CSE 6363 - *Machine Learning*

Homework 1- Spring 2023

Due Date: Feb 17 2023, 11:59pm Central Time

Data Set Generation

This assignment consists of theoretical and implementation questions. Some of them (including 1 b), 1 c), and 2 require data. To obtain the data for these problems you need to do the following:

- Go to https://ranger.uta.edu/∼huber/cse6363/Hwk1/Hwk1gen.php
- Enter your student ID number (the 1000... number on your student ID) and hit submit
- Save the generated web page and submit it with your assignment
- Copy the generated data to files on your computer and use them with the corresponding questions

Make sure that you enter your own student ID. Results on data for other student ID numbers will not be considered correct solutions.

MLE and MAP

1. In class we covered the derivation of basic learning algorithms to derive a model for a coin flip task. Consider a similar problems where we monitor the number of boot failures of an unstable computer that occur before it finally starts up. Assuming that the boot failures are purely random with a fixed probability for the system to boot successfully, q, the number of boot failures before starting follows a geometric distribution. Assume we want to learn the best model for the parameter, q, of the distribution given a data set of observed failed reboot counts counts. The geometric distribution for a successful boot after $k-1$ failed attempts is given by:

$$P_q(k) = (1-q)^{k-1}q$$

   We are assuming here that the different data points are independent, i.e. that the time intervals that were used to determine them were independently distributed over a longer time interval.

   a) Derive the performance function and the optimization result for analytic MLE optimization for a model learning algorithm that returns the MLE for the parameter p of the model given a data set $D = \{k_1, ...k_n\}$. Make sure you show your steps.

   b) Apply the learning algorithm from a) to the dataset for Question 1 that you generated previously.

c) Derive the optimization for a MAP approach using the conjugate prior, the Beta distribution. The Beta distribution is:

$$p_{\alpha,\beta}(q) = \frac{q^{\alpha-1}(1-q)^{\beta-1}}{B(\alpha,\beta)}$$

Note that α and β are constants and that there still is only one parameter, q, to be learned. Show your derivation and the result for the data in part b) and values for α and β that you obtained from data generation web page.

K Nearest Neighbor

2. Consider the problem where we want to predict the type of material (among 3 material types) of a mug based on four measurements, namely the height, diameter, weight, and hue (color). [1]

a) Using Cartesian distance as the similarity measurements show the results of the material type prediction for the Evaluation data that you generated above for Question 2 a) based on the corresponding generated training data for values of K of 1, 3, and 5. Include the intermedia steps (i.e. distance calculation, neighbor selection, prediction).

b) Implement the KNN algorithm for this problem. Your implementation should work with different training data sets and allow to input a data point for the prediction.

c) To evaluate the performance of the KNN algorithm, implement a leave-one-out evaluation routine for your algorithm. In leave-one-out validation, we repeatedly evaluate the algorithm by removing one data point from the training set, training the algorithm on the remaining data set and then testing it on the point we removed to see if the label matches or not. Repeating this for each of the data points gives us an estimate as to the percentage of erroneous predictions the algorithm makes and thus a measure of the accuracy of the algorithm for the given data.

Apply your leave-one-out validation with your KNN algorithm to the dataset for Question 2 c) for values for K of 1, 3, and 5 and report the results. For which value of K do you get the best performance ?

d) Modify your KNN algorithm to use Manhattan distance (L_1) as the similarity measure and repeat the experiment from part c). Which similarity measure gives you a better performance for each of the values of K ? Show your intermediate results.

e) Repeat the prediction experiment from part c) using KNN and Cartesian distance when the fourth attribute in the data is removed (i.e. when only the first three features in the input data are available). Which data gives you better predictions? Show your intermediate results.

[1] For your submission you should use the training data as provided and not normalize it. You are, however, encouraged to try the experiments with normalized data outside the submission to get a feel whether it makes a difference for this dataset.