Answer to the questions:

1. Describe a data-science problem that you think would benefit from predictive or generative modeling. It can be something that you are interested in, relevant to your work, or have experienced working in the past. [1 pt]

Answer: I think predicting customers behaviour using ML is a very useful model for e-commerce overall. For example, NLP can predict and identify which words are the most popular when it comes to the particular product: “good”, “bad”, “useful” or “disappointing”. It can help the e-commerce business to better build their product line or customer services around products.

1. For the problem you identified in part 1, identify potential data sources and describe what they might look like. Hint: You don’t have to actually identify data sources, instead focus more on a very high-level hypothetical description (eg. images showing/from X and Y, tabular data with X, Y features, time series data, etc.) [4 pts]

Answer: For the problem of modelling customers reviews and feedback on the products, we can use the following data, for example from the website:

a) Product reviews. We can include data on product reviews, ratings and data on the reviewers. We can get this info from the official website or from Twitter as an option
b) Sales data. Analyze products that had been sold, what type and the amount, dates on when the products were purchased and demographics of those who purchased the times.
c) Information on the customers. More details on the customers like demographics,, purchasing habits, time spent on website, possible items that were reviewed by the customer could be very helpful for the business as well.

To analyze and present the above data we can use different methods such as tabular data with columns for each feature, or unstructured data in the form of text reviews or ratings from the customers. We might also include data images, such as product images or customer profile pictures. Whichever the data will be selected at the end, it would need to be preprocessed and cleaned before using it for predictive or generative modeling.

3. Do you think AI/ML would work well for the problem you’re describing? Why or why not? [5 pts]

Answer: Yes, I think ML and AI is the best way to analyze the above problem and provide the best solution for the business. ML will provide comprehensive tools and speed to do it fast and accurately which could not be done so manually. It would take much longer and the results might not be as useful by the time they will be available.

2. [5 pts] We would like for you to understand the differences between supervised and unsupervised learning. While near-infinite examples are available on Google and other search tools, we would like you to think about a personal area of interest and provide an example of a problem that would be well-suited for:

a. **Supervised learning**. Can be helpful to banks and other financial institutions, for example when trying to detect any possible credit card fraud activities. In this case we present the learning algorithm with transactions that are normal or possibly fraudulent.

b. **Unsupervised learning**. If we talk about customers and their behaviours, unsupervised learning would be beneficial when we are trying to investigate customer persona or detect any anomaly. For example, detecting bot activity which might act as a customer but certain patterns will help to determine it's not an actual customer.

3. It is important to understand the limitations of your data when choosing which ML models to implement. For this part of the assignment, we would like you to provide two examples of situations for k-Nearest Neighbours: one where it is a good choice, and one where it is a poor model choice.

a. Draw a two-dimensional example (two input features, one for each the x and y axes) of a dataset that would: [4 pts]

i. Perform well using k-NN

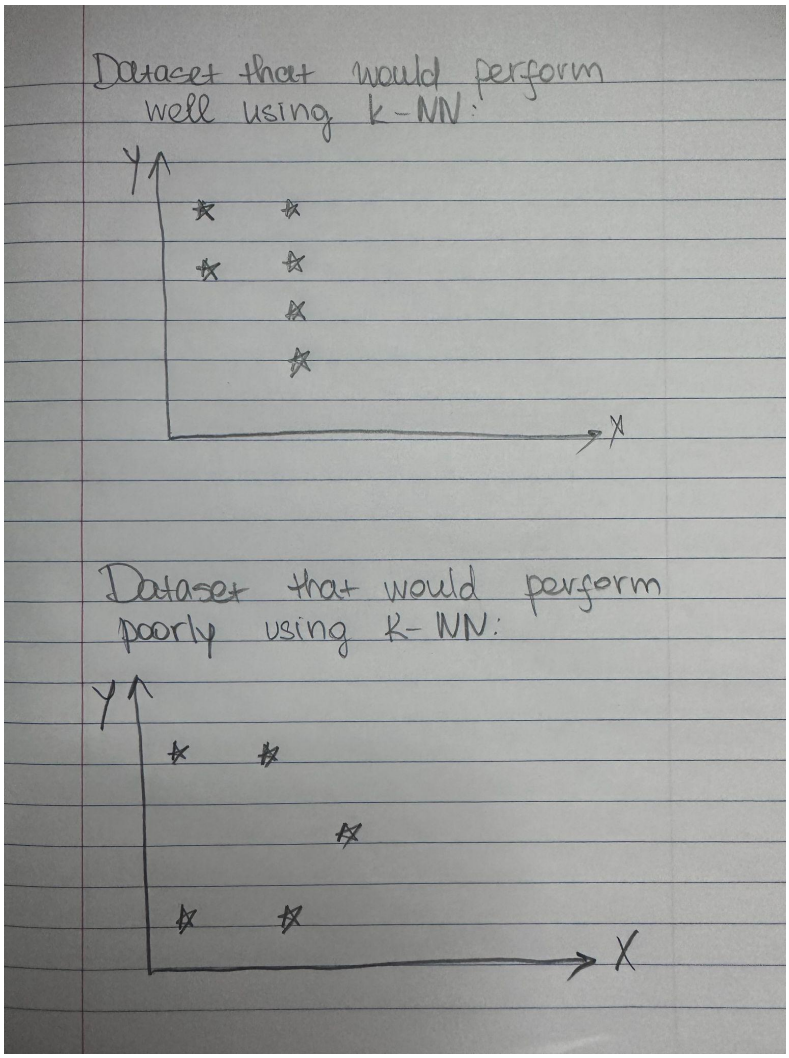
ii. Perform poorly using k-NN

b. For your above example, describe why you drew what you did (1-2 sentences). Would this depend on the value of k? [3 pts]

c. Why is it a good idea to use an odd number for k? [1 pt]

d. Why can't you simply pick a value of k that performs well on your training data? How do we commonly address this problem in ML? [2 pts]

Answer:



i. In the first dataset, data points are well separated and form distinct clusters. It is easy to determine which class the point belongs to based on the majority of k-NN clusters.

Ii. In the second dataset, the data points are not well separated and k-NN will have difficulties to determine and classify points close to the boundaries as the nearest neighbor can be evenly split between the 2 classes.

b. k-NN works perfectly when the data points are clearly separated in different classes as per my example 1. In the second poorly performing example, I presented two overlapping clusters and k-MM will have some problems separating it.

c. It's a good idea to have an odd k-value; it is avoiding a possible tie when we are finding the k-nearest neighbour.

d. If I simply pick a value of k that performs well on my training data, there will be a possibility of overfitting. To avoid that, we use techniques like cross-validation to address an issue.

GITHUB link with the notebooks:

4. [15 pts] Using the provided notebook re-implement this algorithm using the Iris dataset (more details in Colab notebook).

HINT: The Iris dataset is not a dataset of images and is instead tabular data with 4 features (the
petal length, the sepal length, the petal width, the sepal width). This will require you to comment out any sections of code that involve images (to comment out a line in Python, start that line with the # character).

a. Provide your Jupyter notebook so we can see your implementation [6 pts] - **Provided**

b. What was the optimal value of k for the Iris dataset and what was the test accuracy when it was evaluated?:

The best accuracy was 0.9166666666666666

The K values where the accuracy is highest are (array([5, 6, 7, 8, 9, 10, 11, 12, 13, 17, 18, 19, 20]),)

c. How does the value of k compare to what we observed for MNIST? Why do you think this is?

The optimal value of k in the Iris dataset is smaller compared to the MNIST results.

The best accuracy was 0.9925925925925926

The K values where the accuracy is highest are (array([0, 1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]),)

This may happen because the Iris dataset is a way simpler with fewer dimensions and classes. As a result, a smaller value of k may be sufficient to obtain high accuracy. Moreover, the Euclidean distance metric used for the Iris dataset may be more effective than the MNIST dataset.

d. When k becomes larger and close to the size of the training set the validation accuracy drops, why is that? [2pts]

When k becomes larger and close to the size of the training dataset, the validation accuracy drops because the model becomes more general and less flexible. It might lead to underfitting and the model will not catch complex relationships in data.

5. [10 pts] Using the same notebook you implemented for the IRIS dataset we would like to understand what happens when we work with lower dimensionality data. To simulate this only use the first two columns/features (sepal length and width) and re-run your model

HINT: This will involve changing your dataset immediately after using the sklearn load function.

a. What is the shape of the input data when you use only the first two features? [2 pt]

Answer: 150 * 2: 150 rows and 2 columns

b. What was the optimal value of K for the Iris dataset and what was the test accuracy when it was evaluated? [5 pts]

The best accuracy was 0.75

The K values where the accuracy is highest are (array([15, 16, 17, 53, 55, 56, 57]),)

c. How did this compare to when we previously used all of the data? [3 pts]

The accuracy is slightly lower.