Assignment 2

1. We can build unsupervised data mining models when we lack labels for the target variable in the training data, multiple choice, choose one:

   - ✓ True
   - False

2. For supervised data mining the value of the target variable is known when the model is used, multiple choice, choose one:

   - True
   - ✓ False

3. A company is attempting to predict customer churn, they have the following columns and their corresponding possible values. Each of the following is converted, using One-Hot Encoding.

   Employed: Yes/No
   Education_Level: None, HighSchool, CollegeDegree, GraduateDegree, Other
   Health_Insurance: None, EmployerProvided, SelfObtained

   How many new columns will be created (multiple choice choose one):

   - 2
   - 3
   - 5
   - ✓ 10

4. You have 9500 customer records, you split these into a training set containing 5000 records and a test set containing 4500 records. You train a classifier to predict churn (Yes/No) and 4000 of your predictions in-sample are correct. You discover that your classifier simply predicted “No” for all records. What is your in-sample base rate (provide a percentage between 0 and 100)?

   4000 / 4500 ≈ 88.89 %

5. In the customer churn example above, you try to use the following features to predict churn, mark each feature as most likely to be either “categorical” or “continuous”:

Age_in_years: continuous
Days_since_last_purchase: continuous
Number_of_orders_last_12_months: continuous
Married_yes_no: categorical
Has_logged_in_last_1_month: categorical
Annual_income_dollars: continuous
Is_a_US_citizen_yes_no: categorical
Has_children: categorical
Homeowner_or_renter: categorical
Price_of_last_item_purchased: continuous
Number_of_logins_last_12_months: continuous
Percentage_of_orders_with_coupon_code_used: continuous
Customer_acquisition_channel: categorical
Number_of_returns_last_12_months: continuous
Called_customer_service_last_1_month_yes_no: categorical

6. You try to build a model to predict the probability of customer churn. Your historical data consists of customers who left (churn=1) or customers who didn’t leave (churn=0). Your model predicts a probability value p, between 0 and 1 representing the probability of churn. You choose to measure the performance of your model using a metric called log loss, defined as follows (y is 1/0 i.e. whether or not your customer churned):

$$-(y\log(p)+(1-y)\log(1-p))$$

For the following customers, you are provided with the actual label (churn=1, no churn=0), and the probability of churn predicted by your model. Write a function in Python (called **logloss)** that takes two arguments, y and p and computes the log loss (numpy has a “log” method), compute and report the log loss for each using this function (log here refers to natural log):

(see jupyter)

churn	predicted_probability	log loss
1	0.77	0.261364764...
0	0.24	0.2744368
0	0.88	2.120263536
1	0.63	0.462035459 59
0	0.41	0.52763274

7. I have built a linear regression model to predict a home's price based on a variety of attributes. Write a python function called **mse** that takes in two lists, actual_values and predicted_values and returns the mean_squared error (in your method, calculate the mean squared error by taking the difference between the items of the two lists, squaring them and taking the average, use numpy functions where appropriate). Call that function with the following two lists and report the mean squared error below.

   Note that the actual and predicted values below are in 1000's (i.e. 228 represents $228,000) but you should compute your MSE on the raw values:

   actual = [228, 115, 154, 150, 174]
   predicted = [109, 56, 122, 86, 210]

   4811.6 * 1000 = 4811600

8. Now load the attached CSV file (linear_regression_predictions.csv) into a Jupyter notebook Python session, using pandas.read_csv(). This CSV has 3 columns: actual (the actual values), pred_model1 (predictions made by linear regression model #1) and pred_model2 (predictions made by linear regression model #2). Using the method **mse** you wrote for the prior question, calculate the mean squared error of model1 and model 2 and report them below

   MSE of model 1: 38240.2344

   MSE of model 2: 202376.1846

   Based on these MSE values alone, which model has better performance? model 2

   Now observe the data in the CSV file manually, which model would you choose between the two models and discuss why (with the context of the two MSE values computed above):

   I choose the model 2. Actually, predicted values of model2 is closer except for the value of 2510.001, when that of model1 is 3358.702 and the actual value is 3688. Since the difference between 2510.001 and 3688 is 1188, the square of 1188 is quite large, which results in the bigger mse of model2. Leaving 2510.001 alone, performance of model2 is better.

9. You decide to just use the attributes of your customers (the ones listed in the question above where you marked attributes as categorical/continuous) to create 4 segments of customers. Based on what these groups may look like, you may create marketing campaigns targeted to each segment regardless of whether you think they will churn or not. This is an example of (multiple choice, choose one):

   - ○ Supervised Learning
   - ✓ ○ Unsupervised Learning

   prior

10. You take historical data that you intend to build a supervised model from. One of your colleagues helps you to create a number of features from this data and to split it into training and test sets. One of these features has leakage (i.e. it uses future data). Your out-of-sample performance is likely to be (multiple choice, choose one):

    - ✓ ○ Overestimated
    - ○ Underestimated

11. Using cross-validation across many folds of your data to measure model performance is better than creating a single training/test split, because (select as many as you think are correct):

    - ✓ ○ You can get a more reliable measure of performance
    - ✓ ○ You can measure stability of your model's performance
    - ○ You don't have to worry about missing data
    - ○ Outliers won't affect any of your performance numbers

12. Mark each of the following as either classification or regression

    - ○ A system that marks emails as spam or not: classification
    - ○ A system that labels a customer as a frequent purchaser or not: classification
    - ○ A system that predicts a farmer's crop yield for the upcoming growing season: regression
    - ○ A system that labels a credit card swipe as fraud or not: classification
    - ○ A system that predicts a potential customer's lifetime value: regression
    - ○ A system that takes an image of a handwritten digit and predicts whether it is a 0, 1, 2, 3, 4, 5, 6, 7, 8 or 9: classification

13. Please paste below the code for your function "**mse**" written for questions 7 and 8 above.

see Assignment 3.ipynb #Question 7, #Question 8

14. Please paste below the code for your function "**logloss**" written for question 6 above.

see Assignment 3.ipynb #Question 6

15. You have been provided with the file **nearest_neighbors_example.csv**. The file contains 3 attributes (Age, in years; Annual Income, in 1000's; Number of Credit Cards) on 7 customers.

    Load the file into a Dataframe using Pandas.

    Create a column that includes the euclidean distance of each customer from "David", based on the 3 attributes provided.

    Note that the entry for "David" (i.e. distance from David to David) should be 0. See Table 6-1 (Page 147) in the Provost/Fawcett textbook for an example.

    Display this Dataframe in your notebook after having created the column.

see Assignment 3.ipynb # Question 15.