Ronald Lee

CDMA 3654

Inclass 8_2

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.

    a) 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 use Linear Regression since it is good for predicting the sale size based on the customer's demographic and income characteristics.

    b) 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

    The type of information I would collect is the categorical variables. I would assign an observation to one of two classes: spam or non-spam. So, I would use Naïve Bayes algorithm.

    c) 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 type of information I would collect is the return on investment based on stock fundamentals and market return. So, I can use Linear Regression algorithm.

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

    I would use Nearest neighbor recommender systems that predict next item to add to cart based on similar customers' choices.

2. 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
   (264 + 158) / (264 + 14 + 22 + 158) = 0.92
   With 92%, non-spam mails are classified correctly.
2) Precision
   264 / (264 + 22) = 0.92
   If we say some mails are non-spam mails, at least 92% of them better be that.
3) Recall
   264 / (264 + 14) = 0.95
   We can identify 95% of the true non-spam mails.
4) F1
   2 * (0.92) * (0.95/(0.92 + 0.95)) = 0.93
   This score is the combination of precision and recall so we can predict and identify non-spam emails with 93%.
5) Sensitivity
   264 / (264 + 14) = 0.95
   We can identify 95% of the true non-spam mails like Recall method.
6) Specificity
   158 / (158 + 22) = 0.88
   We can see 88% of spam mails correctly.
7) In your opinion, is it more important to have good recall or precision?
   I think good precision is more important than good recall. Prediction is just people's guess, but the precision is the statistical dataset which can be truly believed as objective information.