Worksheet 4 – Similarity-based Learning

Theory

Review the videos from week 4. Answer the following questions based on those lectures

1. What is meant by an N-dimensional feature space?
2. What is a distance metric and what does it measure? Give an example.
3. How does the Nearest Neighbor algorithm work? What is there no training step?
4. How does the K Nearest Neighbor (kNN) variant of this algorithm work and how is it an improvement?
5. What is the difference in the way we make predictions between categorical and continuous target features?
6. Why is it recommended to normalise continuous feature values before using kNN?
7. The kNN algorithm performs poorly over large datasets. Explain one way we can improve its O(N) performance
8. Explain why the choice of distance metric can be important when using similarity-based learning

Practice

Follow the tutorial videos from week 4 and carry out the following steps

1. Download the code archive and extract the file from the week 4 learning materials. Make sure that you can run the examples code as provided.
2. Load the breast cancer data set from sklearn and run the model again. You will notice that it takes a relatively long time to classify and measure the accuracy. Extend this implementation by adding K-D-Tree optimization to your model search. Retest with the breast cancer data set and compare your new performance with the old implementation.

Answer: I wrote a KD-Tree implementation to check the result, it is OK and has no salient difference.



1. The model as provided in this week’s source code is a classifier. Modify the code to build a regressor, this is a model capable of predicting continuous target values. Test your implementation on the diabetes dataset from sklearn.

Answer: The accuracy\_score to judge the performance is mean square error loss. And based on the regressor, we compute the mse\_loss is approximately 3450.

1. Compare your regressor implementation with the sklearn built-in implementation.

Answer: If we use the built-in linearRegreesion regressor, we can get the result with the mean\_square\_error as 3608.