COMP3010J Tutorial

Nearest Neighbour Classifiers

1. 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

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 Example	
Gender	male
Weight	70
Amount	1
Meal	snack
Duration	30
Class	???

features for this system are:

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

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

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

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

2. 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**?

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

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

Example 007	
Manufacturer	Ford
Model	Fiesta
Engine Size	1,100
Fuel	Petrol
Mileage	65,000
Bodywork	Excellent
Price	€3,100

Example 014	
Manufacturer	Citroen
Model	BX
Engine Size	1,800
Fuel	Diesel
Mileage	37,000
Bodywork	Fair
Price	€4,500

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 Bodywork is an ordinal feature that has the possible values {Poor, Fair, Good, Excellent},

c) Use the proposed distance function to calculate the distance between the two examples above.