Assignment 5 (Knowledge Discovery and Data Mining)

Number of problems/points: Five problems for total of 100 points
Out: February 20, 2020
Due: February 27, 2020 in class (hard copy).

Problem 1: (20 points)
Download Census Income data set from UCI repository - link is
https://archive.ics.uci.edu/ml/datasets/census+income
Use training data do develop a model aimed to determine whether a person makes over 50K a year.
Solve this problem using a Decision Tree (ID3. CART or C4.5 algorithm) and report ROC on test data.

To train a Decision Tree Classifier, you can use the following SciPy built-in function (but you are not required to use this specific toolbox):
Details:

- call the "fit(X, y)" method to train the decision tree
- call the "predict(X)" method to get the predicted class labels (represented by discrete numbers in \{0,...,#classes\})
- call the "predict_proba(X)" method to get the predicted class probabilities (or probability scores)

After building the decision tree, you can calculate the ROC AUC score using the function described in the following link:
Details:

- call the "roc_auc_score(y_true, y_score)" method, where:
  - y_true contains the true class labels
  - y_score contains the predicted class probabilities

Problem 2: (20 points) Now, use a logistic regression model to solve Census Income problem and report ROC.

Problem 3: (20 points) Solve Census Income problem using feed-forward neural network and report ROC.

Problem 4: (20 points)
(a) Suppose the fraction of undergraduate students who smoke is 15% and the fraction of graduate students who smoke is 23%. If one-fifth of the college students are graduate students and the rest are undergraduates, what is the probability that a student who smokes is a graduate student?

(b) Given the information in part (a), is a randomly chosen college student more likely to be a graduate or undergraduate student?

(c) Repeat part (b) assuming that the student is a smoker.
(d) Suppose 30% of the graduate students live in a dorm but only 10% of the undergraduate students live in a dorm. If a student smokes and lives in the dorm, is he or she more likely to be a graduate or undergraduate student? You can assume independence between students who live in a dorm and those who smoke.

**Problem 5: (20 points)**
Consider a data set containing four points located at the corners of the square. The two points on one diagonal belong to one class, and the two points on the other diagonal belong to the other class. Is this data set linearly separable? Provide a proof.