Inf2b - Learning

Lecture 4: Classification and nearest neighbours

Hiroshi Shimodaira
(Credit: Iain Murray and Steve Renals)

Centre for Speech Technology Research (CSTR)
School of Informatics
University of Edinburgh

http://www.inf.ed.ac.uk/teaching/courses/inf2b/
https://piazza.com/ed.ac.uk/spring2020/infr08028
Office hours: Wednesdays at 14:00-15:00 in IF-3.04

Jan-Mar 2020

Today's topics

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

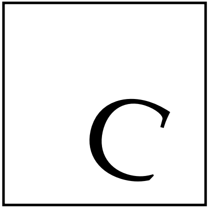
Types of learning problems

Data	System		Type of problem	Type of learning
	input	output		
$\mathbf{x}$	$\{\mathbf{x}\}$	groups (subsets)	clustering	unsupervised learning
$(\mathbf{x}, y)$	$\mathbf{x}$	y: discrete category	classification	supervised learning
$(\mathbf{x}, y)$	$\mathbf{x}$	y: continuous value	regression	supervised learning

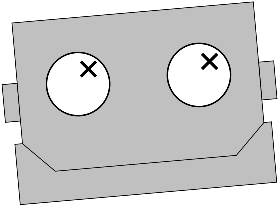
where $\mathbf{x} = (x_1, \ldots, x_D)^T$: feature vector
y : target vector or scalar

Test mode

Classification



N/A



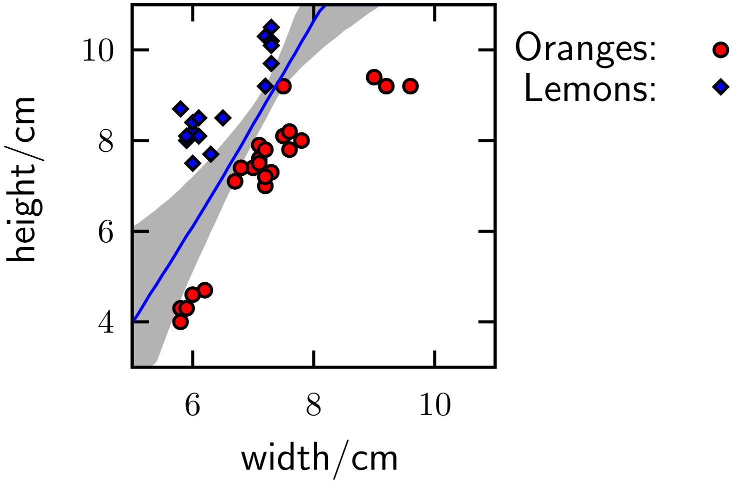
G ?

New data

Label

Goal of training: develop a classifier of good generalisation

Supervised learning



Classification

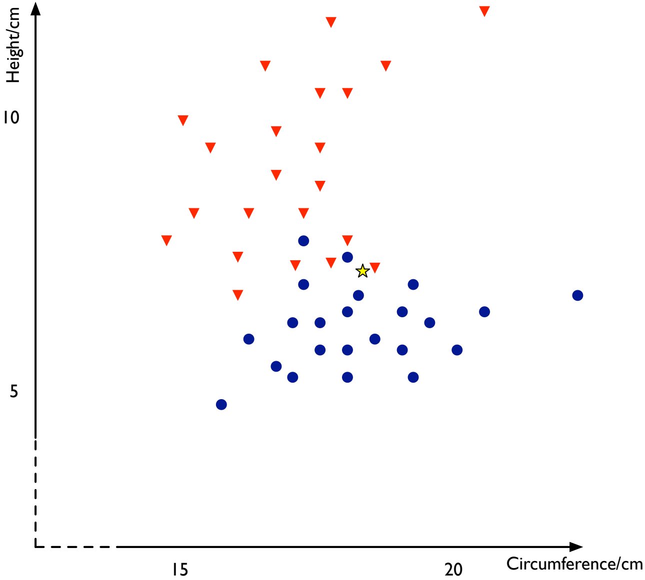
- The data has a feature vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_D)^T$ and a label $c \in \{1, \ldots, C\}$
- Training set: A set of N feature vectors and their labels $(\mathbf{x}_1, c_1), \ldots, (\mathbf{x}_N, c_N)$
- Use a learning algorithm to train a classifier from a training set
- Test 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:

$$\text{Error rate} = \frac{\#\text{ of misclassified samples}}{\#\text{ of test samples}}$$

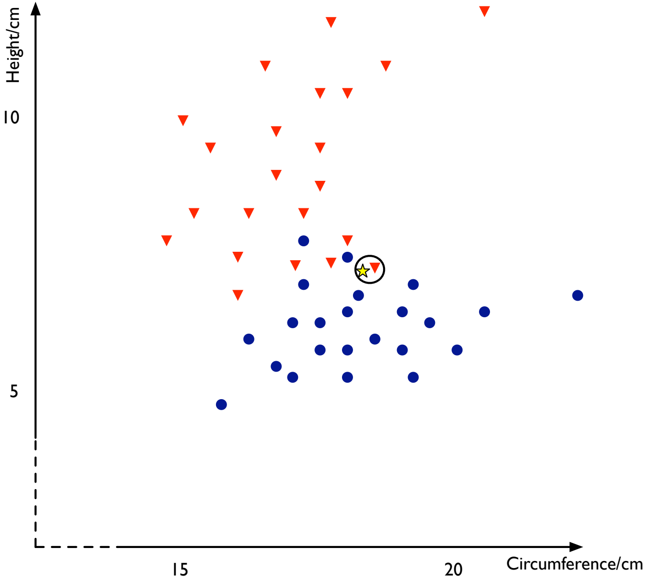
Nearest-neighbour classifier

- 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)

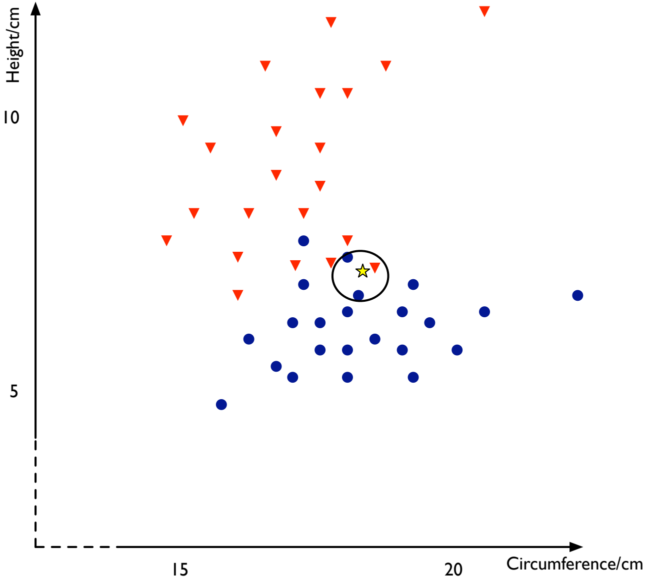
Classifying test data with K-nearest neighbours



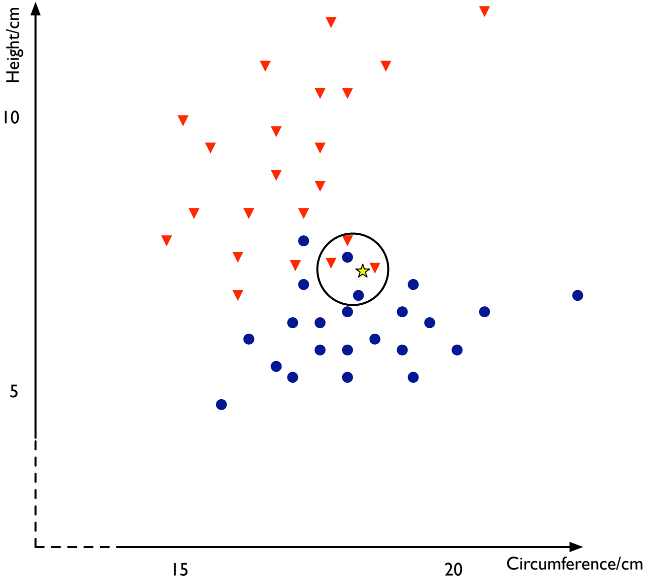
1-nearest neighbour



3-nearest neighbour



5-nearest neighbour



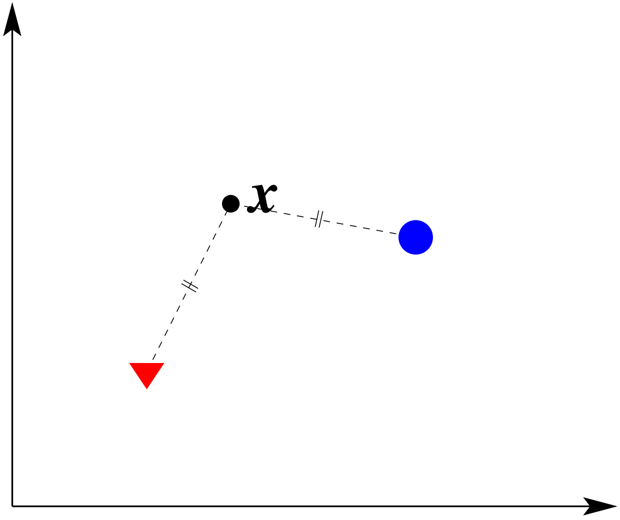
K-NN classification algorithm

For each test example $\mathbf{z} \in Z$:

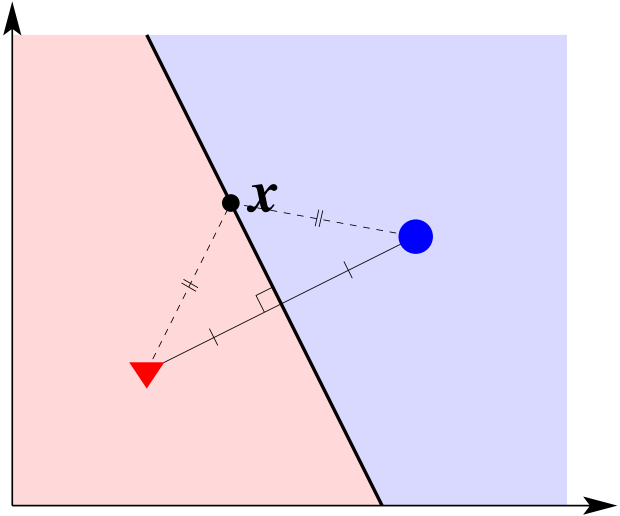
- Compute the distance $r(\mathbf{z}, \mathbf{x})$ between $\mathbf{z}$ and each training example $(\mathbf{x}, c) \in X$
- Select $U_k(\mathbf{z}) \subseteq X$, the set of the k nearest training examples to $\mathbf{z}$
- Decide the class of $\mathbf{z}$ by the majority voting:

$$c(\mathbf{z}) = \underset{j \in \{1,\ldots,C\}}{\arg\max} \sum_{(\mathbf{x},c) \in U_k(\mathbf{z})} \delta_{j\,c}$$

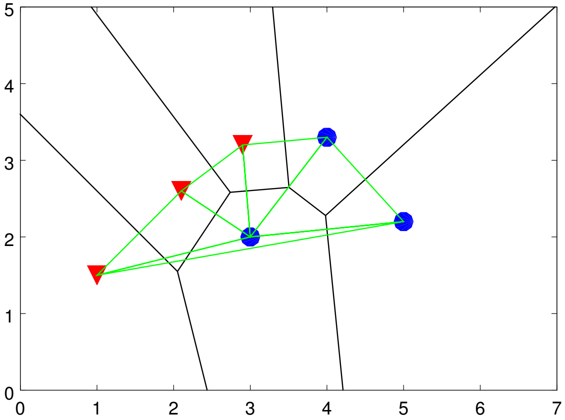
Geometry of nearest neighbour



Geometry of nearest neighbour – decision boundary and decision regions

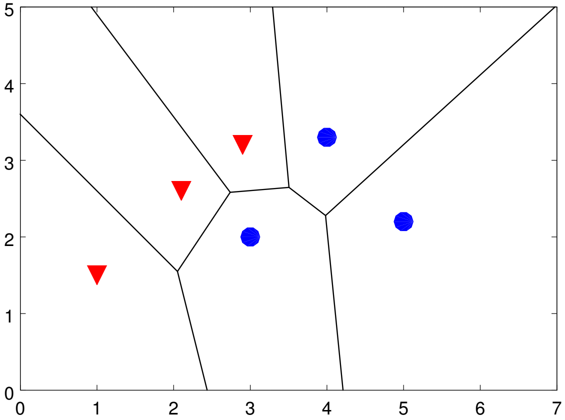


Geometry of nearest neighbour



Delaunay triangulation

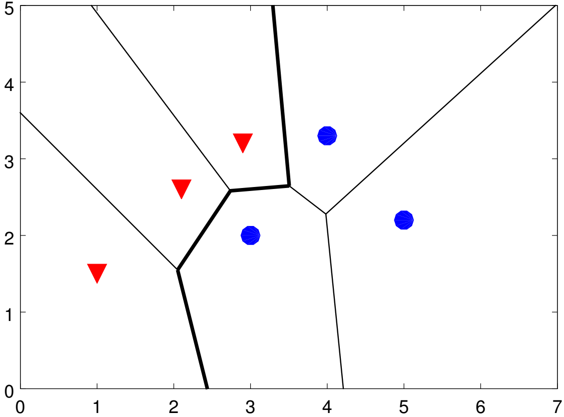
Voronoi tessellation



Voronoi diagram

Decision boundary

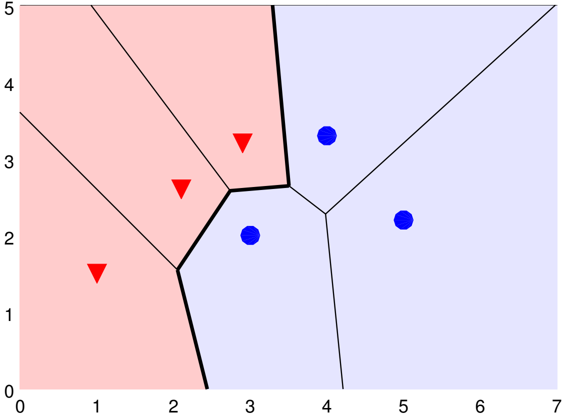
Decision boundary: boundary (surface) that partitions the vector space into subsets of different classes.



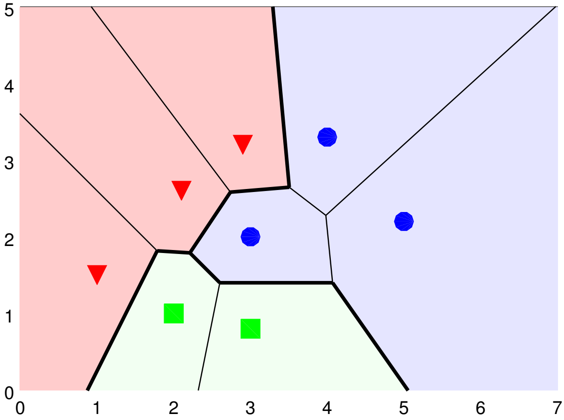
K-NN classification forms *piecewise-linear decision boundary*.

Decision regions

Decision regions: regions separated by the decision boundaries

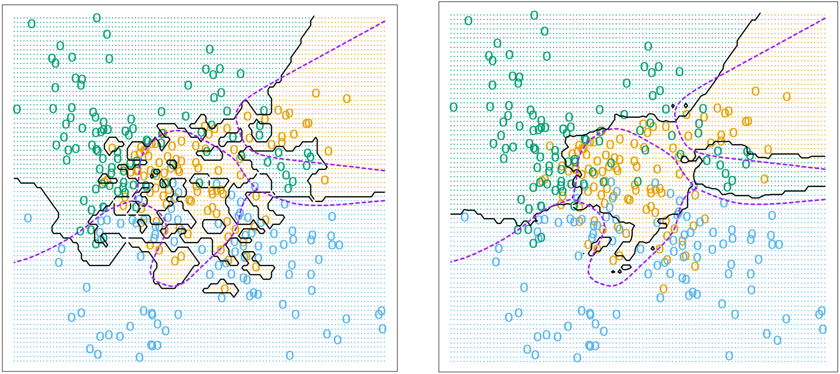


Decision boundaries for $C = 3$



What K should we use?

An example where a large K reduces noise

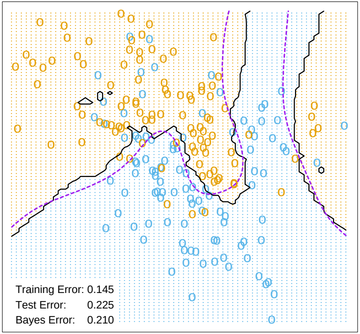


$K = 1$ $K = 15$

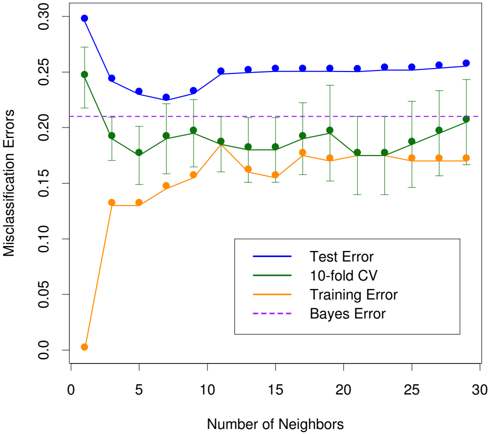
(Black curve: KNN decision boundary, broken purple curve: the Bayes decision boundary

The Elements of Statistical Learning (2nd Ed.)
Hastie, Tibshirani, Friedman. §13.3 p463–
http://statweb.stanford.edu/~tibs/ElemStatLearn/

Learning curves

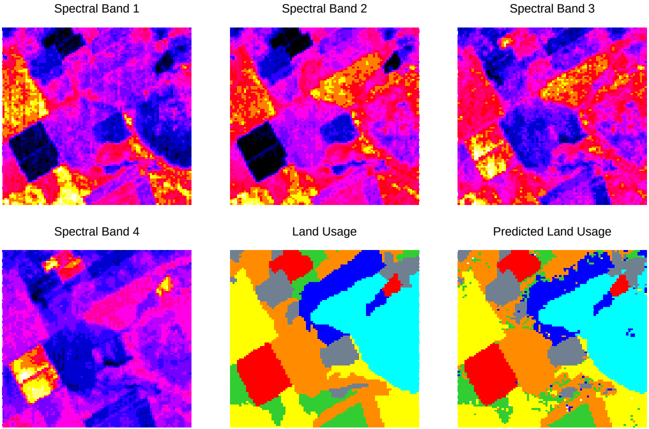


$K = 7$



The Elements of Statistical Learning (2nd Ed.)
Hastie, Tibshirani, Friedman. §13.3 p463–

LANDSAT Application

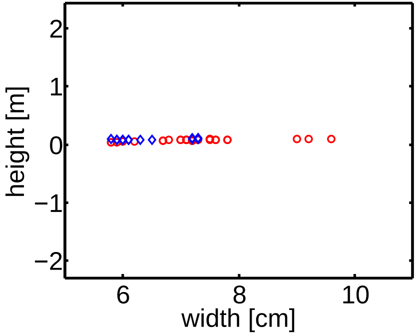


Predict land-usage from satellite data

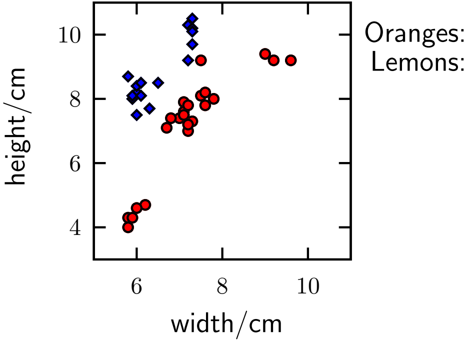
KNN applied to 9 pixel patch in 4 spectral bands, with $K = 5$

The Elements of Statistical Learning (2nd Ed.)
Hastie, Tibshirani, Friedman. §13.3 p463–

Tips on pre-processing data



different units



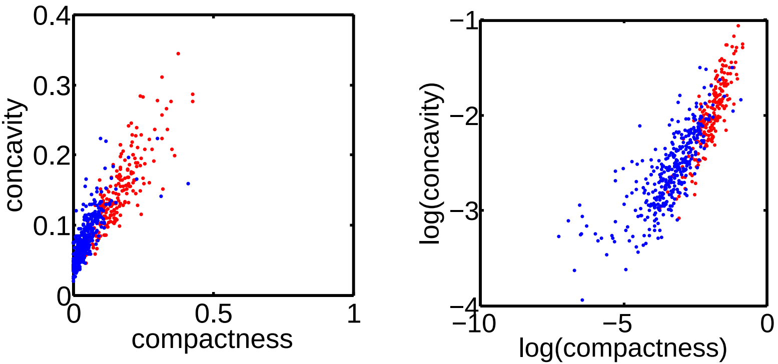
same unit

$\Rightarrow$ Standardise features unless understand units

Tips on pre-processing data

Wisconsin Diagnostic Breast Cancer (WDBC) data set

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

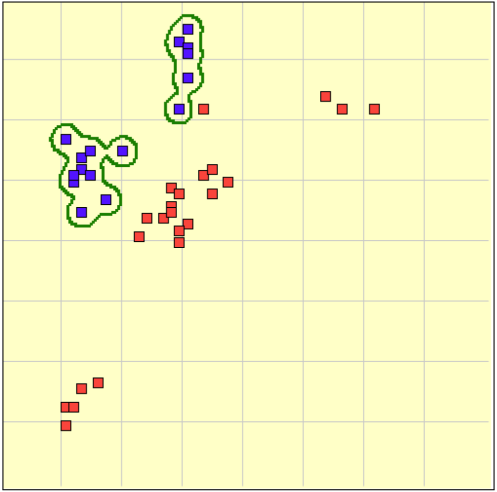


Linear scale   Log scale

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

Generalisation and over-fitting

How reasonable is this decision boundary?



Poor generalisation: stories

In a competition:

http://blog.kaggle.com/2012/07/06/the-dangers-of-overfitting-psychopathy-post-mortem/

Classic stories:

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

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

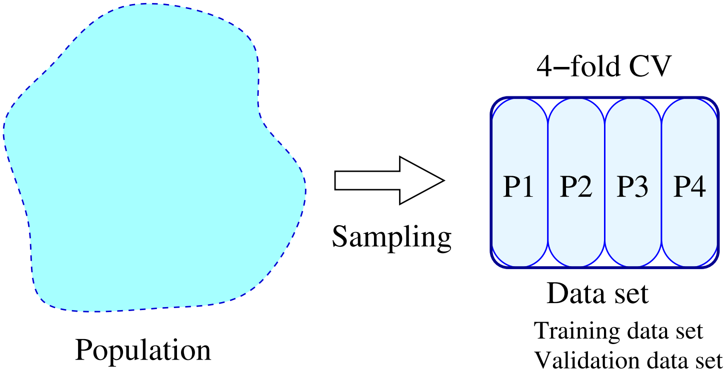
How reliable is the error rate?

- Error rate on training data set:
  can be $\sim$ 0 %
  $\Rightarrow$ 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)

Cross validation



Training data set
Validation data set
Test data set

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 closet 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)

Further reading *(NE)*

- L. Jiang, Z. Cai, D. Wang, S. Jiang, "Survey of Improving K-Nearest-Neighbor for Classification," Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)
- M.R. Abbasifard, B. Ghahremani, H. Naderi, "A Survey on Nearest Neighbor Search Methods," International Journal of Computer Applications (0975 – 8887), Vol.95, No.25, June 2014.
- Hand-Drawn Voronoi Diagrams
- Roberto Tamassia, "Introduction to Voronoi Diagrams," Lecture notes of C.S. 252, Computational Geometry, University of Brown, 1993.
- Steven Fortune, "A sweepline algorithm for Voronoi diagrams," Algorithmica 2, 153 (1987).

Labs

04th, 05th Feb.	Lab-3	K-means clustering and PCA
11th, 12th Feb.	Lab-4	K-NN classification