Today's topics

1. Classification
2. Nearest neighbour classification
3. Decision boundary
4. Tips on pre-processing data
5. Generalisation and over-fitting

Types of learning problems

<table>
<thead>
<tr>
<th>Data</th>
<th>System</th>
<th>Type of problem</th>
<th>Type of learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>$(x)$</td>
<td>groups (subsets)</td>
<td>clustering</td>
</tr>
<tr>
<td>$x$</td>
<td>$y$</td>
<td>discrete category</td>
<td>classification</td>
</tr>
<tr>
<td>$x$</td>
<td>$y$</td>
<td>continuous value</td>
<td>regression</td>
</tr>
</tbody>
</table>

where $x = (x_1, \ldots, x_D)^T$: feature vector  
$y$: target vector or scalar

Supervised learning

Test mode

Classification

New data: Label

Goal of training: develop a classifier of good generalisation

Classifying test data with $K$-nearest neighbours

1-nearest neighbour
K-NN classification algorithm

For each test example \( z \in Z \):
\[ \text{Compute the distance } r(z,x) \text{ between } z \text{ and each training example } (x,c) \in X \]
\[ \text{Select } U_k(z) \subseteq X, \text{ the set of the } k \text{ nearest training examples to } z \]
\[ \text{Decide the class of } z \text{ by the majority voting:} \]
\[ c(z) = \arg \max_{j \in \{1,...,C\}} \sum_{(x,c) \in U_k(z)} \delta_j c \]

Decision boundary: boundary (surface) that partitions the vector space into subsets of different classes.

Decision regions: regions separated by the decision boundaries.
Classification and nearest neighbours

**Decision boundaries for $C = 3$**

![Decision boundaries for $C = 3$](image)

**What $K$ should we use?**

![An example where a large $K$ reduces noise](image)

$K = 1$

$K = 15$

(Black curve: KNN decision boundary, darkened purple curve: the Bayes decision boundary)

**LANDSAT Application**

Predict land-usage from satellite data

KNN applied to 9 pixel patch in 4 spectral bands, with $K = 5$

**Tips on pre-processing data**

![Tips on pre-processing data](image)

- Different units
  - Standardise features unless understand units

- Consider transformation, e.g. log-transform.

**Generalisation and over-fitting**

How reasonable is this decision boundary?

**Poor generalisation: stories**

In a competition:


Classic stories:

http://neil.fraser.name/writing/tank/

http://www.j-paige.org/dobbs/neural_net_urban_legends.html

**How reliable is the error rate?**

- Error rate on training data set:
  - can be $\sim 0\%$
  - useless to estimate generalisation error

- Error rate on a test data set (exclusive to the training set)
  - How large should the data set be?
  - How should it be collected?

  *Cross validation* is used to estimate generalisation error (swapping test and training data sets)

  - $k$-fold cross validation ($k$-fold CV)
    - 2-fold CV is sometimes called ‘holdout method’
    - leave-one-out cross validation (LOO CV)
## Summary

- **Classification with similarity based methods**
  - Represent items as feature vectors
  - Compute distances to other items and sort
  - Assign a class label to the feature vector
  - \(k\)-NN: an example-based approach that classifies a test point based on the classes of the closest training samples
  - Larger \(k\) results in a smoother solution
  - Decision boundaries/regions, Voronoi diagram

- **Generalisation**
  - Overfitting: tuning a classifier to closely to the training set can reduce accuracy on the test set
  - Compare methods on held out data (validation set)
  - Estimate final performance on really new data (test set)

## Lab 4

**Friday 09 Feb. Lab-4 K-NN classification**

- Fridays at 14:10-15:00 in AT-5.05

## Tutorial Week 04

**Simple recommender system and clustering**

- Work on the questions in advance to identify what you understand and what you don’t. (avoid attending the tutorial without any preparation)
- Be active/positive - prepare topics that you’d like to discuss at the tutorial
- Try writing Matlab code of your own