Assignment1

Yunhua Zhao

February 26, 2021

0.1 Write a "Data Summary" section. [10 pts]

- Describe the dataset and the variables. What is the target? What are you calculating it from?
  Answer: The dataset is 4898*12, if drop the duplicate rows, it is 3961*12. "quality" is the target from these features ['fixed_acidity', 'volatile_acidity', 'citric_acid', 'residual_sugar', 'chlorides', 'free_sulfur_dioxide', 'total_sulfur_dioxide', 'density', 'pH', 'sulphates', 'alcohol', 'quality'].

- Include a list of each variables' descriptive statistics (mean, standard deviation, quartiles).
  Answer:

fixed_acidity	[6.839346, 0.866860, 6.300000]
volatile_acidity	[0.280538, 0.103437, 0.210000]
citric_acid	[0.334332, 0.122446 , 0.270000]
residual_sugar	[5.914819, 4.861646, 1.600000]
chlorides	[0.045905, 0.023103, 0.035000]
free_sulfur_dioxide	[34.889169, 17.210021, 23.000000]
total_sulfur_dioxide	[137.193512, 43.129065, 106.000000]
density	[0.993790, 0.002905 , 0.991620]
pH	[3.195458, 0.151546 , 3.090000]
sulphates	[0.490351, 0.113523, 0.410000]
alcohol	[10.589358, 1.217076 , 9.500000]

- Describe whether or not you used feature scaling and why or why not.
  Answer: I used feature scaling, because the accuracy and F1 all improved, especially F1

- Describe whether or not you dropped any feature and why or why not.
  Answer: I droped two features, residual_sugar and alcohol, because these Pearson correlation coefficient with desity are very big, 0.82 and 0.76

0.2 Write a "Methods" section. [10 pts]

- Describe the runtime complexity of the KNN_Classifier model.
  Answer: $O(n^2)$: there is a for loop in the for loop

- Explain the effects of increasing k. When is and isn't it (increasing k) effective?
  Answer: When k increases, the accuracy and F1 all increase, but 7 neighbors and 9 neighbors are similar result maybe a little grows but not significant, but from 3 to 7, accuracy and F1 are increasing.

- Describe whether or not you used inverse distance weighting in the features and why.

Answer: I will not use the distancing weighting, from my result(in the table) it is not as good as uniform, and uniform has less calculation.

0.3 Write a "Results" section. [10 pts]

- Describe the performance of the model with respect to the different levels of k and the different distance metrics. Include a table of performances, bolding the best.
  Answer: Uniform works better and $k = 7$ is best; the worst is the distance and 3 folds; the result is becoming better when k grows and uniform works better than distance.

neighbor	distance	Euclid/Manhattan	accuracy	F1
3	uniform	Euclid	0.6363636363636364	0.7377049180327868
5	uniform	Euclid	0.6464646464646465	0.7445255474452555
7	uniform	Euclid	**0.6603535353535354**	**0.757875787578758**
9	uniform	Euclid	0.6603535353535354	0.758744394618834
3	distance	Euclid	0.5845959595959596	0.6857688634192932
5	distance	Euclid	0.6287878787878788	0.7277777777777777
7	distance	Euclid	0.6325757575757576	0.7308048103607769
9	distance	Euclid	0.6426767676767676	0.7396504139834407
3	uniform	Manhattan	0.6388888888888888	0.7376146788990826
5	uniform	Manhattan	0.6502525252525253	0.7456382001836547
7	uniform	Manhattan	**0.6691919191919192**	**0.7648114901256732**
9	uniform	Manhattan	0.6616161616161617	0.7598566308243728
3	distance	Manhattan	0.5997474747474747	0.6983824928639392
5	distance	Manhattan	0.625	0.723720930232558
7	distance	Manhattan	0.6338383838383839	0.7309833024118739
9	distance	Manhattan	0.6439393939393939	0.7398523985239852

- Characterize the overall performance of your model.
  Answer: I think my model works well, also I tried the model which uses "kwargs" for high parameters, I use 0.5 as the threshold to convert the probability to binary,($p(labelis1) > 0.5$ then it is predicted as 1)

- Discuss which quality values led to good performance of your model and those that resulted in poor performance. Include a table of average error (e.g., F1 score) to support your claims.
  Answer: Please Check the previous table, I summary everything there, Euclid and Manhattan work similiar results, Manhattan is a little bit better, uniform works better, and the best is 7 folds.

- Give any final conclusions.
  Answer: The hw is so helpful, in future I can add the feedback to my model to choose a better threshold.