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, y))</td>
<td>(x)</td>
<td>y: discrete category</td>
<td>classification</td>
</tr>
<tr>
<td>((x, y))</td>
<td>(y)</td>
<td>y: 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**

- **Classification**
  - Test mode
  - New data
  - Label
  - Goal of training: develop a classifier of good generalisation

**Classification**

- The data has a feature vector \(x = (x_1, x_2, \ldots, x_D)^T\) and a label \(c \in \{1, \ldots, C\}\)
- Training set: A set of \(N\) feature vectors and their labels \((x_1, c_1), \ldots, (x_N, c_N)\)
- Use a learning algorithm to train a classifier from a training set
- Test set: a set of feature vectors to which the classifier must assign labels — used for evaluation. (NB: training and test sets should be mutually exclusive)
- Error function: how accurate is the classifier? One option is to count the number of misclassifications:
  \[
  \text{Error rate} = \frac{\# \text{ of misclassified samples}}{\# \text{ of test samples}}
  \]

**Nearest-neighbour classifier**

- Nearest neighbour classification: label a test example to have the label of the closest training example
- \(K\)-nearest neighbour (\(K\)-NN) classification: find the \(K\) closest points in the training set to the test example; classify using a majority vote of the \(K\) class labels
- Training a \(K\)-nearest neighbour classifier is simple! — Just store the training set
- Classifying a test example requires finding the \(K\) closest training examples
  - This is computationally demanding if the training set is large — potentially need to compute the Euclidean distance between the test example and every training example
  - Data structures such as the kD-tree can make finding nearest neighbours much more efficient (in the average case)

**Classifying test data with \(K\)-nearest neighbours**

**1-nearest neighbour**

---

**Supervised learning**

![Graph showing decision boundary and classification]

**Graph showing decision boundary and classification**

- Oranges:
- Lemons:

---

**Graph showing decision boundary and classification**

- Oranges:
- Lemons:
K-NN classification algorithm

For each test example \( z \in Z \):
1. Compute the distance \( r(z, x) \) between \( z \) and each training example \( (x, c) \in X \).
2. Select \( \mathcal{U}_k(z) \subseteq X \), the set of the \( k \) nearest training examples to \( z \).
3. Decide the class of \( z \) by the majority voting:
   \[
   c(z) = \arg \max_{j \in \{1, \ldots, C\}} \sum_{(x,c) \in \mathcal{U}_k(z)} \delta_{jc}
   \]
13.4. k-nearest-neighbors on the two-class mixture data. The upper panel shows the misclassification boundary.


How reasonable is this decision boundary?

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

Tips on pre-processing data

Wisconsin Diagnostic Breast Cancer (WDBC) data set http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin(Diagnostic)

Tips on pre-processing data

Generalisation and over-fitting

Poor generalisation: stories

How reliable is the error rate?

In a competition:


Classic stories:

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

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

Tips on pre-processing data

Wisconsin Diagnostic Breast Cancer (WDBC) data set http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin(Diagnostic)

Tips on pre-processing data

Wisconsin Diagnostic Breast Cancer (WDBC) data set http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin(Diagnostic)

Tips on pre-processing data

Wisconsin Diagnostic Breast Cancer (WDBC) data set http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin(Diagnostic)
Cross validation

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)

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

Lab 4 & Lab 5

08 Feb. Lab-4 K-means clustering and visualisation with PCA.

15 Feb. Lab-5 K-NN classification and Naive Bayes

- Wednesdays at 16:10-17:00 in FH-3.D01