
Question 1.1.2 While Euclidean distance is a more commonly known method to measure distance, it's not the only one! Another method is the Manhattan distance. How do we calculate the Manhattan distance between two points? What situation might Manhattan distance be more suitable than Euclidean distance? **(4 points)**

Hint: Consider why it's named *Manhattan distance*—think of a city grid!

- Manhattan distance is calculated as the sum of the absolute differences of the points' coordinates. Supposed we are given 2 points (x1, y1), (x2, y2), their distance is: |x1 - x2| + |y1 - y2|.
- Manhattan distance is more suitable than Euclidean distance is situations where the movements are restricted to 4 directions up, down, left, right only, i.e., moving in grid-like spaces such as across building blocks, etc.

Question 1.7.1 When doing Knn classification we split our data into training and test sets.

Why do we divide our data into training and test sets? Or in other words what is the point of the training set? What is the point of the test set? Answer both questions. **(7 points)**

Hint: Check out this section in the textbook.

- Training set is used to build model that fits to the patterns of seen data, i.e. we repeatedly comparing the preditions to the true labels in training set, and changing the model's parameters until we obtain a model that gives overall good predictions for the samples in the training set.
- Test set is used to evaluate the performance of model. In reality, when deploying a model, it is likely that we would encounter an entirely new sample with values and features that are never seen before, or missing. Thus, test set is split to act as such unseen data, and give objective evaluation on the model's performance.

Question 1.7.2 Why do we only want to use the test set once? **(3 points)**

If we expose the test set to the model more than once, there is a high chance of leaking the data to the model, and as a consequence, the model would also try to fit to the test set and might give inauthentic good performance on test data. Using the test set more than once also goes against the point of train/test splitting.

Question 1.8. Why do we choose `k` to be an odd number in k-NN? Explain. **(10 points)**

`k` is usually an odd number to avoid ties when voting.