Nearest Neighbour Classification (kNN)

Nearest Neighbour Classification (kNN)

- An instance based learner.
- Uses k “closest” points (nearest neighbors) for performing classification.
- Find the k nearest neighbours in the ‘training’ data
- Take the majority class of these neighbours.
  - (Or weight the vote according to distance weight factor, $w = 1/d^2$)
- There is no model so no training but classification is slow.

Typical Applications

- Computer Vision applications.
  - OCR, facial recognition etc.
- Patterns in genetic data.
- Medical diagnosis (tumors)
- Difficult to define concepts but you 'know it when you see it'.

Distance Metric

- Use Euclidean distance to compute distance between two points:

$$d(p,q)=\sqrt{\sum_i (p_i-q_i)^2}$$

Scaling

- Attributes may have to be scaled to prevent distance measures from being dominated by one of the attributes
- Example:
  - area_mean: 143.5 -> 2501
  - smoothness_mean: 0.052 -> 0.163
- If not scaled, smoothness_mean will have minimal contribution.

Scaling Methods

min-max normalization

$$x_{new} = \frac{x - min(x)}{max(x) - min(x)}$$

z-score normalization

$$x_{new} = \frac{x - \mu}{\sigma} = \frac{x - mean}{std.dev.}$$

Scaling Methods

- min-max normalization scales to between 0 and 1.
- z-score normalization counts the numer of standard deviations from the mean.
- For a normal distribution, 99% of values are within 3 standard deviations of the mean.

Choosing the value of k:

- If k is too small, sensitive to noise points
- If k is too large, neighborhood may include points from other classes.
- k can be obtained using cross validation.

Non-parameteric methods

- kNN is an example of a non-parameteric method.
- It stores the ‘training’ data and in effect does not build a model.
- Hence it does not need to determine any parameters of a model.
- Note that the value of k is a meta-parameter.
- It can only be determined by looking at performance on validation/testing data.

KNN - Summary

- kNN classifiers are lazy learners
- Does not build models explicitly
- Classifying unknown records is relatively expensive.
- Distances to every point in the ‘training’ data has to be calculated.
- The bigger the ‘training’ data the more expensive prediction is.