COMP47750/COMP47990 Tutorial

Nearest Neighbour Classifiers

1. The table below shows three examples from the Penguins dataset. The two labelled examples are one Adelie and one Gentoo. The type of the query example is not known.
   The four descriptive features are:
   - Bill Length: numeric, with range [30,60]mm
   - Bill Depth: numeric with range [10,20]mm
   - Flipper Length: numeric with range [170, 230]mm
   - Body Mass: numeric with range [3,000, 6,000]g

Example: ***x1***	
Bill Length	39.1
Bill Depth	18.7
Flipper Len	181
Body Mass	3,750
Type	Adelie

Example: ***x2***	
Bill Length	50.2
Bill Depth	14.3
Flipper Len	218
Body Mass	5,700
Type	Gentoo

Query: ***q***	
Bill Length	39.5
Bill Depth	17.4
Flipper Len	186
Body Mass	3,800
Type	???

a) Normalise all numeric features to the range [0,1]

b) Propose an appropriate global distance function for comparing examples such as the above.

c) Use your proposed distance function to calculate the distances between the query example *q* and the two labelled examples. Which class label would a 1-NN classifier assign to the query based on the distances?

<Google Sheet>

2. The table below reports the pairwise distances between a set of 9 labelled training examples and a new query example *q*.

Example	**Class**	**Distance to *q***
x1	over	1.5
x2	under	2.8
x3	over	1.8
x4	under	2.9
x5	under	2.2
x6	under	3.0
x7	under	2.4
x8	over	3.2
x9	over	3.6

   a. What class label would a 3-NN classifier assign to *q*?
   b. What class label would a 4-NN classifier assign to *q*?
   c. What class label would a weighted 4-NN classifier assign to *q*?

3. Two different examples from a *k*-NN system for estimating the price of second-hand cars are shown in the tables below. Each example is described by 6 features.

Example: ***x1***	
Manufacturer	Ford
Model	Fiesta
Engine Size	1,100
Fuel	Petrol
Mileage	65,000
Condition	Excellent
Price	€3,100

Example: ***x2***	
Manufacturer	Citroen
Model	C3
Engine Size	1,800
Fuel	Diesel
Mileage	37,000
Condition	Fair
Price	€4,500

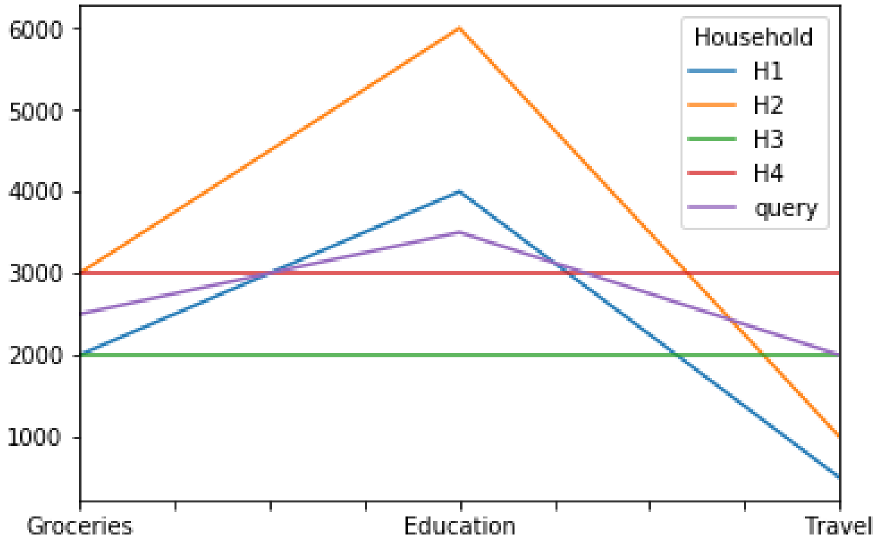
a) Normalise all numeric features to the range [0,1]. Assume that the feature ranges are:
- Engine Size 1,000 to 3,000
- Mileage 1,000 to 100,000

b) Propose a suitable global distance function that might be used in a *k*-Nearest Neighbour case retrieval system for this data. Assume that “Condition" is an ordinal feature that has the possible values {Poor, Fair, Good, Excellent},

c) Use the proposed global distance function to calculate the distance between the examples *x1* and *x2* above.

4. The data below shows households classified by how budget is allocated (‘Household.csv’).
   The notebook ’03 kNN Tutorial’ contains code to classify the query example using 1-NN and Euclidean distance.
   Modify this code so that correlation is used rather than Euclidean distance.

Household	Groceries	Education	Travel	Category
H1	2000	4000	500	C1
H2	3000	6000	1000	C1
H3	2000	2000	2000	C2
H4	3000	3000	3000	C2
query	2500	3500	2000	?



5. In the Data Normalisation example (Athlete Selection) in the 02-kNN Notebook replace the N(0,1) scaler with a min-max scaler.

6. The notebook 03 kNN Tutorial contains code to load the Sepsis dataset from the UCI repository (https://archive.ics.uci.edu/dataset/827/sepsis+survival+minimal+clinical+records)

   This dataset is divided into train and test sets and scaled. Then a *k*NN classifier is trained and tested. The time to classify the test data is also recorded.

   scikit-learn provides two strategies to speed up *k*-NN, `ball_tree` and `kd_tree`, see details here:

   https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

   Compare the performance of these two algorithms with brute force search `brute`.