COMP30120 Tutorial

Nearest Neighbour Classifiers

Derek Greene

School of Computer Science and Informatics

Autumn 2015

Tutorial Q1

- Three examples are shown below from the Iris dataset:
  - Each represented by a vector of 4 numeric features.
  - Example 1: Class A
  - Example 2: Class B

Example 1	
Sepal length	4.4
Sepal width	2.9
Petal length	1.4
Petal width	0.2
Class	A

Example 2	
Sepal length	5.6
Sepal width	3.0
Petal length	4.5
Petal width	1.5
Class	B

Target Example	
Sepal length	6.1
Sepal width	3.0
Petal length	4.6
Petal width	1.4
Class	???

a. What type of distance measure might be appropriate for comparing the examples above?

b. Use this distance measure to calculate the distances between the target example and the labelled examples. What class label should be assigned to the target?

Tutorial Q1

- Euclidean distance: Common distance measure between two numeric inputs.

$$\mathrm{ED}(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{f \in F} (q_f - p_f)^2}$$

ED (Target, Example1)

$$\sqrt{(6.10 - 4.40)^2 + (3.00 - 2.90)^2 + (4.60 - 1.40)^2 + (1.40 - 0.20)^2} = 3.82$$

ED (Target, Example2)

$$\sqrt{(6.10 - 5.60)^2 + (3.00 - 3.00)^2 + (4.60 - 4.50)^2 + (1.40 - 1.50)^2} = 0.52$$

- Distance to Example 2 is smaller

  ➡ Assign to Class B

Tutorial Q2

- Three cases from a regression system for estimating Blood-Alcohol Content (BAC). The 5 input features are:
  - Gender, Framesize (i.e. weight), Amount of alcohol in units, Meal type, Duration of drinking session.

Case 1

Gender	Male
FrameSize	1
Amount	1
Meal	snack
Duration	60
BAC	0.2

Case 2

Gender	Female
FrameSize	3
Amount	4
Meal	full
Duration	85
BAC	0.8

Case 3

Gender	Male
FrameSize	1
Amount	3
Meal	snack
Duration	120
BAC	0.7

a. Normalise all numeric features to the range [0,1]. Note that you can assume that the feature ranges for: Framesize is [1,6], Amount is [1,16], Duration is [20,300].

b. Propose a suitable custom similarity measure for comparing cases such as the above. The similarity measure should take account of the fact that Meal is an ordinal feature taking values {None, Snack, Lunch, Full}.

c. Use your proposed similarity measure to calculate the similarities between Case 1 and the other two cases. Which case is the nearest neighbour of Case 1?

Tutorial Q2a

a. Normalise all numeric features to the range [0,1]. Note that you can assume that the feature ranges for: Framesize is [1,6], Amount is [1,16], Duration is [20,300].

- Min-max normalisation:
  Use min and max values for a given feature to rescale to the range [0,1]

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Case 1

Gender	Male
FrameSize	(1-1)/(6-1) = 0
Amount	(1-1)/(16-1) = 0
Meal	snack
Duration	(60-20)/(300-20) = 0.143

Case 2

Gender	Female
FrameSize	(3-1)/(6-1) = 0.4
Amount	(4-1)/(16-1) = 0.2
Meal	full
Duration	(85-20)/(300-20) = 0.232

Case 3

Gender	Male
FrameSize	(1-1)/(6-1) = 0
Amount	(3-1)/(16-1)= 0.133
Meal	snack
Duration	(120-20)/(300-20) = 0.357

Tutorial Q2b

b. Propose a suitable custom similarity measure. The similarity measure should take account of the fact that Meal is an ordinal feature taking values {None, Snack, Lunch, Full}.

$$d(q, x_i) = \sum_{f \in F} w_f \cdot \delta(q_f, x_f) \qquad d(q_f, x_{if}) = \begin{cases} 0, & \text{if } f \text{ discrete and } q_f = x_{if} \\ 1, & \text{if } f \text{ discrete and } q_f \neq x_{if} \\ |q_f - x_{if}|, & \text{if } f \text{ continuous} \end{cases}$$

Numeric features: Calculate the absolute value of the difference between the feature values.

Ordinal features: the distance can be the absolute difference between the two positions in the ordinal list of possible values.

- Meal: {None, Snack, Lunch, Full} = {1, 2, 3, 4}

e.g. `difference(Snack,Full) = |2-4| = 2`

Note: In practice, we may often normalise with respect to ordinal list length

e.g. `|2-4|/4 = 0.5`

Tutorial Q2b

Feature	Type	Difference Measure
Gender	Categorical	Binary
FrameSize	Numeric	Absolute difference (after normalisation)
Amount	Numeric	Absolute difference (after normalisation)
Meal	Ordinal {None, Snack, Lunch, Full}	Absolute relative rank difference
Duration	Numeric	Absolute difference (after normalisation)

Calculate weighted sum of difference measure values.

Weight features to contribute equally *w=1/5=0.2*

$$d(q, x_i) = \sum_{f \in F} w_f \cdot \delta(q_f, x_f)$$

Convert a distance to a similarity using inverse

$$sim(q, x_i) = 1 - d(q, x_i)$$

Tutorial Q2c

Sum over features. Weight features to contribute equally *w=1/5=0.2*

$$d(q, x_i) = \sum_{f \in F} w_f \cdot \delta(q_f, x_f)$$

Convert a distance to a similarity using inverse

$$sim(q, x_i) = 1 - d(q, x_i)$$

Sim (Case1,Case2)

Feature	Difference
Gender	1
FrameSize	|0-0.4| = 0.4
Amount	|0-0.2| = 0.2
Meal	|2-4| = 2
Duration	|0.143-0.23| = 0.089

```
dist = (0.2*1) + (0.2*0.4) +
(0.2*0.2) + (0.2*2) +
(0.2*0.089) = 0.7379
```

```
sim = 1 - 0.7379 = 0.2621
```

Sim (Case1,Case3)

Feature	Difference
Gender	0
FrameSize	|0-0|=0
Amount	|0-0.133| = 0.133
Meal	|2-2| = 0
Duration	|0.143-0.357| = 0.214

```
dist = (0.2*0) + (0.2*0) +
(0.2*0.133) + (0.2*0) +
(0.2* 0.214) = 0.0695
```

```
sim = 1 - 0.0695 = 0.9305
```

Tutorial Q3

- Two cases from a system for estimating the price of second-hand cars are shown below.

Case 007

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

Case 014

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

a. Normalise all numeric features to the range [0,1]. Note that you can assume that the feature ranges for: Engine Size is 1,000 to 3,000; Mileage is 1,000 to 100,000.

b. Propose an appropriate custom similarity measure. Assume that Bodywork is an ordinal feature that has the possible values {Poor, Fair, Good, Excellent},

c. Use this measure to calculate the similarity between the two cases above.

d. How might this system be improved by using feature weights?

Tutorial Q3a

a. Normalise all numeric features to the range [0,1]. Note that you can assume that the feature ranges for: Engine Size is 1,000 to 3,000; Mileage is 1,000 to 100,000.

- Min-max normalisation:
  Use min and max values for a given feature to rescale to the range [0,1]

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Case 007

Manufacturer	Ford
Model	Fiesta
Engine Size	(1100-1000)/ (3000-1000) = 0.05
Fuel	Petrol
Mileage	(65000-1000)/ (100000-1000) = 0.646
Bodywork	Excellent

Case 014

Manufacturer	Citroen
Model	BX
Engine Size	(1800-1000)/ (3000-1000) = 0.4
Fuel	Diesel
Mileage	(37000-1000)/ (100000-1000) = 0.364
Bodywork	Fair

Tutorial Q3b

Feature	Type	Difference Measure
Manufacturer	Categorical	Binary difference
Model	Categorical	Binary difference
Engine Size	Numeric	Abs difference
Fuel	Categorical	Binary difference
Mileage	Numeric	Abs difference
Bodywork	Ordinal	Abs difference of position

Calculate weighted sum of difference measure values.

Weight features to contribute equally *w=1/6=0.167*

$$d(q, x_i) = \sum_{f \in F} w_f \cdot \delta(q_f, x_f)$$

Convert a distance to a similarity using inverse

$$sim(q, x_i) = 1 - d(q, x_i)$$

Tutorial Q3c

Case 007 (Normalised)

Manufacturer	Ford
Model	Fiesta
Engine Size	0.05
Fuel	Petrol
Mileage	0.646
Bodywork	Excellent

Case 014 (Normalised)

Manufacturer	Citroen
Model	BX
Engine Size	0.4
Fuel	Diesel
Mileage	0.364
Bodywork	Fair

Sim(Case 007, Case 014)

Feature	Difference
Manufacturer	1
Model	1
Engine Size	|0.05-0.4| = 0.35
Fuel	1
Mileage	|0.646-0.364| = 0.283
Bodywork	|4-2| = 2

```
dist = (0.167*1) + (0.167*1)
+ (0.167*0.35) + (0.167*1) +
(0.167*0.283) + (0.167*2)
= 0.9388
```

```
sim = 1 - 0.9388 = 0.0612
```

* subject to rounding

Tutorial Q3d

- *Feature weighting* could be used to emphasise certain features that are more important in the domain:

$$d(q, x_i) = \sum_{f \in F} w_f \cdot \delta(q_f, x_f)$$

Feature	w_f
Manufacturer	0.1
Model	0.1
Engine Size	0.3
Fuel	0.1
Mileage	0.3
Bodywork	0.1

Example:
Provide higher weighting to the features “car mileage” and “engine size”