COMPSCI 762 2022 S1 Week 10 Questions

Luke Chang

May 6, 2022

Question 1 - K-Nearest Neighbour (KNN)

1. When building a *K-Nearest Neighbour* (KNN) model, there are many options for selecting distance metrics, e.g., Euclidean distance (L_2-norm), Manhattan distance (L_1-norm), and Mahalanobis distance (L_∞-norm). In term of the type (e.g., continuous, discrete, and categorical) and the dimension of the data (the number of input features), give examples for how do you choose a distance metric, and explain why you think that is the suitable distance metric for the data set.
2. The parameter k controls the number of neighbours each indexed point are returned. How do you chose the value of k? Do you think using $k = 1$ is a good idea? Is there an application where $k = 1$ is a suitable choice? Explain why that is the case.

Question 2 - K-Nearest Neighbour (KNN)

You are given the following data set:

	F_1	F_2	F_3	$Target$
Train	0	100	High	A
	0	50	High	A
	1	5	High	A
	1	0	Low	B
	1	10	Low	B
Test	0	10	High	B
	1	5	Low	B
	1	50	High	A

1. Train a 1-Nearest Neighbour classifier on the training set and apply it on the test set. Use the following distance metric as a measure of closeness in the input space:

$$D(\vec{x}, \vec{y}) = \sum_{i=1}^{N} |x_i - y_i|$$

2. What distance metric is the formula above? Why do you think it is a good choice for this problem?
3. Compare its performance on the training set and the test set in terms of accuracy and discuss if you trained a good classifier. Show the modelling process step-by-step and explain your answer.

Question 3 - Support Vector Machine (SVM)

1. You are given a dataset X with class attribute C. Explain the general procedure to train and evaluate a SVM model when the parameters of the SMV need to be optimised. SVM is just an example model here, the procedure would be the same with any model.

2. What strategies are SVM use when the data have more than two classes? Explain how **One-Vs-Rest** (OVR) and **One-Vs-One** (OVO) work. For example, a dataset contains 3 output classes: A, B, C. How many binary classifiers do you train for OVR and OVO? What are the output labels for each classifier?

3. What are the differences between SVM, logistic regression and linear regression?

Note:

Questions should be solved by hand. The explanation for each sub-question should be 200 words or less. Using figures to illustrate your solution is highly encouraged.