Nearest neighbour methods

1. Nearest neighbour
2. Similarity measures
3. Scoring: Look-Alike models
4. Combining scores
5. Nearest neighbour issues

1. Nearest neighbour methods

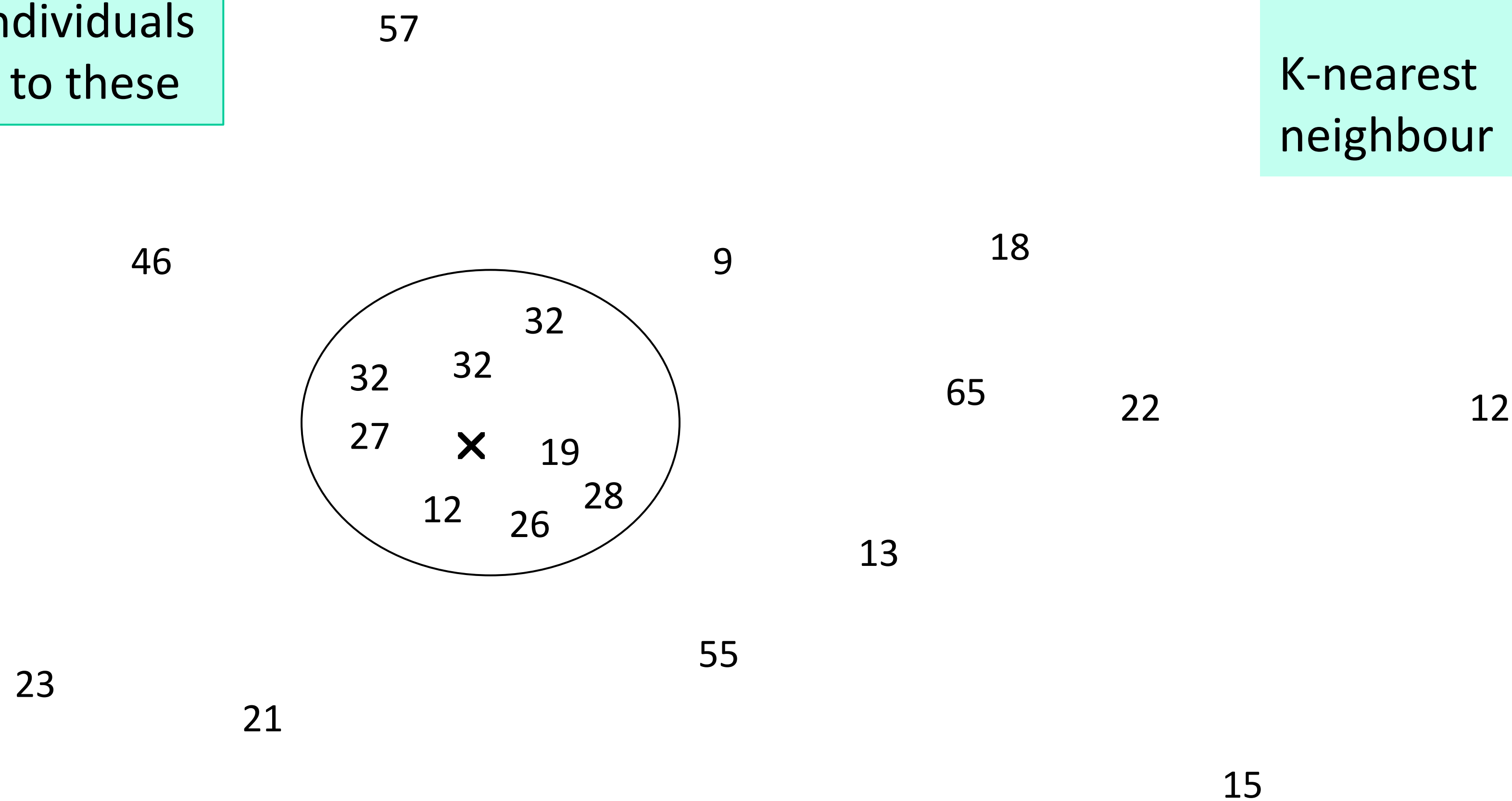
- Suppose we are trying to predict a new customer's income from information we have about their age, employer, type of employment, postcode, type of house, . . .
- We have a database of known customers (for whom the income is known)
- Can we predict the income of the new employee by *combining* the incomes of those (known) employees who are *similar* to the new employee (with respect to the known attributes – age, employer, …)
  - a rationale we've already employed with decision trees

Predicting income

Suppose we have two numeric input variables – and we plot the individuals according to these

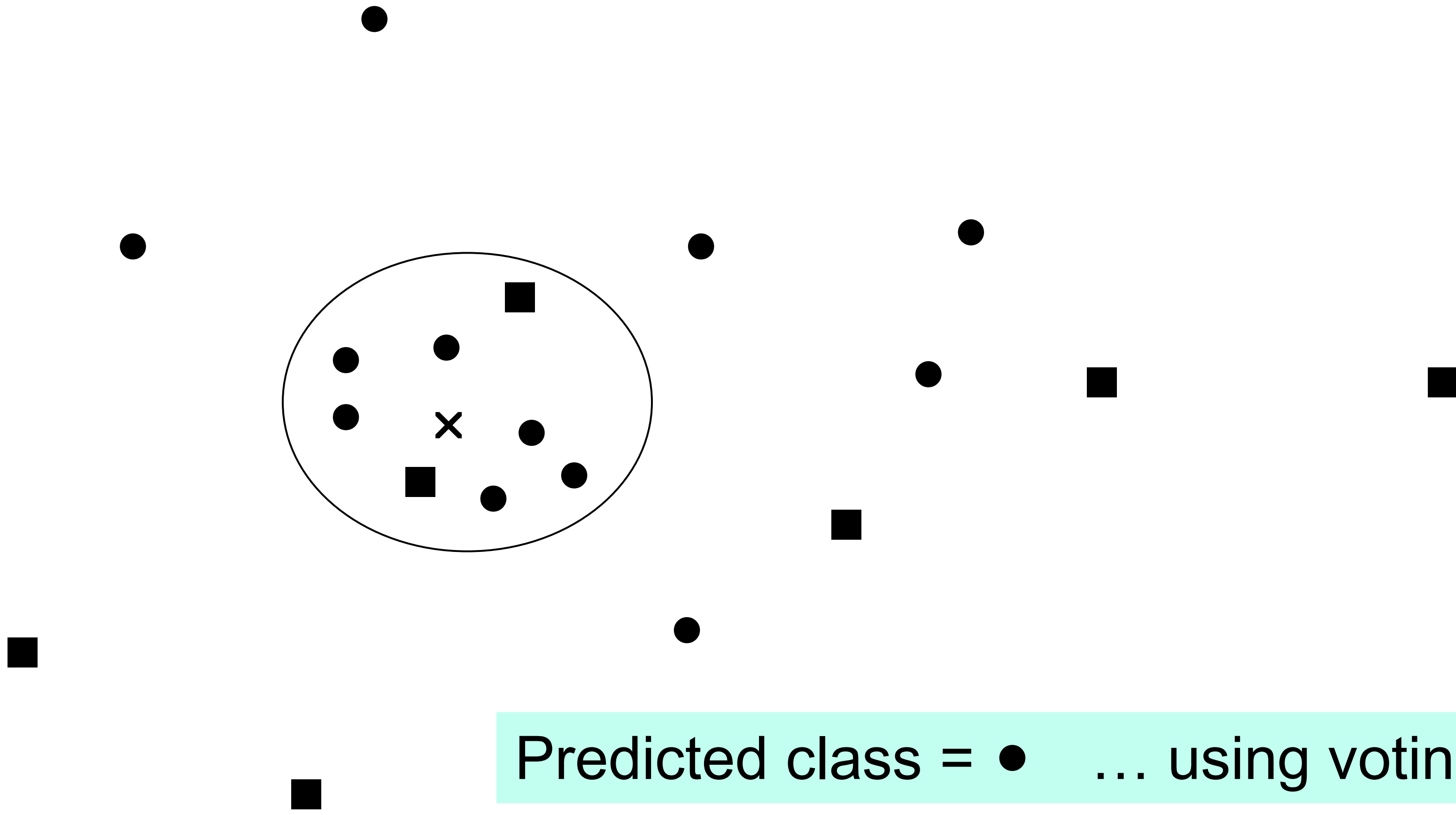
Estimation

K-nearest neighbour



Perhaps then predicted income = avg {32, 32, 32, 27, 12, 26, 28, 19}

Nearest neighbour applied to classification



Predicted class = ● … using voting

Applications

- Insurance claims that are similar to historical fraudulent claims …
- Customer response modelling
- Predicting customer preferences
- Medical treatment
- Medical image diagnosis
- Predicting property prices
- …

Nearest neighbour methods

Require:

- A collection of historical data values
  - The training set
- A similarity (or distance) measure
- A combination function

2. Similarity measures

- Suppose we have just two attributes, Age and Salary
- We can then measure the distance between two individuals as

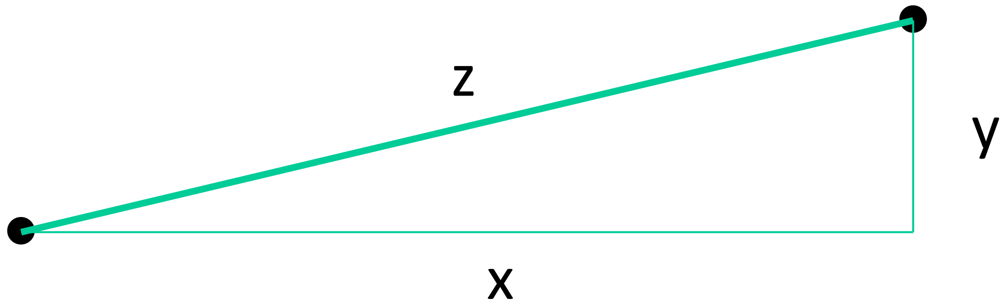
$$\text{sqrt}[\ (\text{Age}_1 - \text{Age}_2)^2 + (\text{Salary}_1 - \text{Salary}_{\mathbf{2}})^2\ ]$$

difference in age = $|\text{Age}_1 - \text{Age}_2|$

difference in salary = $|\text{Salary}_1 - \text{Salary}_2|$

Euclidean distance

A
S
I
D
E



By Pythagoras' theorem: $z^2 = x^2 + y^2$ … or if you prefer

$$z = \mathrm{sqrt}[x^2 + y^2]$$

Example

- Suppose our two individuals have Age and Salary as:

  (61, 43000) & (27, 40000)

  then the simple Euclidean distance between the two individuals is

  $$\mathrm{sqrt}[34^2 + 3000^2] = 3000.2$$

- … the difference in Age is not reflected in the distance between the individuals

Normalising the data ranges

- Instead of taking the raw difference in values, we can normalise the difference by looking at the difference divided by the maximum possible difference
- E.g., $|Age_1 – Age_2| / 80$
- … and $|Salary_1 - Salary_2| / 150000$

Normalising the data ranges

- E.g., $|Age_1 – Age_2| / 80$
- … and $|Salary_1 - Salary_2| / 150000$

then the difference between the two individuals becomes

$$sqrt[(34/80)^2 + (3/150)^2] = 0.4255$$

34/80 = 0.425; 3/150 = 0.02

Normalising the data

- More rigorously we can use the z-value
- Given two ages: Age_1 and Age_2 their (z-value) difference is measured as

$$|Age_1 - Age_2|/\sigma$$

where σ is the standard deviation of the Age-values

Similarity measures

For each attribute A we need a distance function $dist_A$ that measures the distance between two A-values

- Given n attributes $A_1, A_2, \ldots, A_n$ and two individuals $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$, then the distance between them can be defined as

$$\mathrm{sqrt}[\, dist_{A1}(x_1, y_1)^2 + dist_{A2}(x_2, y_2)^2 + \ldots + dist_{An}(x_n, y_n)^2 \,]$$

or as

$$dist_{A1}(x_1, y_1) + dist_{A2}(x_2, y_2) + \ldots + dist_{An}(x_n, y_n)$$

Similarity measures

- For each attribute A we need a distance function $dist_A$ that measures the distance between two A-values
  - for numeric variables this is easy
- For a simple categorical variable like colour we could define

$$dist_{col}(Col_1, Col_2) = 0 \text{ if } Col_1 = Col_2$$

$$dist_{col}(Col_1, Col_2) = 1 \text{ if } Col_1 \neq Col_2$$

Similarity measures - postcodes

We can see a geographic hierarchy of (at least) 3 levels

- Postcode: e.g., TR9 2DW
- District: e.g., TR9
- Area: e.g., TR

In reality it is more complicated than this

A postcode distance measure

- $dist(p_1,p_2) = 0$ if $p_1=p_2$
- $dist(p_1,p_2) = 0.2$ if $p_1 \neq p_2$ but $district(p_1)=district(p_2)$
- $dist(p_1,p_2) = 0.4$ if $district(p_1) \neq district(p_2)$ but $area(p_1)=area(p_2)$
- $dist(p_1,p_2) = 1$ if $area(p_1) \neq area(p_2)$

In reality it is more complicated than this

A web visitor similarity measure

- E.g., we can define a distance function between 2 visitors via:
- CP = # pages visited by both
- CA = # pages visited by neither
- P = total # of pages at the site, then

$$\text{Similarity} = (CP + CA)/P$$

…requires some ingenuity

Other similarity measures

… can become more complicated when looking at data types for

- Images
- Audio
- Text
- Time series

3. Scoring: Look–Alike models

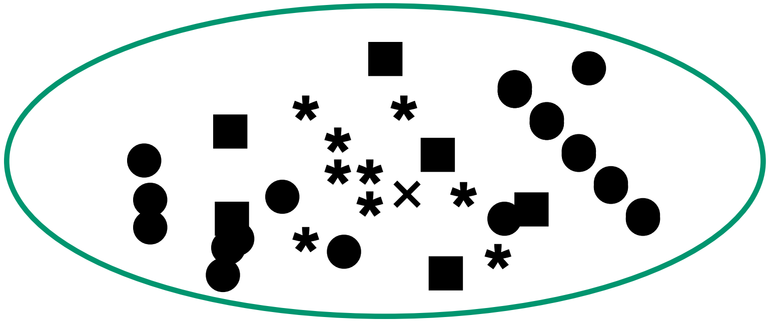
- A simplified form of nearest neighbour
- Given a new individual for scoring, pick the *single* individual in the training set that is the most similar
  - Its score then provides the score for the new individual
- Simple – and in many cases overly simple
  - E.g., predicting income

Look-Alike for paired testing

- e.g., a company wishes to measure the effectiveness/impact of a business strategy – how can it be sure that the observed effects are indeed down to the strategy?
- Subdivide stores into *similar* pairs – and for each pair apply the business strategy to just one of the pair

4. Scoring: Using voting

- Scoring requires us to combine the scores of the nearest neighbours
- For categorisation, we can employ "voting"
  - Obviously with many classes, there is a danger that the prediction will be weak, and more dependent upon K
  - Alternatively can use weighted voting



Scoring: using averaging

- For estimation we can – as earlier - take the average, or weighted average, of the scores of the nearest neighbours

Collaborative filtering

- Product recommendation
- Each individual has known preferences
- Individuals can be compared with other individuals
- Recommendations for a new individual are based upon weighted preferences of similar individuals

Properties of the score

- The score falls within the range of existing values … it is in some sense "reasonable"
  - But not good for extrapolation
- For estimation - the set of possible values is very, very large
- For classification – all classes are possible
  - Given an appropriate training set

5. Nearest neighbour issues

- Training set needs to provide coverage of underlying population
  - Representative samples may not
- When applied to classification, we may oversample so that each class is sufficiently represented
  - May need tens/hundreds/thousands of data points for each class

Properties of nearest neighbour

- Requires *interval data*
  - where you can define intervals – and hence measure distance
- But conversely only requires interval data … there are no other constraints
  - It can deal with complex underlying data types

High dimensional spaces

- High dimensional spaces
  - may not be sensible to use nearest neighbours
    - A million points in a 2-d square will fill the space
    - A million points in a 20-d cube will be sparse --- and moreover the distance between any two points becomes more uniform
- Suggests the use of preliminary or statistical analysis to reduce the number of input variables

Nearest neighbour issues

- Similar to decision trees in rationale
  - But here the subsets used are not disjoint
- Generates no underlying theory
  - black box; no model or rule is generated
  - with decision trees we use the historical data to build a model that is then used to score the data

Nearest neighbour issues

- Computational complexity
  - Scoring a new individual requires us to compare the new individual against all records in the historical data set
    - can become very expensive when we have many new records to score
  - methods exist to prune the training set

Choosing K

- Choosing K is obviously important
  - Too small and idiosyncrasies will be picked up
  - Too large and non-related individuals fall in the neighbourhood
- If the predictions are not stable (wrt K) then we may have to accept that the training data is insufficient to make predictions