Lab Sheet

Nearest Neighbour Classifiers

0. Work through the "02 k-NN" notebook.

1. Three examples are shown below from the "penguins" dataset. Each example is represented by a vector of 4 numeric features. Example *x1* has been manually labelled as belonging to "Class A", while Example *x2* has been labelled as belonging to "Class B".

Example: *x1*	
Bill length	4.4
Bill depth	2.9
Flipper length	1.4
Body Mass	0.2
Class	A

Example: *x2*	
Bill length	5.6
Bill depth	3.0
Flipper length	4.5
Body Mass	1.5
Class	B

Query: *q*	
Bill length	6.1
Bill depth	3.0
Flipper length	4.6
Body Mass	1.4
Class	???

   a) What type of distance function might be appropriate for comparing the examples above?

   b) Use this 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?

2. The table below shows three examples from a system for predicting whether a person is over or under the drink driving limit. The 5 input features for this system are:

- Gender: categorical feature {male, female}
- Weight: numeric, with range [50,150]
- Amount of alcohol in units: numeric, with range [1,16]
- Meal type: ordinal feature {None, Snack, Lunch, Full}
- Duration of drinking session: numeric, range [20,230]

Example: *x1*	
Gender	female
Weight	60
Amount	4
Meal	full
Duration	90
Class	over

Example: *x2*	
Gender	male
Weight	75
Amount	2
Meal	full
Duration	60
Class	under

Query: *q*	
Gender	male
Weight	70
Amount	1
Meal	snack
Duration	30
Class	???

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

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

d) 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?

3. The table below reports the pairwise distances between a set of 9 labelled training examples and a new query example *q*, for the system described in Question 2.

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*?

4. Case-based Reasoning (CBR) is a reasoning approach that uses k-NN to retrieve the most similar examples to query cases and uses these to make decisions about the query case.
(For information on CBR see Aamodt and Plaza 2001 seminal paper available in Brightspace under the Reading Unit).
Two different examples from a CBR 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	BX
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.

5. The data below shows households classified by how budget is allocated ('Household.csv').
The notebook '02 kNN Lab Sheet' contains code to load in this dataset. Add in code to classify the query example using 1-NN and Euclidean distance. In this example households are classified based on how budget is allocated, correlation would be a better measure of similarity. 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	?

6. In the Data Normalisation example in the "02-kNN" Notebook replace the N(0,1) scaler with a min-max scaler. Are there any differences?

7. Download the zip file '02-BYO kNN-Python focus'. It contains a notebook that takes you through building your own kNN classifier in Python with a significant focus on writing good Python code. Work through this notebook if you are interested.