COMP 5450 Machine Learning

Homework 1

1. Give 2 examples of a good application of machine learning, and 2 examples where machine learning should not be used. (**5pts**)
2. Explain the difference between overfitting and underfitting, give examples. (**5pts**)
3. The mean monthly rent for a random sample of 100 studio apartments is $1200. Given the population standard deviation of $200, construct a 95% confidence interval. (**5pts**)
4. When a vaccine is given to patients, 96% of the patients got the vaccine test positive for antibodies and 5% of the patients who did not get the vaccine test positive for antibodies. Suppose that 90% of patients get the vaccine. What is the probability that a patient who tests positive for antibodies got the vaccine? (**5pts**)
5. Consider the hypothesis space defined over instances shown below, each hypothesis is represented by 4-tuples. Use the naive Bayes classifier to predict the target value PlayTennis=Yes/No to the following instances. (**15pts**)
   a) <Sunny, Cool, Normal, Weak>
   b) <Rain, Mild, High, Strong>

Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	Normal	Weak	Yes
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	No
D4	Rain	Mild	Normal	Strong	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Mild	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Rain	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Strong	Yes
D10	Rain	Hot	High	Weak	No
D11	Sunny	Mild	High	Weak	No
D12	Sunny	Mild	Normal	Strong	Yes
D13	Sunny	Cool	Normal	Strong	No
D14	Overcast	Cool	High	Weak	Yes

6. See the provided ‘Weather’ dataset, use simple linear regression to model the prediction of ‘Temperature..C.’ based on ‘Humidity’. Then perform the simple linear regression by gradient descent (hint: play around with the parameter settings). Finally check the performance of multiple regression with additional parameters ‘Wind.Speed..km.h.’, and ‘Pressure..millibars.’. Which model would you choose, simple or multiple, why? (**20pts**)
7. Given the transactions in the table below, use the Apriori algorithm to find the frequent itemsets. Minimum support is 60%, minimum confidence is 80%. Identify 2 strong association rules and list the support and confidence for each. (**20pts**)

TID	Items
T01	{Milk, Banana, Eggs}
T02	{Oats, Banana, Eggs, Grapes}
T03	{Milk, Eggs}
T04	{Milk, Banana, Cookies, Eggs}
T05	{Oats, Banana, Eggs, Cookies}

8. Please cluster the following points, (18, 10), (21, 11), (22, 22), (24, 15), (26, 12), (26, 13), (27, 14), (30, 33), (31, 39), (35, 37), (39, 44), (40, 27), (41, 29), (42, 20), (44, 28), (46, 21), (47, 30), (48, 31), (49, 23), (54, 24) use the numbers as pairs of x and y values which represent their locations. All distances are measured with Euclidean distance. (**25pts**)
   a. Use the k-means algorithm, with (24, 15), (30, 33), and (54, 24) as the initial cluster centers, process the dataset until convergence, choose a method of your choice to evaluate the cluster result quality.
   b. Use the agglomerative hierarchical clustering to cluster the data, use single link, complete link, and average link. Discuss the similarities and/or the dissimilarities.

**** Steps of calculation should be shown. Visualize your results for Q6 and Q8 as part of the written homework that you turn in. You are also required to submit your code (R or Python), include any libraries you use. Type your work when possible, unintelligible work will not be graded.***