any need to clean it. The columns of interest are: Survived (0 for no, and 1 for yes), Sex (gender of the passenger, male or female), Pclass (the class in which the passenger was traveling, 1, 2, or 3). Write a Python program which, based on user's choice, will perform either of the two chi-square tests of independence:

   (a) To determine whether the survival is independent of the passenger's gender. Here H_0: Survival is independent of passenger's gender, and H_a: Survival is not independent of passenger's gender. The contingency table for this test will be 2×2 (Survived $\times$ Sex)

   (b) To determine whether the survival is independent of the passenger's class. Here H_0: Survival is independent of passenger's class, and H_a: Survival is not independent of passenger's class. The contingency table for this test will be 2×3 (Survived $\times$ Pclass)

   Compute the Chi-Square statistic, degrees of freedom, and p-value. Using the significance level $\alpha = 0.05$, print your decision regarding whether or not we reject H_0. (10 marks)

2. * Consider the dataset contained in the file Mall_Customers.csv which has 201 rows, and 5 columns: (a) customer ID, integer 1-200, (b) customer gender (Male or Female), (c) age, (d) annual income in the units of thousand (1000) US dollars, and (e) spending score, integer 1-100. Write a Python program which will:

   (a) Make a scatter plot of the dataset based on the last two features, i.e., the annual income, and spending score.

   (b) Assuming that the total number of clusters for this data is $K = 5$, implement a K-means clustering algorithm and plot the scatter plot of the data after the algorithm converges. Make sure to use different colors for the data points belonging to different clusters, and also identify the centroid for each cluster using the "cross" (X) symbol. For this part also, use only the last two features, as in part (a). (10 marks)