

COMP47460 Tutorial

Nearest Neighbour Classifiers

1. Three examples are shown below from the Iris botanical dataset. Each example is represented by a vector of 4 numeric features. Example 1 has been manually labelled as belonging to “Class A” and Example 2 has been labelled as belonging to “Class B”.



Example x1		Example x2		Query Example	
<i>Sepal length</i>	4.4	<i>Sepal length</i>	5.6	<i>Sepal length</i>	6.1
<i>Sepal width</i>	2.9	<i>Sepal width</i>	3.0	<i>Sepal width</i>	3.0
<i>Petal length</i>	1.4	<i>Petal length</i>	4.5	<i>Petal length</i>	4.6
<i>Petal width</i>	0.2	<i>Petal width</i>	1.5	<i>Petal width</i>	1.4
<i>Class</i>	A	<i>Class</i>	B	<i>Class</i>	???

- 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 and the two labelled examples. Which class label would a 1NN 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 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]

Example x1		Example x2		Query Example	
Gender	female	Gender	male	Gender	male
Weight	60	Weight	75	Weight	70
Amount	4	Amount	2	Amount	1
Meal	full	Meal	full	Meal	snack
Duration	90	Duration	60	Duration	30
Class	over	Class	under	Class	???

- Normalise all numeric features to the range [0,1]
- Propose an appropriate global distance function for comparing examples such as the above.
- 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?

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. Two different examples from a Case-based reasoning (CBR) system for estimating the price of second-hand cars are shown in the tables below. Each example is described by 6 features.

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

Example 014	
<i>Manufacturer</i>	Citroen
<i>Model</i>	BX
<i>Engine Size</i>	1,800
<i>Fuel</i>	Diesel
<i>Mileage</i>	37,000
<i>Bodywork</i>	Fair
<i>Price</i>	€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 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.