Introduction to Data Analytics

Xin Gao
Xin.gao@kaust.edu.sa
July 28, 2022
SDU

Machine Learning Goals

- We want to develop an algorithm that is able to improve the performance P at some task T with experience E

- Learning is defined by <P, T, E>

Machine Learning Goals

- Machine learning seeks to develop theories and computer systems for
  - Representing;
  - Classifying, clustering and recognizing;
  - Reasoning under uncertainty;
  - Predicting;
  - Reacting to

  Complex and real-world information, based on the system's own experience with data, and (hopefully) under an explicit model or mathematical framework, that
    - Can be formally characterized and analyzed
    - Can take into account human prior knowledge
    - Can generalize and adapt across data and domains
    - Can operate automatically and autonomously
    - Can be interpreted and perceived by human

Growth of Machine Learning

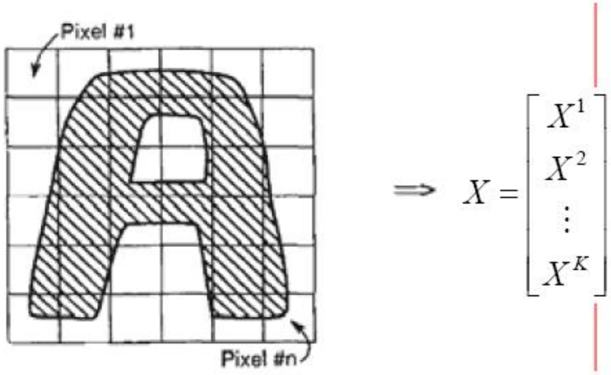
- Machine learning has been the dominant approach to
  - Robot control
  - Speech recognition
  - Computer vision
  - Etc
- It is still growing quickly
  - Increased data capture ability
  - Improved learning algorithms
  - Increased massive data

Paradigms of Machine Learning

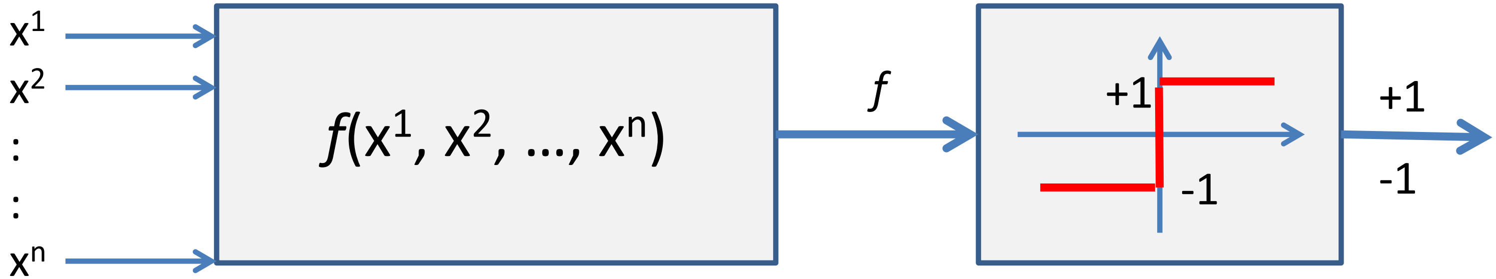
- Supervised learning
  - Given $D=\{X_i, Y_i\}$, learn F, s.t., $Y_i=F(X_i)$
- Unsupervised learning
  - Given $D=\{X_i\}$, learn F, s.t., $Y_i=F(X_i)$
- Semi-supervised learning
  - Given $D=\{X_i, Y_i, X_j\}$, where $0<=i<=k$, $k+1<=j<=N$, learn F, s.t., $Y_l=F(X_l)$, where $0<=l<=N$
- Reinforcement learning
  - Given D={env, actions, rewards}, learn policy: e, r -> a
    utility: a, e -> r
- Active learning
  - Given $D=\{X_i, Y_i\}$, iteratively learn $F|(D \text{ and } D^{new})$, where $F \rightarrow D^{new}$

Classification

- A problem in statistics of identifying the sub-population to which new observations belong
- Representing data

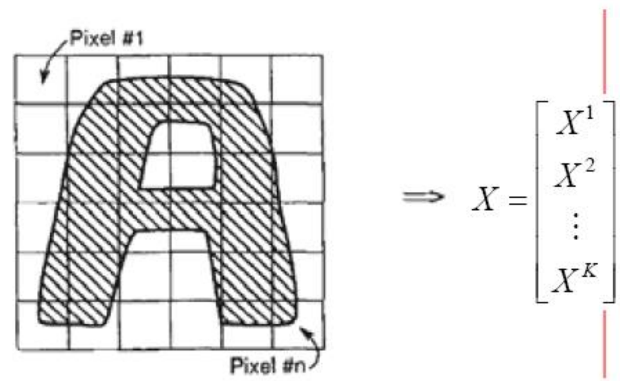


- Hypothesis (classifier)

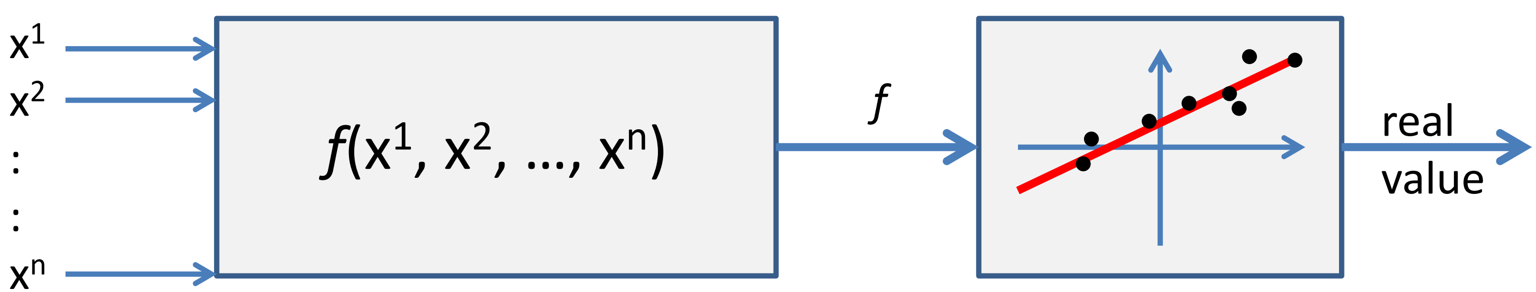


Regression

- A problem of modeling and analyzing the relationship between a dependent variable and one or more independent variables
- Representing data



- Model



K Nearest Neighbors

- KNN is a simple algorithm that stores all available instances and classifies new instances based on a distance metric to the available ones

- KNN is also called
  - Case-based learning
  - Memory-based learning
  - Lazy learning
  - Instance-based learning

KNN Algorithm

- Training process:
  - Store the available training instances
- Predicting process:
  - Find the K training instances that are closest to the query instance
  - For classification, return the most frequent class label among those K instances
  - For regression, return the average of those K instances

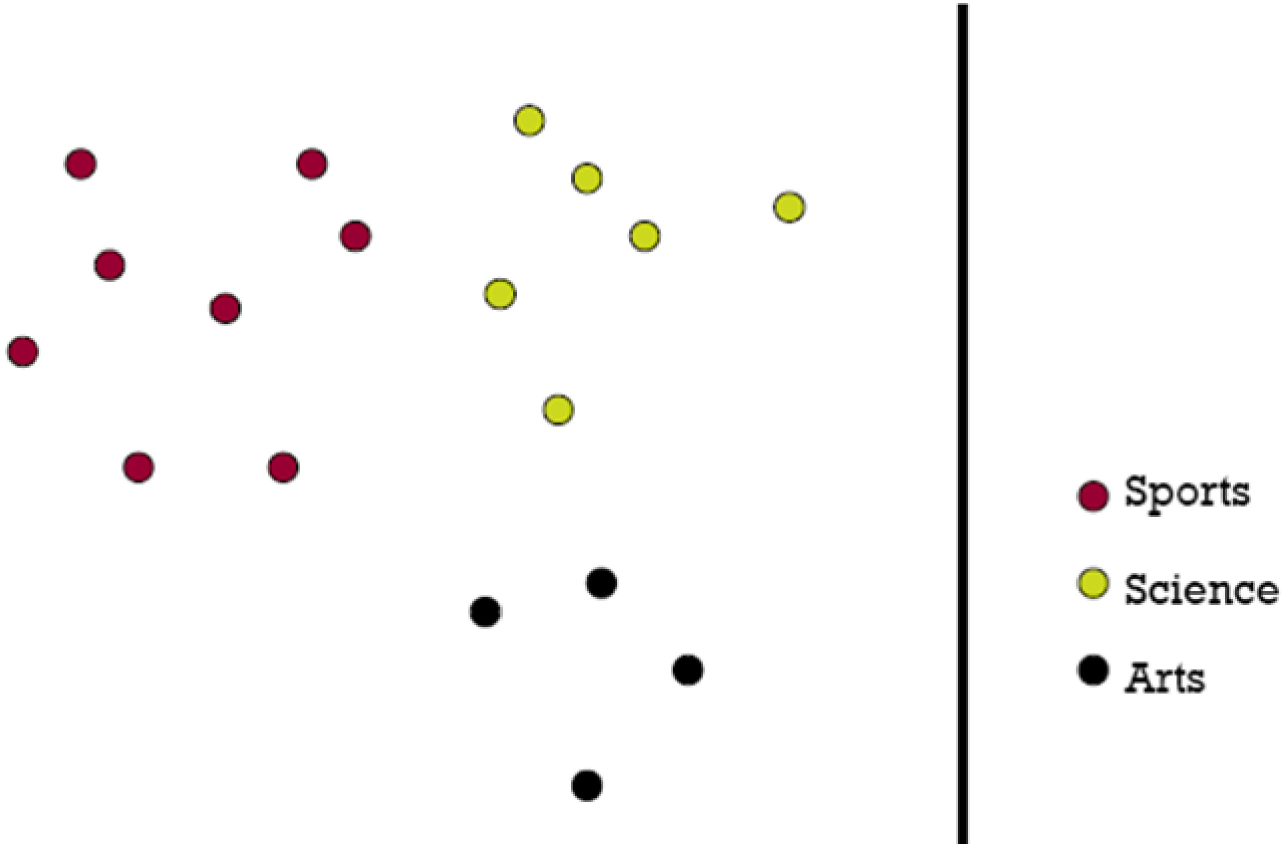
Example

- Text classification: classify documents into different classes/categories/labels
- Representation:

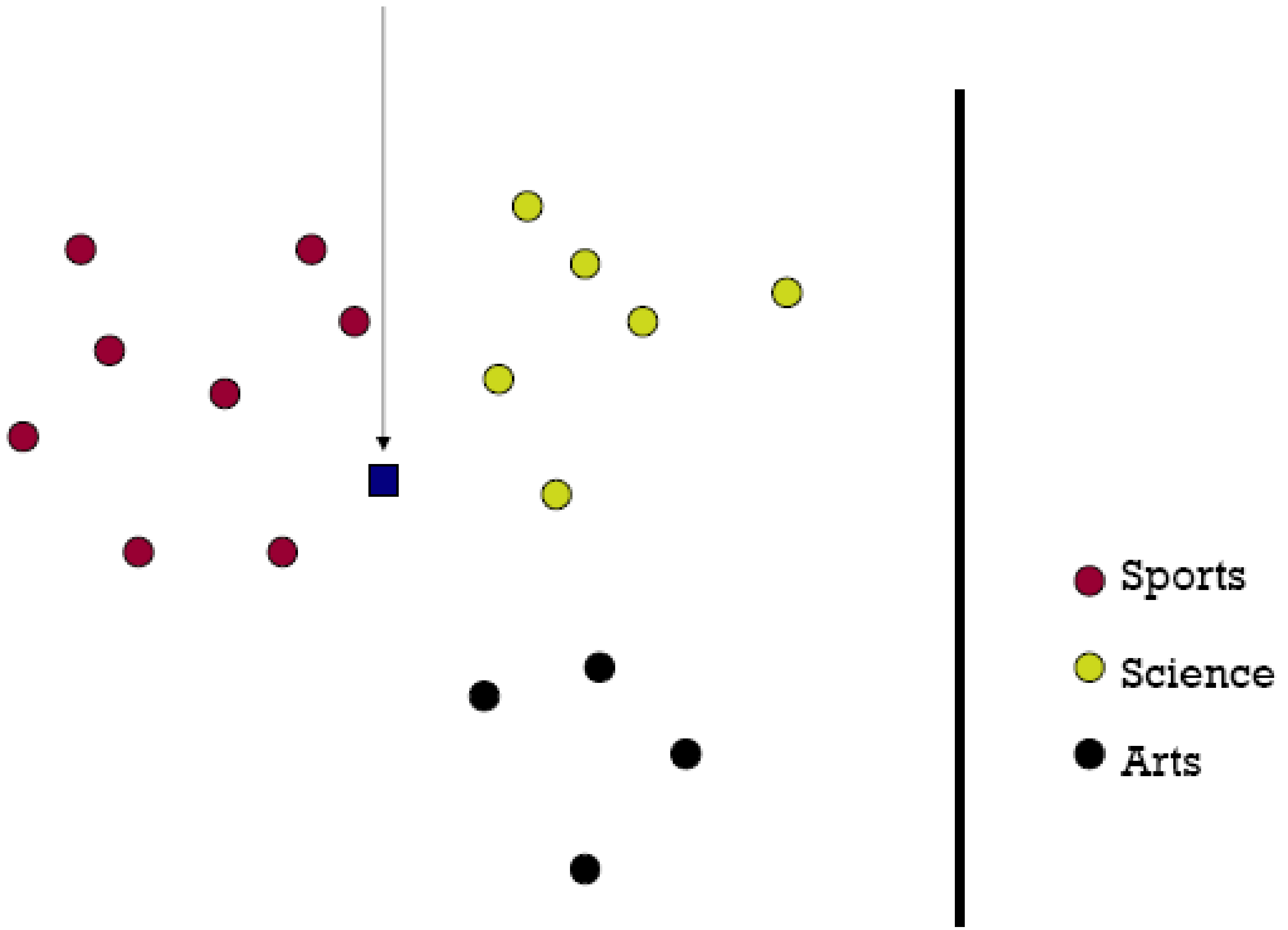
	Doc 1	Doc 2	…	Doc N
Word 1	5	3		0
Word 2	0	0		7
…				
Word M	3	7		4

  – Documents are represented by vectors
  – Need to be normalized

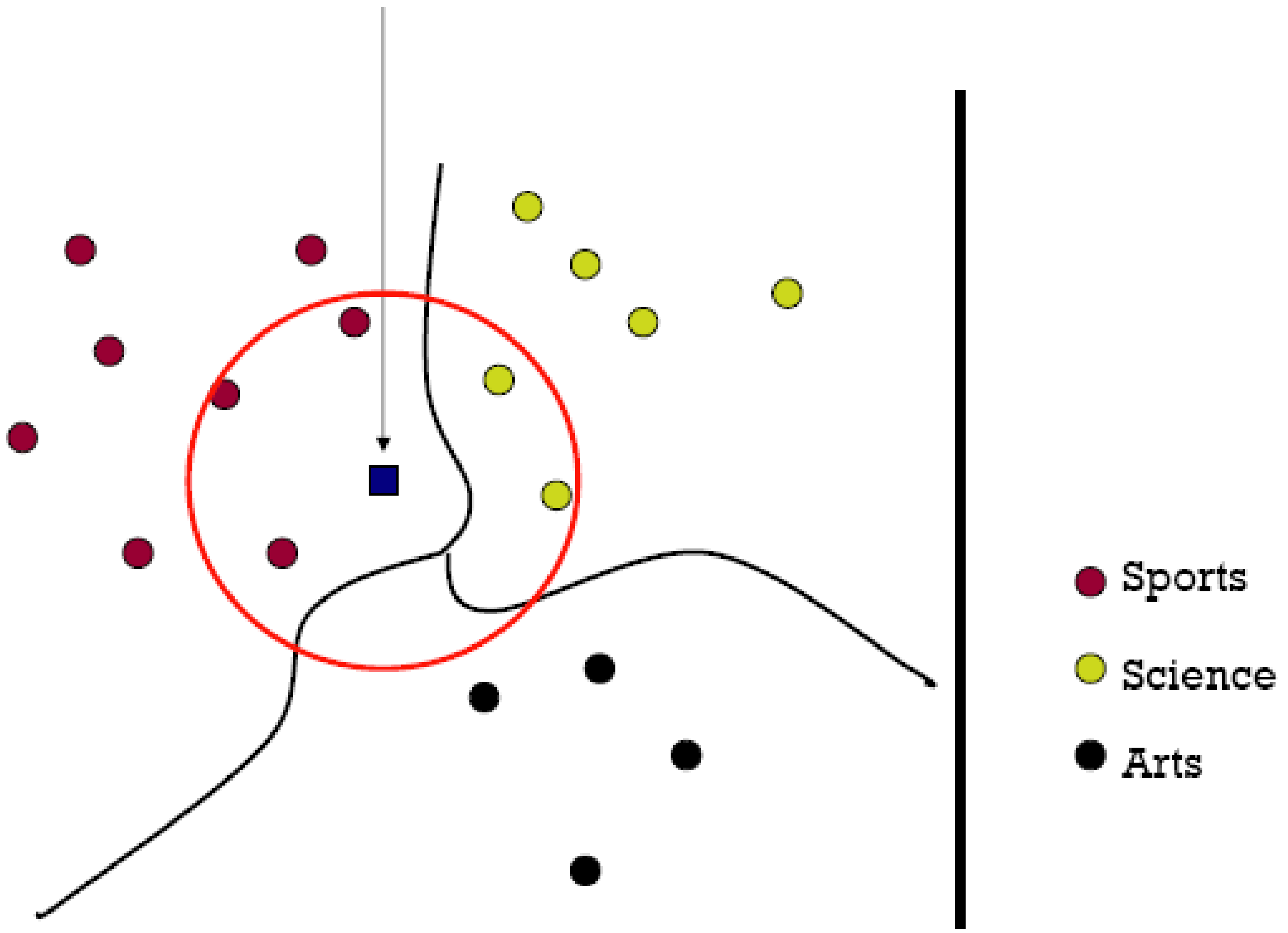
Example



Example



Example



Performance Measure

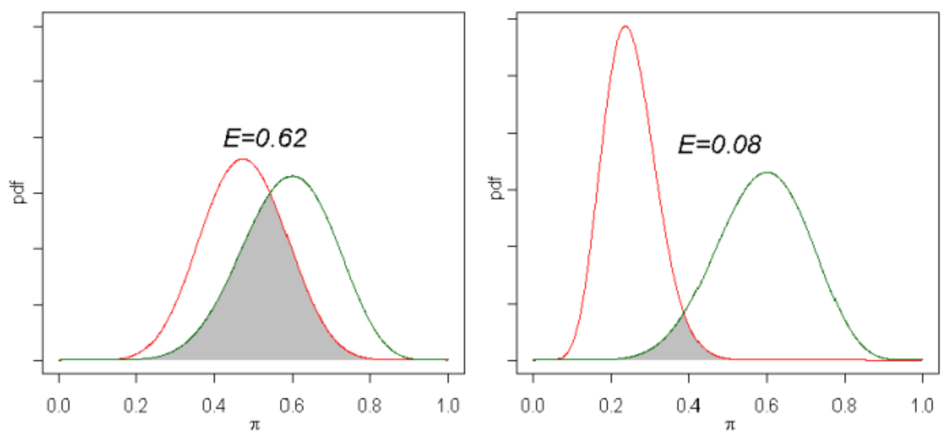
- Time and space complexity is both O(K)
- KNN is conceptually simple, yet able to solve complex problems
- Can work with relatively little information
- Learning is extremely simple (no learning at all!)
- Does not explicitly compute a generalization or category prototypes
- KNN is close to optimal!

Bayes Error

- We want to calculate the probability of error for a binary classifier
  - The probability that a sample is assigned to the wrong class
- Given an instance x, the risk is

$$r(x) = \min[p_1(x), p_2(x)]$$

- Bayes error is the lower bound of probability of classification error



KNN Is Close to Optimal

- Cover and Hart 1967
- Asymptotically, the error rate of 1-nearest-neighbor classification is less than twice the Bayes error (error rate of classifier knowing model that generated data)
- Asymptotic error rate is 0 if Bayes rate is 0

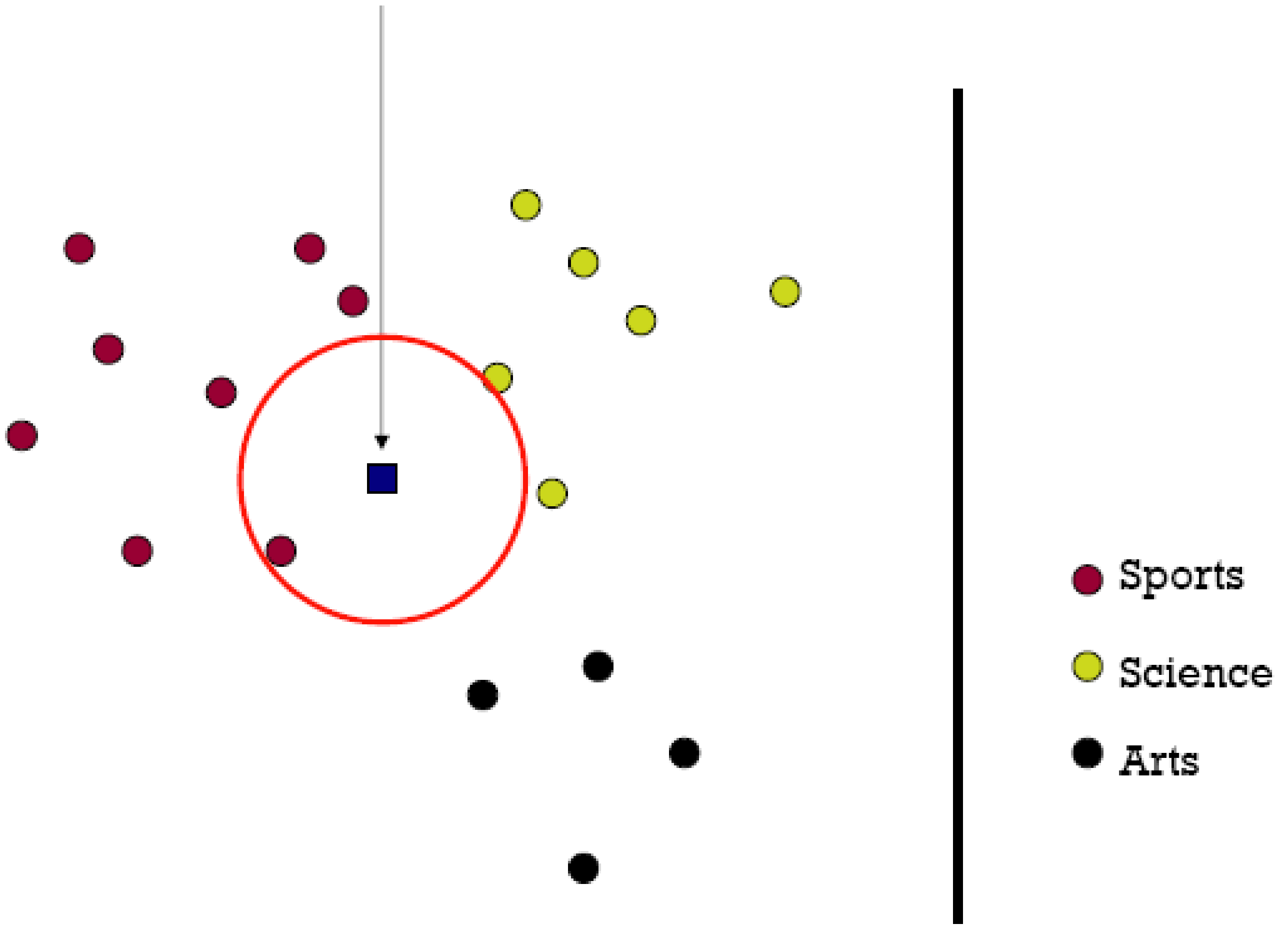
KNN Is an Instance-based Learning

- What determines an instance-based learning?
  - A distance metric
  - How many neighbors to look at
  - A weighing function (optional)

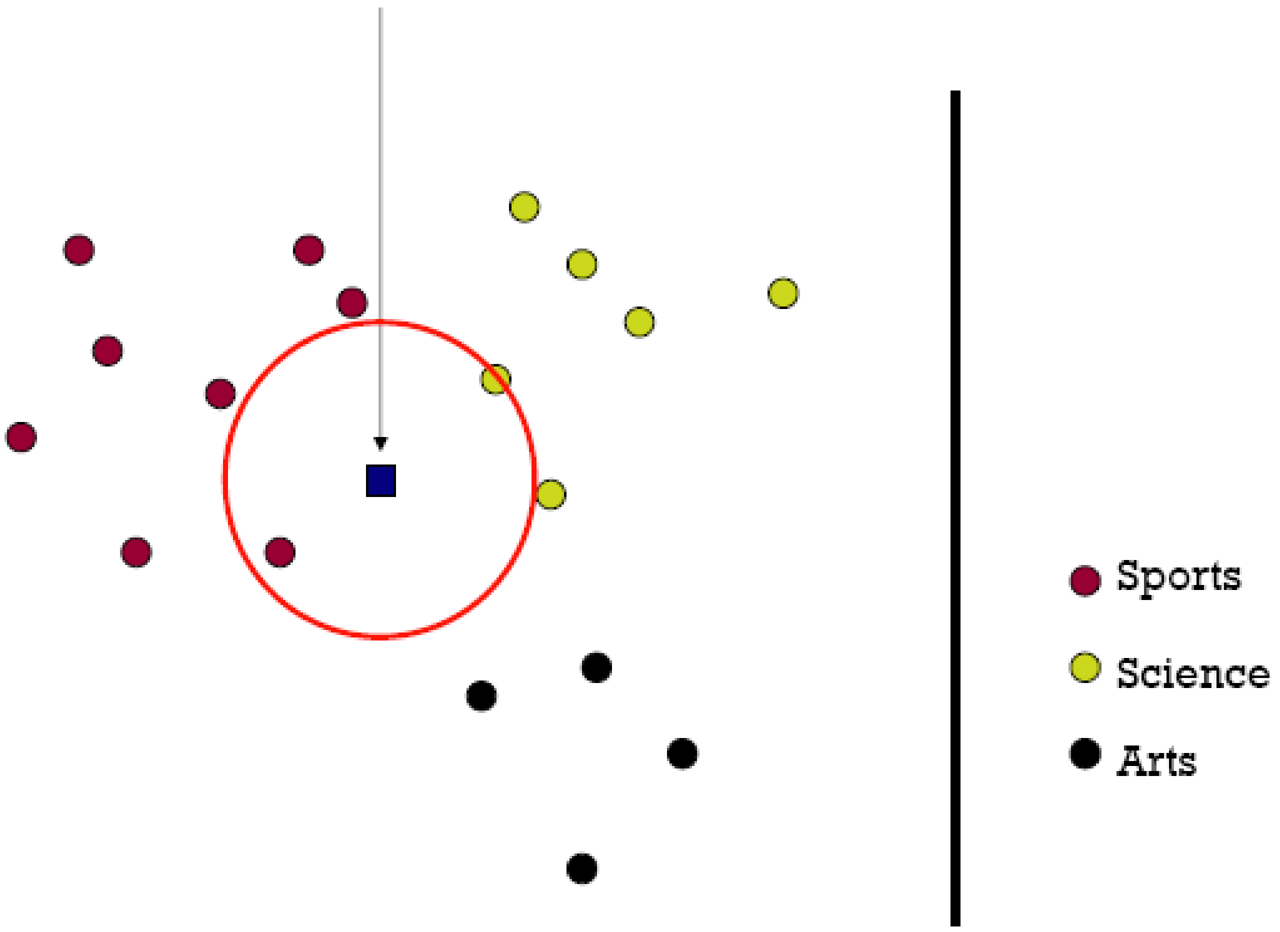
Distance Metrics

- Euclidean distance
  - $D(x, x') = \sqrt{\sum_i (x_i - x'_i)^2}$
- L_1 norm
  - $D(x, x') = \sum_i |x_i - x'_i|$
- L_∞ norm
  - $D(x, x') = \max|xi - x'_i|$
- Mahalanobis distance
  - $D(x, x') = \sqrt{(x - x')^T S^{-1} (x - x')}$
- Correlation
- Hamming distance
- Manhattan distance
- etc

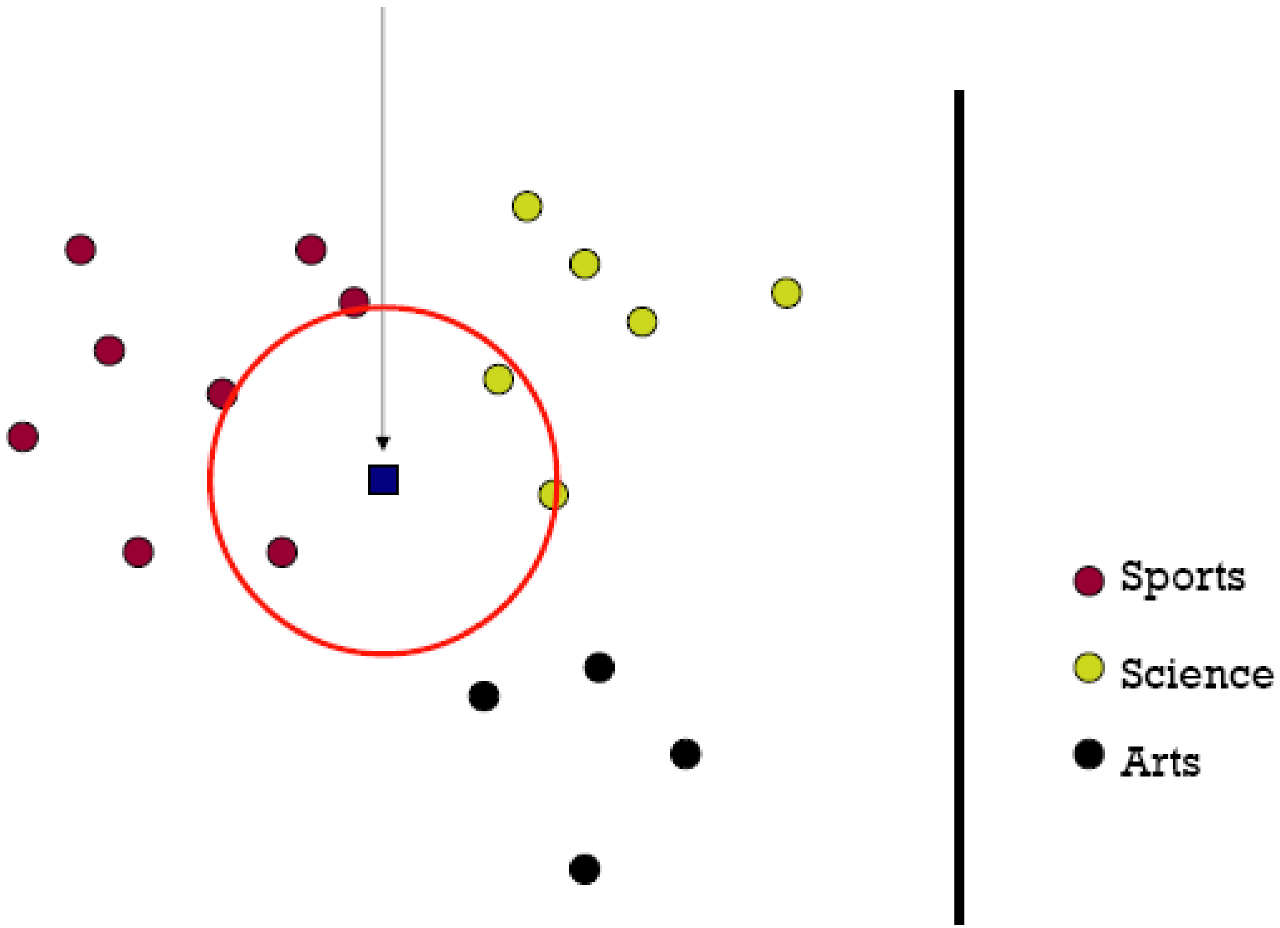
Number of Neighbors



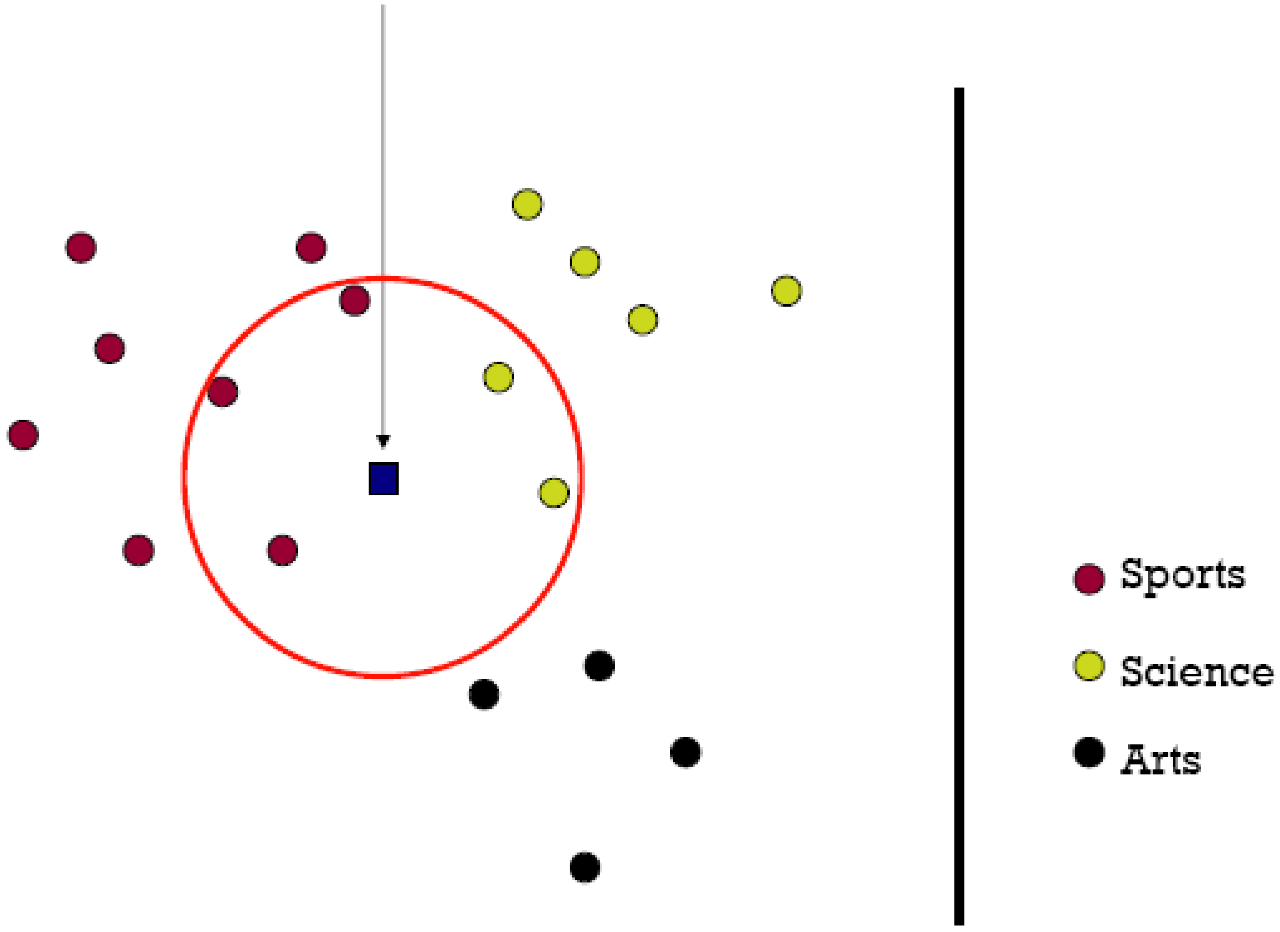
Number of Neighbors



Number of Neighbors



Number of Neighbors



Number of Neighbors

- In practice, using a value of K somewhere between 5 and 10 gives good results for most low-dimensional data sets
- A good K can also be chosen by using cross-validation

Cross-validation

- **Cross-validation/rotation estimation**, is a technique for assessing how the results of a statistical analysis will generalize to an independent data set
- Mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice
- One round of cross-validation involves partitioning a set of data into complementary subsets, performing the analysis on one subset (called the *training set*), and validating the analysis on the other subset (called the *testing set*)
- Cross-validation is a powerful way to deal with overfitting

k-fold Cross-validation

- **Idea:** train multiple times, leaving out a disjoint subset of data each time for validation. Average the validation set accuracies
- **Process:**
  - Randomly partition data into K disjoint subsets
  - For k = 1 to K
    - ValidationData = k-th subset
    - h <- classifier trained on all data except for ValidationData
    - Accuracy(k) = accuracy of h on ValidationData
  - End
- FinalAccuracy = mean of the K recorded accuracies

Leave-one-out Cross-validation

- **Idea:** a special case of k-fold cross-validation, where k = K
- **Process:**
  - Partition data into K disjoint subsets, each containing on data point
  - For k = 1 to K
    - ValidationData = kth subset
    - h <- classifier trained on all data except for ValidationData
    - Accuracy(k) = accuracy of h on ValidationData
  - End
- FinalAccuracy = mean of the K recorded accuracies