**Problems Inclass 8\_2.** You can comment in this document and submit a pdf of your work. Please mark clearly all your answers and answer problems in the order provided.

1. Think through and answer the following problems to the best of your abilities.
2. Valentine Day is approaching. A restaurant is trying to decide if to organize a singles’ night or if to offer a special romantic menu. The restaurant has an established base of customers and collects demographic, income, social media and behavioral information on its customers. They decide to use the help of a data scientist to make sense of their Valentine’s day menu in order to maximize sales (Valentine’s days tend to be cash cows for restaurants). What algorithm would you use?

I would say it is best to use logistic regression in this case. For a given scenario (singles night or couples dinner), a probability of participation for a given person can be predicted. Based on this and average meal cost, the restaurant can have a good estimate of which would provide them with more money.

1. Describe the type of information you would collect (what features) to decide if an email is spam or non-spam and what machine learning algorithm you would use.

Some relevant information for whether an email is spam or not includes: number of capital letters, word repetition, who the email was sent to, and whether the email contains links or attachments. Based on this, I would once again choose a logistical regression model, which can give a probability of an email being spam and a decision can then be made based on that.

1. Describe the type of information you would collect (what features) and from what sources to decide if to buy or sell a stock (financial investment). What machine learning algorithm can you use?

The information collected would be the current value, recent change in the pricing of the stock in question, recent changing of industry stock as well as the stock market as a whole, data on who is buying/selling, and any relevant social media buzz. I would choose linear regression for this, and the choice on whether to buy/sell would be determined by whether the output meets a certain threshold.

1. How would you use Facebook to recommend certain products to people and what machine learning algorithm would you use?

For this problem I would use k-means clustering. Based on user’s “likes” and groups and data from their posts, they can be grouped together based on shared interests, and product purchasing habits and recommendations can then be grouped together as well.

1. A classification algorithm classifies emails into spam and non-spams. The following confusion matrix was returned by using the classifier on the testing set:

|  |  |
| --- | --- |
| 264 | 14 |
| 22 | 158 |

Consider “non-spam” = “positive” class. The matrix has the organization described in class. Calculate and interpret the following:

1. Accuracy rate = TP+TN/TP+TN+FP+FN = 92%
2. Precision = TP/TP+FP = ~92%
3. Recall = TP/TP+FN = 88%
4. F1 = 2\*Precision\*(recall/precision+recall) = 2\*0.91\*0.49 = 0.89
5. Sensitivity = TP/(TP+FN) = 88%
6. Specificity = TN/(TN+FP) = 95%
7. In your opinion, is it more important to have good recall or precision?

In this case, precision. We want to ensure that we have the fewest number of false positives as possible, because users can manually deal with the few false negatives that will get through as a result.