



Classification (I)

Never Stand Still

COMP9417 Machine Learning & Data Mining

Aims

This lecture will introduce you to machine learning approaches to the problem of classification. Following it you should be able to reproduce theoretical results, outline algorithmic techniques and describe practical applications for the topics:

- a framework for solving classification problems
- the use of a linear model as a 2-class classifier
- logistic regression classification algorithm
- describing issues of generalisation and evaluation for classification
- describing distance measures and how using them in classification
- the basic k-nearest neighbour classification method

Introduction

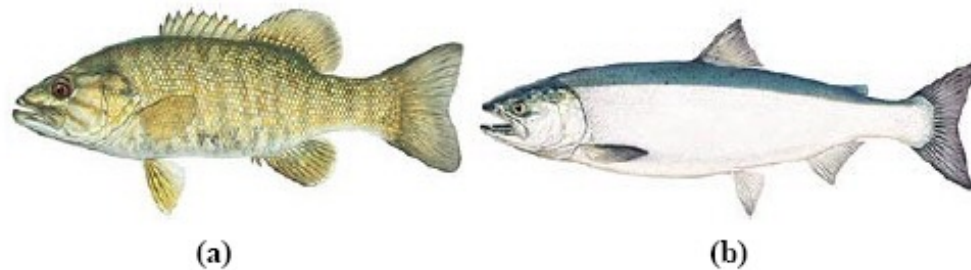
Classification (sometimes called *concept learning*) methods dominate machine learning . . .

. . . however, they **often don't have convenient mathematical properties** like regression, so are more complicated to analyse. The idea is to learn a *classifier*, which is usually **a function mapping from an input data point to one of a set of discrete outputs**, i.e., the classes.

We will mostly focus on classifier advantages and disadvantages as learning methods first and point to unifying ideas and approaches where applicable.

Classification

Example: Imagine that we want to automate the process of sorting incoming fish in a fish-packing plant. And as a pilot, we start by separating sea bass from salmon using some information collected through sensing.

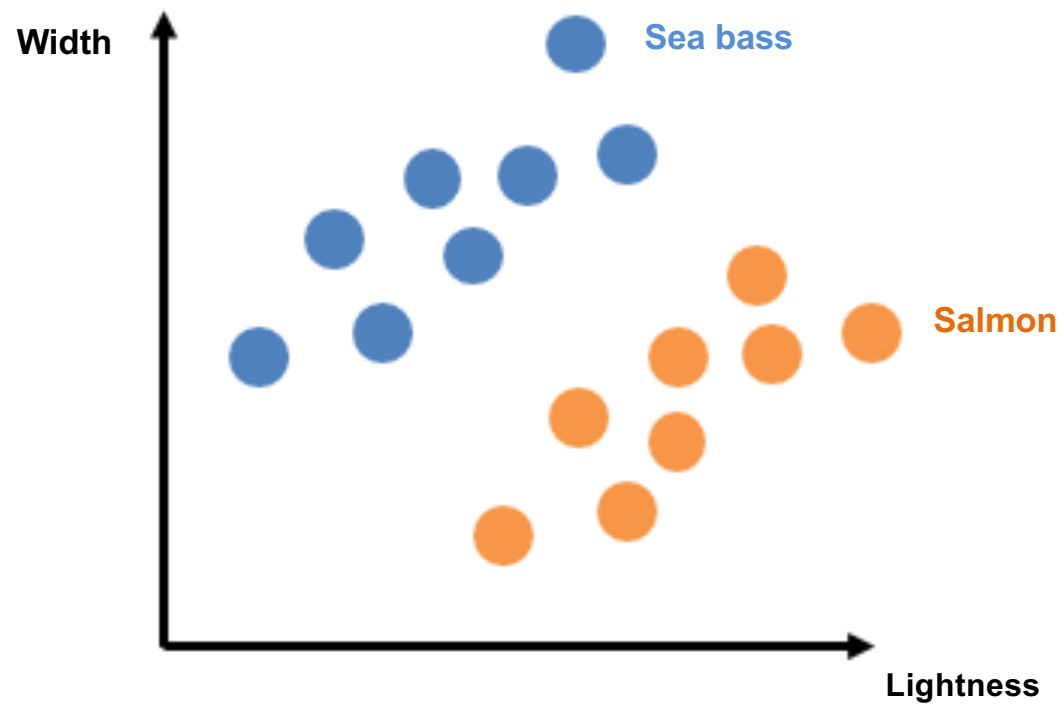


Classification

Example: classifying sea bass vs. salmon

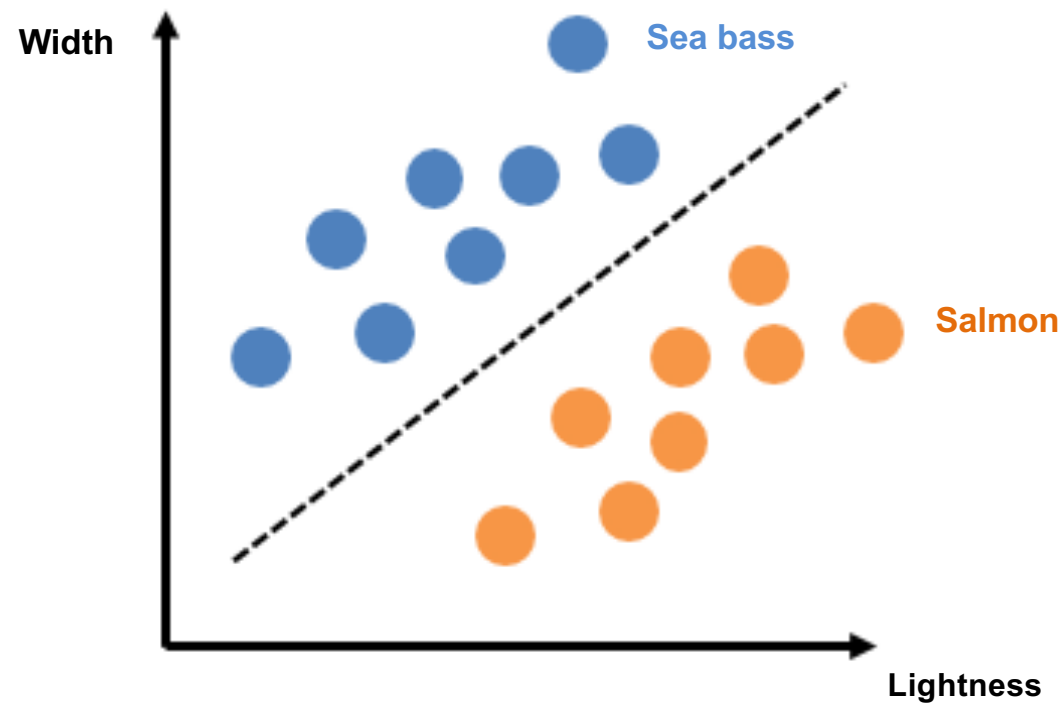
Features that can be used: width, length, weight, lightness, fins, eyes/mouth position, etc.

Question: how to separate these two classes?



Classification

Example: Maybe we can find a line that separates the two classes.



Classification

Example: If we find the line that separated the two classes, then how our algorithm makes prediction?

The line equation will look like:

$$ax_1 + bx_2 + c = 0$$

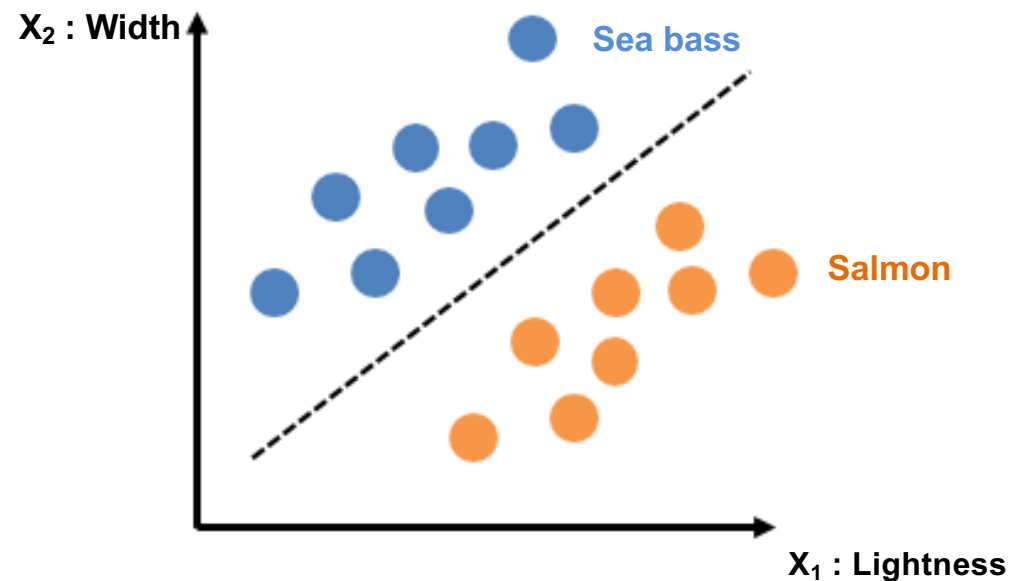
We can define a , b & c such that:

for any point above the line:

$$ax_1 + bx_2 + c > 0$$

and for any point below the line:

$$ax_1 + bx_2 + c < 0$$

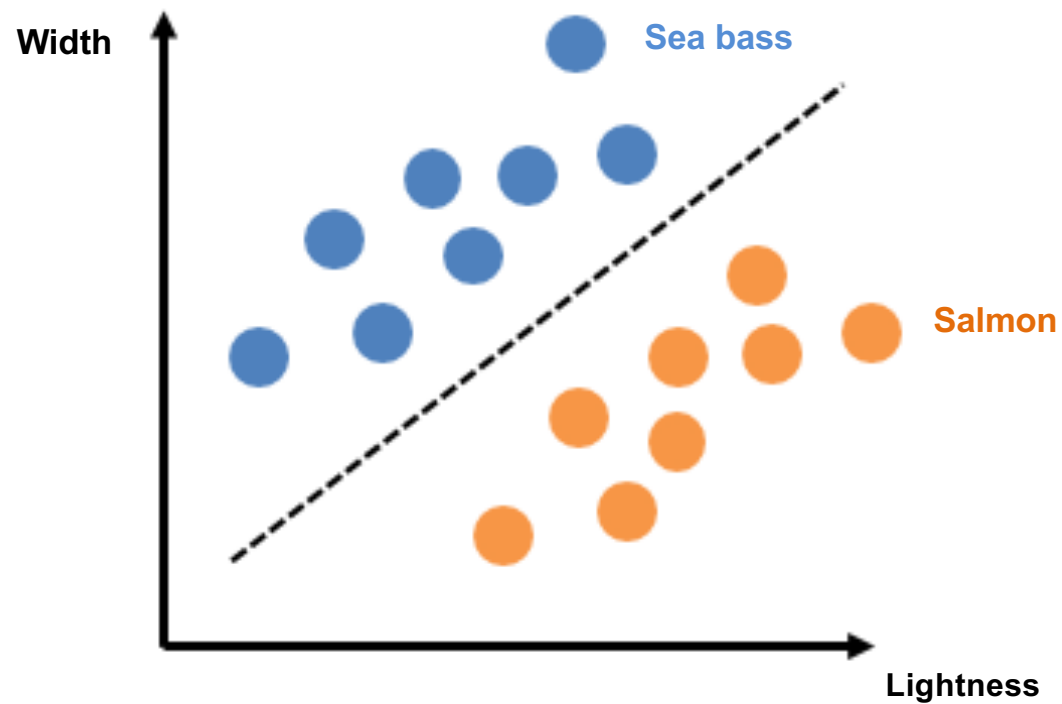


This type of classifier is called *linear classifier*. It is also a type of *discriminative learning* algorithm.

Classification

Example:

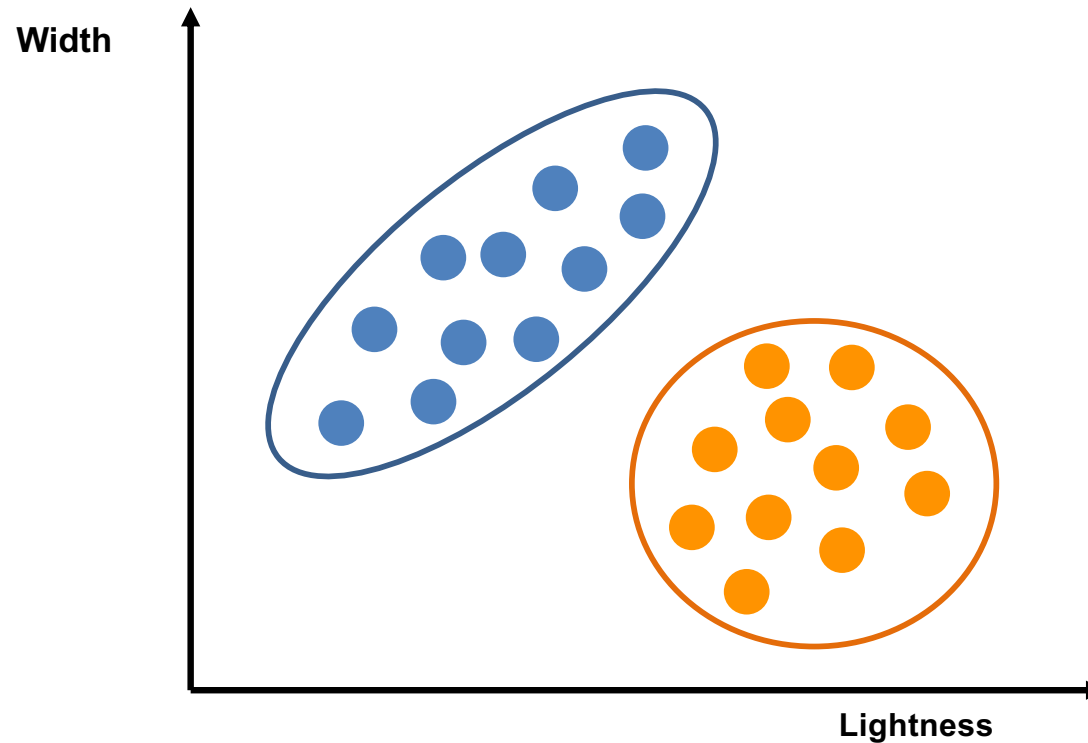
Can we do something different than finding the discriminative line (or some boundary) to be able to separate the two groups?



Classification

Example:

Instead of finding a discriminative line, maybe we can focus on one class at a time and build a model that describes how that class looks like; and then do the same for the other class. This type of models are called *generative learning algorithm*.



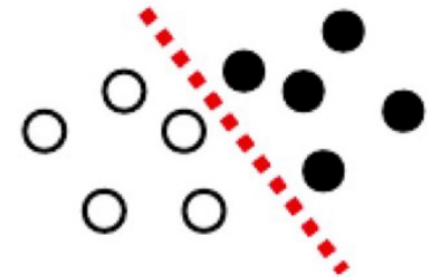
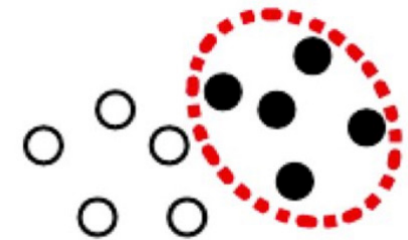
Classification

Generative algorithm: builds some models for each of the classes and then makes classification predictions based on looking at the test example and see it is more similar to which of the models.

- Learns $p(x|y)$ (and also $p(y)$, called class prior)
- So, we can get $p(x, y) = p(x|y)p(y)$
- It learns the mechanism by which the data has been generated

Discriminative algorithm: Do not build models for different classes, but rather focuses on finding a decision boundary that separates classes

- Learns $p(y|x)$



Classification

- In a generative algorithm, to predict the output for sample x , we have to estimate $p(y|x)$:

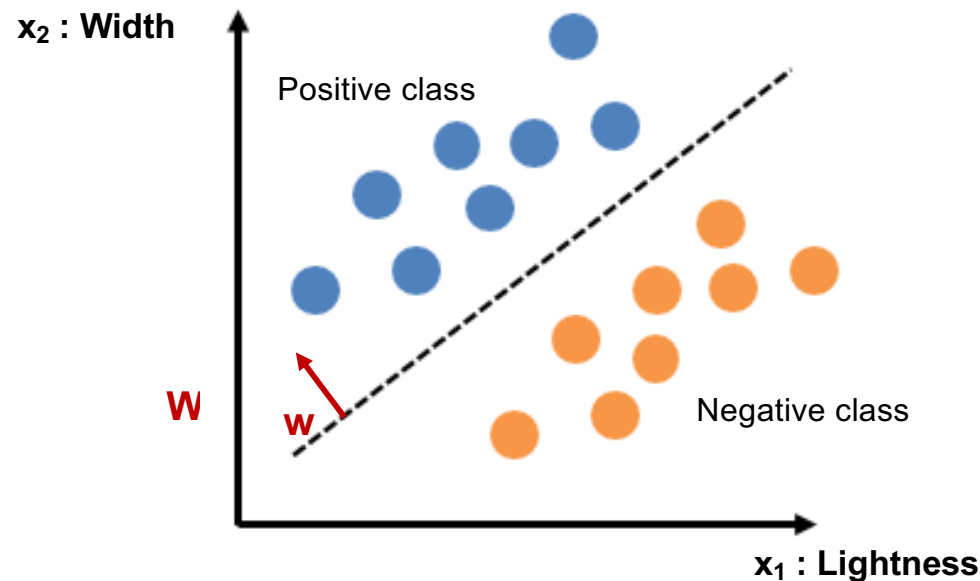
$$p(y = 0|x) = \frac{p(x|y = 0)p(y = 0)}{p(x)}$$

$$p(y = 1|x) = \frac{p(x|y = 1)p(y = 1)}{p(x)}$$

If $p(y = 0|x) > p(y = 1|x)$, then x belongs to class $y = 0$ and otherwise to class $y = 1$.

- In a discriminative algorithm, we can directly have $p(y = 0|x)$ and $p(y = 1|x)$ and similar to above, if $p(y = 0|x) > p(y = 1|x)$, then x belongs to class $y = 0$ and otherwise to class $y = 1$.

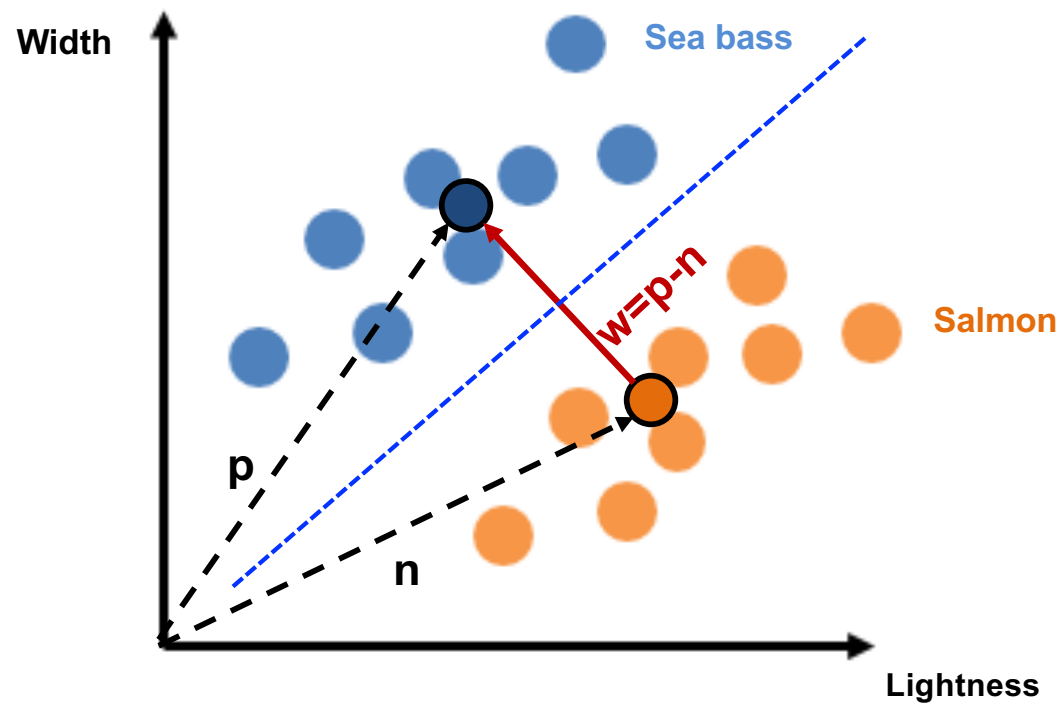
Linear classification in two dimensions



- We find the line that separates the two class: $ax_1 + bx_2 + c = 0$
- We define a weight vector $w^T = [a, b]$, $x^T = [x_1, x_2]$
- So, the line can be defined by $x^T w = -c = t$
- w is perpendicular to decision boundary (in direction of positive class)
- t is the decision threshold (if $x^T w > t$ then x belongs to positive class and if $x^T w < t$ then x belongs to negative class)

Basic Linear Classifier

The basic linear classifier constructs a decision boundary by half-way intersecting the line between the positive and negative centres of mass.



Basic Linear Classifier

The basic linear classifier is described by the equation $x^T w = t$, and $w = p - n$

As we know, $\frac{p+n}{2}$ is on the decision boundary, so we have:

$$t = \left(\frac{p+n}{2}\right)^T \cdot (p-n) = \frac{||p||^2 - ||n||^2}{2}$$

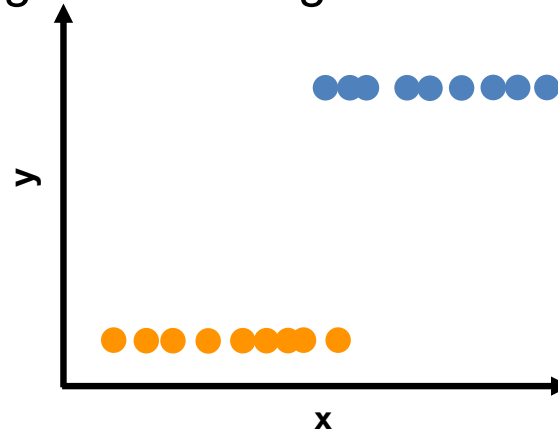
Where $||x||$, denotes the length of vector x

Logistic Regression

a probabilistic linear classifier

Logistic Regression

- In a binary classification problem with one input variable, if we show the output on y-axis we may get something like below:



- Can we use a linear regression to model this data?

Logistic Regression

Why regular regression doesn't work here:

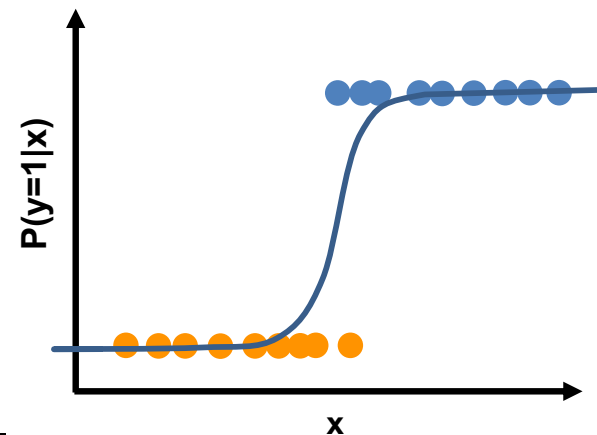
- Univariate or multivariate regressions are to predict a real-valued output from one or more independent variables
- Binary data doesn't have a normal distribution which is a condition needed for most other types of regression
- What we expect to get is an output of 0 or 1, but linear regression obviously can produce values beyond that range.

Logistic Regression

- In binary classification, we can transform the y values into probability values (values are in range $[0,1]$)

- We can model this with a s-curve (sigmoid curve) as above:

$$P(y = 1|x) = \frac{1}{1 + e^{-f(x)}} \Rightarrow f(x) = \log \frac{P(y = 1|x)}{1 - P(y = 1|x)}$$



Now $f(x)$ can have a value between $-\infty$ and $+\infty$ and in Logistic Regression we estimate $f(x)$ with a line.

$$\hat{f}(x) = x^T \beta \Rightarrow \log \frac{P(y = 1|x)}{1 - P(y = 1|x)} = x^T \beta$$

Logistic Regression

Logistic regression seeks to:

- Model the probability of a class given the values of independent input variables
- Estimate the probability that a class occurs for a random observation (versus the probability that the class doesn't occur)
- Classify an observation based on the probability estimations

Logistic Regression

$$\hat{P}(y = 1|x) = \frac{1}{1 + e^{-x^T \beta}}$$

If $P(y = 1|x) \geq 0.5$ (same as saying $x^T \beta \geq 0$) then predict as class 1

If $P(y = 1|x) < 0.5$ (same as saying $x^T \beta < 0$) then predict as class 0

- Interpretation of this model in the input space (feature space) is equivalent of having a linear decision boundary ($x^T \beta = 0$) separating the two class
- Now we have a linear solution to our problem, and this is what makes *Logistic Regression* a **linear model**.

Logistic Regression Parameter Estimation

- We can not use a similar cost function as we used in linear regression here, because it will result a non-convex function with many local minimums and would be very difficult to find the global minimum.
- Instead, the following cost function is used:

Let's define $\hat{P}(y = 1|x) = h_{\beta}(x)$

$$\text{cost}(h_{\beta}(x), y) = \begin{cases} -\log(h_{\beta}(x)) & \text{if } y = 1 \\ -\log(1 - h_{\beta}(x)) & \text{if } y = 0 \end{cases}$$

$$J(\beta) = -\frac{1}{m} \sum_{i=1}^m [y^{(i)} \log(h_{\beta}(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_{\beta}(x^{(i)}))]$$

Now, to find the values of parameters to minimize $J(\beta)$, we can use the Gradient Descent algorithm.

Logistic Regression

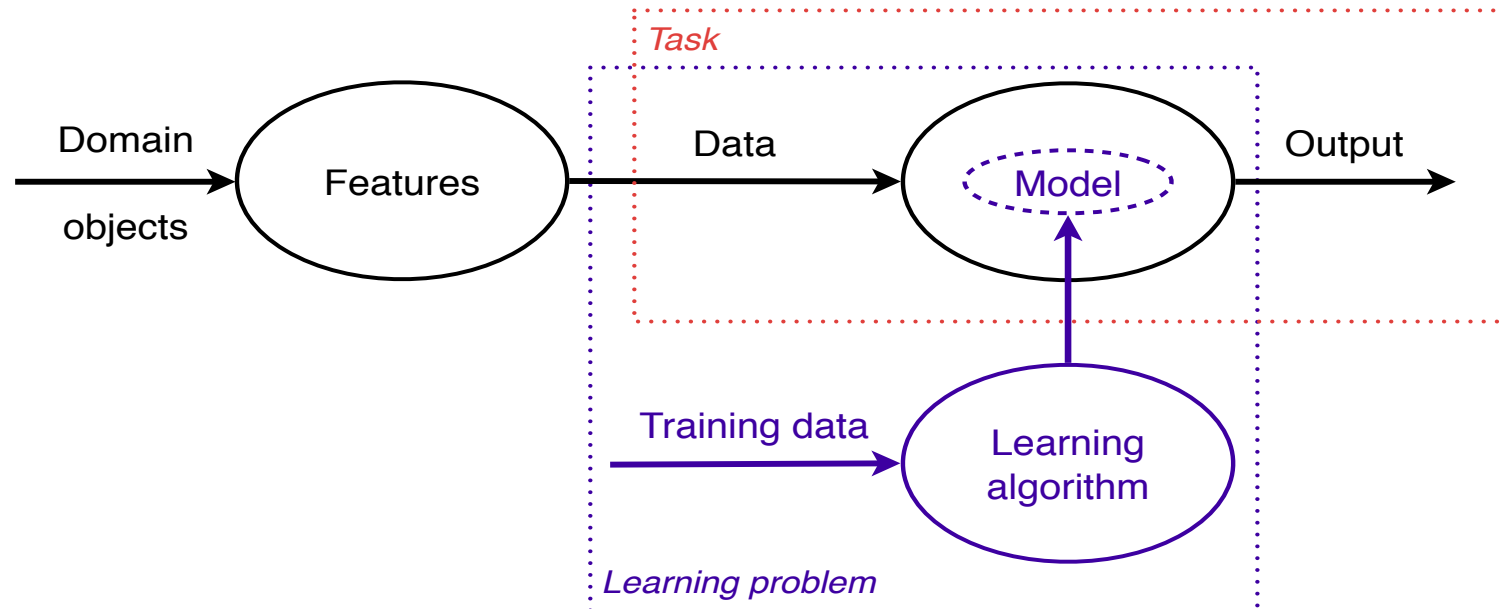
Pros:

- Relatively easy to implement
- Easy to interpret
- Relatively fast at training and very fast at testing
- Can easily extend to multi-classes
- Provide probabilistic predictions

Cons:

- Prone to overfitting in high-dimensional data (one remedy: regularization)
- It provides a linear decision boundary. For non-linear decision boundaries, feature transformation is required
- Requires moderate or no correlation (collinearity) between input variables and may lead to poor model (dimensionality reduction is useful)
- Sensitive to outlier

How solve a task with machine learning



An overview of how machine learning is used to address a given task. A task (red box) requires an appropriate mapping – a model – from data described by features to outputs. Obtaining such a mapping from training data is what constitutes a learning problem (blue box).

Generalisation

Generalisation is a key objective of machine learning.

What we are really interested in is generalising from the sample of data in our training set.

In machine learning, generalisation means how well a trained model can classify or forecast unseen data. Training a generalised machine learning model means, in general, it works for all subsets of unseen data.

E.g., when we train a model to classify between dogs and cats, if the model is provided with a dog image dataset with only two breeds, it may obtain a good performance during training; But, it possibly gets a low classification score when it is tested by other breeds of dogs as well.

Generalisation

There are three basic assumptions for generalization:

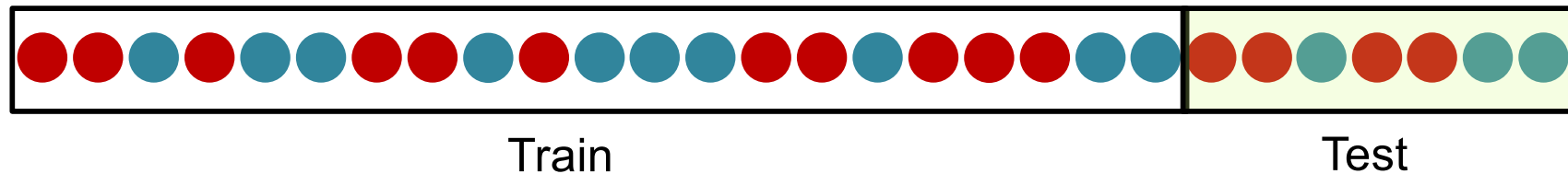
- Examples are drawn independently and identically (i.i.d) at random from the distribution;
- The distribution is stationary; that is , the distribution doesn't change within the data set
- We always pull from the same distribution (for training, validation and test samples)

In practice, we sometimes violate these assumptions.

Cross-validation

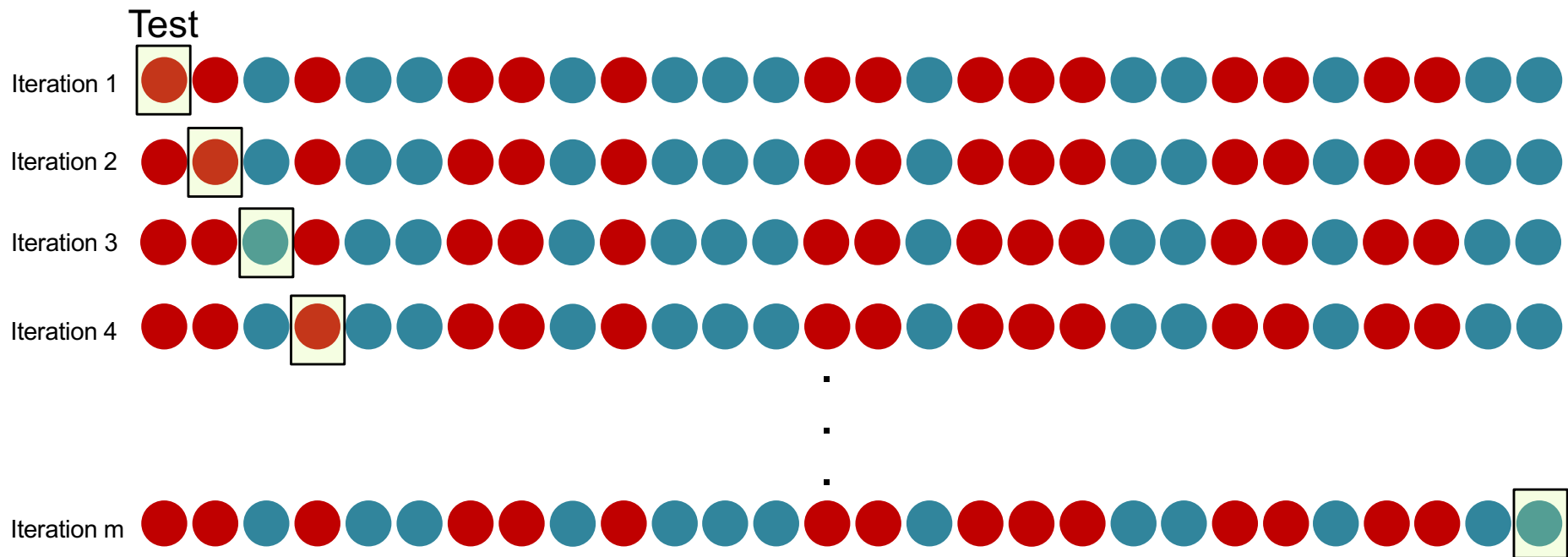
Also known as **out-of-sample testing** is validation technique to assess the results of a model to an independent data set

1. Holdout method:



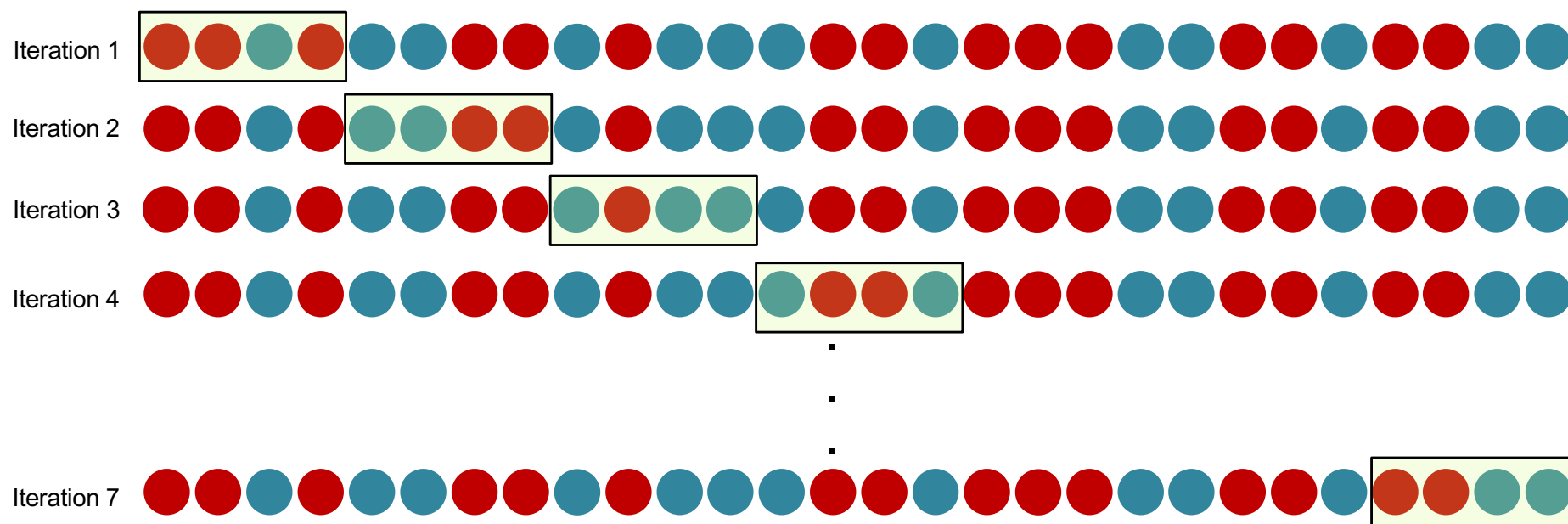
Cross-validation

2. Leave-One-Out Cross validation (LOOCV):



Cross-validation

3. K-fold Cross Validation

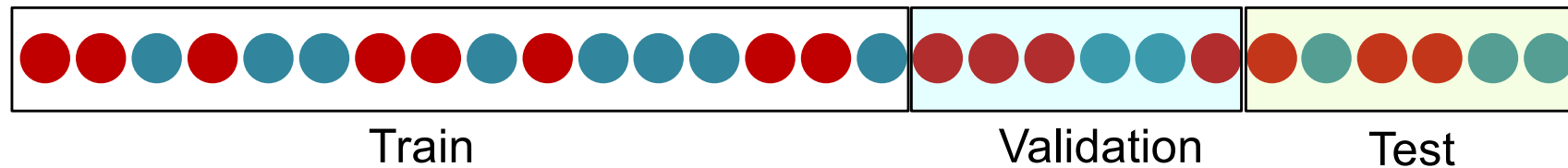


Cross-validation

There are certain parameters that need to be estimated during learning. We use the data, but NOT the training set, OR the test set. Instead, we use a separate *validation* or *development* set.

Cross-validation

Validation set: To make the hyperparameter tuning and model selection independent from the test set, we define another set within the train set



Data Types

In Machine Learning world, in general two types of data is defined:

- **Numerical**: Anything represented by numbers (e.g., integer , floating point)
- **Categorical**: everything that is not numerical (e.g. discrete labeled groups)

In general, for machine learning algorithms, data has to be represented in numeric form

Data Types

Another taxonomy of data types:

- 1. Irrelevant:** it might be represented with strings or numbers but has no relationship with the outcome (e.g. participants name or code)
- 2. Nominal:** discrete values with no numerical relationship between different categories (e.g. animal types, colors, nationality)
- 3. Binary:** discrete data with only two possibilities (e.g. cancerous vs. non-cancerous)

Data Types

4. **Ordinal:** discrete integers that can be ranked, but the relative distance between any two number can not be defined (e.g., students rank based on GPA)
5. **Count:** discrete whole numbers without any negatives
6. **Time:** a cyclical, repeating continuous form of data (e.g days, weeks)
7. **Interval:** data that the we can measure the distance between different values and is measured against a scale. (e.g., temperature, income)

Binary Classification task

In a binary classification (or binomial classification) task, we always want to classify the data of a given set into two groups. We usually define one of the classes as positive and one as negative.

- Sometimes the classes are equally important (e.g., recognition of dog vs cat in image classification)
- Sometimes misclassification in one of the classes is more costly than misclassification in the other class (e.g., predicting that someone has cancer while (s)he doesn't have vs predicting that someone doesn't have cancer while (s)he has) therefore we may prefer to have a better classification in one class in the cost of more errors in the other class.

Evaluation of error

If we have a binary classification, then we have two classes of $y \in \{0,1\}$, where we call the class $y = 1$, *positive class* and $y = 0$, *negative class*.

Some evaluation metrics:

- **True positive:** number of instances from class one that have been predicted as one
- **True negative:** number of instances from class zero that have been predicted as zero
- **False positive:** number of instances from class zero that have been predicted as one
- **False negative:** number of instances from class one that have been predicted as zero

Contingency table

For two-class prediction case:

Actual Class	Predicted Class	
	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

This is also called *confusion matrix*

Classification Accuracy

Classification Accuracy on a sample of labelled pairs $(x, c(x))$ given a learned classification model that predicts, for each instance x , a class value $\hat{c}(x)$:

$$acc = \frac{1}{|Test|} \sum_{x \in Test} I[\hat{c}(x) = c(x)]$$

where $Test$ is a test set and $I[]$ is the indicator function which is 1 iff its argument evaluates to true, and 0 otherwise.

Classification Error is $= 1 - acc$.

Other evaluation metrics

Precision/correctness

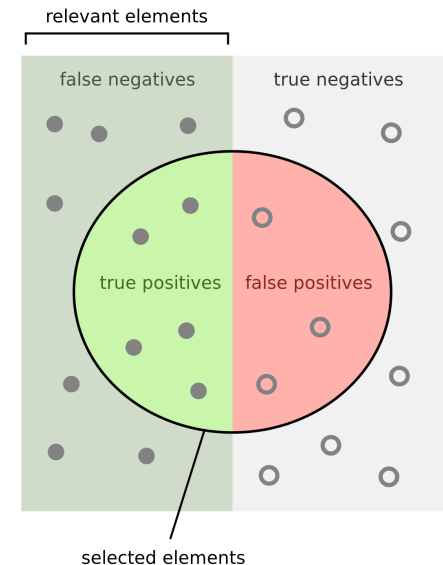
- is the number of relevant objects classified correctly divided by the total number of relevant objects classified

$$Precision = \frac{TP}{TP + FP}$$

Recall/sensitivity/completeness/true positive rate (TPR)

- is the number of relevant objects classified correctly divided by total number of relevant/correct objects

$$Recall = \frac{TP}{TP + FN}$$



How many selected items are relevant?

$$Precision = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

How many relevant items are selected?

$$Recall = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

https://en.wikipedia.org/wiki/Precision_and_recall

Other evaluation metrics

F_1 score: a measure of accuracy, which is the harmonic mean of precision and recall and is defined as:

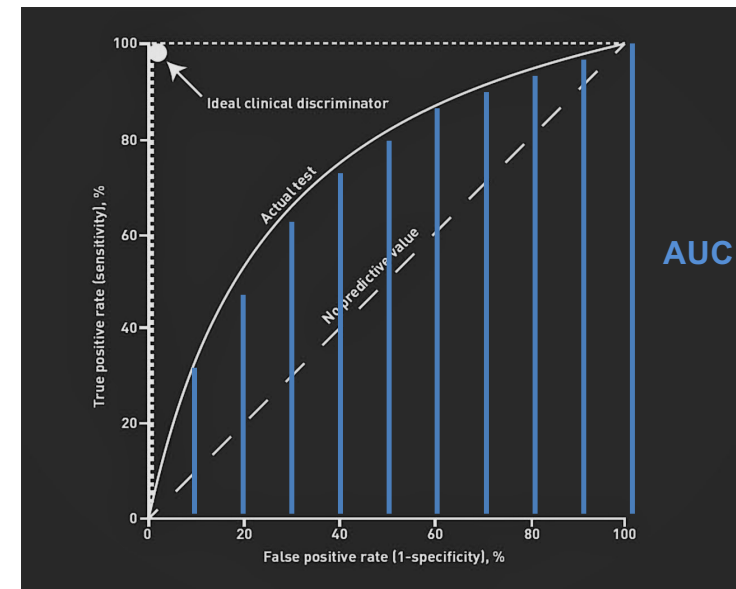
$$F_1 = 2 \frac{\textit{precision} \times \textit{recall}}{\textit{precision} + \textit{recall}}$$

This measure gives equal importance to precision and recall which is sometime undesirable; so, we have to decide which metric to use depending on the task and what's important for the task.

Other evaluation metrics

AUC-ROC curve: Area Under the Curve (AUC) – Receiver Operating Characteristics (ROC) curve is one of the most important evaluation metric for performance of classification models. This metric evaluates the model at different threshold settings and can inform us on the capability of the model in distinguishing between classes.

- $TPR = \frac{TP}{TP+FN}$
- $FPR = \frac{FP}{FP+TN}$
- A good model has *AUC* close to 1
- A very poor model has *AUC* close to 0
- *AUC* = 0.5 means no class separation



Missing Value: An issue to consider

Missing Values

- In practice it rarely happens that the data is complete and homogenous.
- Why data is incomplete:
 - Human errors
 - Sensor errors
 - Software bugs
 - Faulty preprocessing
 - ...

Missing Values

How to handle missing values (common approaches):

- Deleting samples with missing values
- Replacing the missing value with some statistics from the data (mean, median, ...)
- Assigning a unique category
- Predicting the missing values
- Using algorithms that support missing values

Missing Values

Deleting samples with missing values:

- Pros:
 - A robust and probably more accurate model
- Cons:
 - Loss of information and data
 - Works poorly if the percentage of missing values is high

Missing Values

Replacing the missing value with mean/median/mode:

- Pros:
 - When the data size is small, it is better than deleting
 - It can prevent data loss
- Cons:
 - Imputing the approximations adds bias to the model (it reduces the variance of the variable)
 - Works poorly compared to other methods

Missing Values

If categorical, assigning a unique category or the most frequent category:

- Pros:
 - Works well with small datasets and easy to implement
 - No loss of data
- Cons:
 - Unique category works only for categorical features
 - Adding another feature (e.g., a new unique category) to the model may result higher variance in the model
 - Adding the most frequent category can increase the bias in the model

Missing Values

Predicting the missing values:

- Pros:
 - Imputing the missing variable is an improvement as long as the bias from it is smaller than the omitted variable bias
 - Yields unbiased estimates of the model parameters
- Cons:
 - Bias also arises when an incomplete conditioning set is used for a categorical variable
 - Considered only as a proxy for the true values

Missing Values

Using algorithms that support missing values:

- Pros:

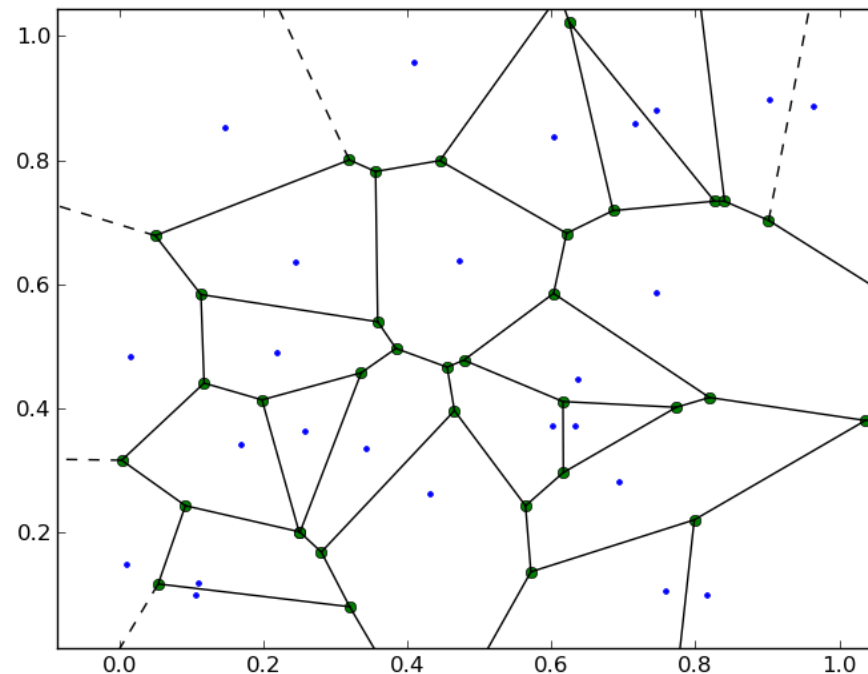
- Does not require creation of a predictive model
- Correlation of the data is neglected

- Cons:

- Some of these algorithms are very time-consuming and it can be critical in data mining where large databases are being extracted

Nearest Neighbor Algorithm for Classification

Nearest Neighbour



- Nearest Neighbour is a regression or classification algorithm that predicts whatever is the output value of the nearest data point to some query.
- To find the nearest data point, we have to find the *distance* between the query and other points. So we have to decide how to define *distance*.

Minkowski distance

Minkowski distance If $\mathcal{X} \rightarrow \mathbb{R}^d$, $x, y \in \mathcal{X}$, the Minkowski distance of order $p > 0$ is defined as:

$$Dis_p(x, y) = \left(\sum_{j=1}^d |x_j - y_j|^p \right)^{1/p} = \|x - y\|_p$$

Where $\|z\|_p = (\sum_{j=1}^d |z_j|^p)^{1/p}$ is the p – norm (sometimes denoted L_p norm) of the vector z .

Minkowski distance

- The *2-norm* refers to the familiar *Euclidean distance*

$$Dis_2(x, y) = \sqrt{\sum_{j=1}^n (x_j - y_j)^2} = \sqrt{(x - y)^T (x - y)}$$

- The *1-norm* denotes *Manhattan distance*, also called *cityblock distance*:

$$Dis_1(x, y) = \sum_{j=1}^n |x_j - y_j|$$

Minkowski distance

- If we now let p grow larger, the distance will be more and more dominated by the largest coordinate-wise distance, from which we can infer that $Dis_{\infty} = \max_j |x_j - y_j|$; this is also called *Chebyshev distance*.
- You will sometimes see references to the *0-norm* (or L_0 norm) which counts the number of non-zero elements in a vector. The corresponding distance then counts the number of positions in which vectors x and y differ. This is not strictly a Minkowski distance; however, we can define it as:

$$Dis_0(x, y) = \sum_{j=1}^d (x_j - y_j)^0 = \sum_{j=1}^d I[x_j \neq y_j]$$

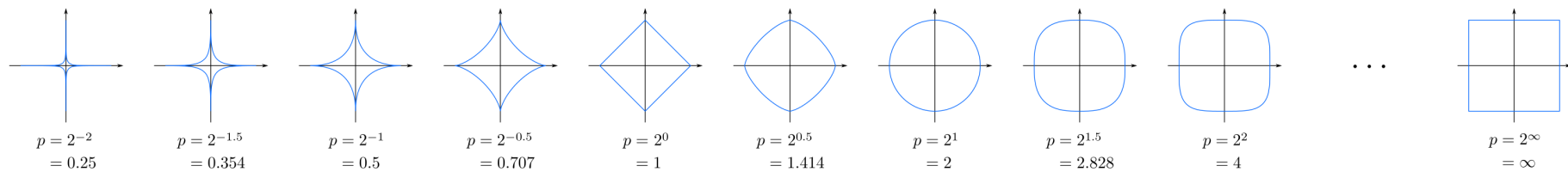
under the understanding that $x^0 = 0$ for $x = 0$ and 1 otherwise.

Minkowski distance

Sometimes the data is not naturally in \mathbb{R}^d , but if we can turn it into Boolean features, or character sequences, we can still apply distance measures. For example:

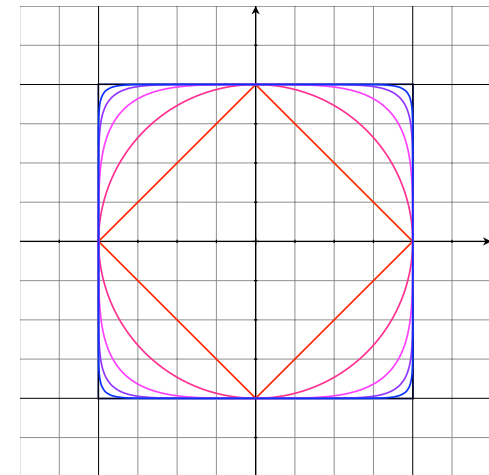
- If x and y are binary strings, this is also called the *Hamming distance*. Alternatively, we can see the Hamming distance as the number of bits that need to be flipped to change x into y .
- For non-binary strings of unequal length this can be generalised to the notion of *edit distance* or *Levenshtein distance*.

Circles and ellipses



Unite circles with different order-p Minkowski distance

- Notice that for points on the coordinate axes all distances agree
- If we require a rotation invariant distance metric, then Euclidean distance is our only choice



Distance metric

Distance metric: Given an instance space \mathcal{X} , a distance metric is a function $Dis : \mathcal{X} \times \mathcal{X} \rightarrow [0, \infty)$ such that for any $x, y, z \in \mathcal{X}$:

1. distances between a point and itself are zero: $Dis(x, x) = 0$
2. all other distances are larger than zero: if $x \neq y$ then $Dis(x, y) > 0$
3. distances are symmetric: $Dis(y, x) = Dis(x, y)$
4. detours can not shorten the distance (triangle inequality):

$$Dis(x, z) \leq Dis(x, y) + Dis(y, z)$$

- It can be shown that triangle inequality does not hold for $p < 1$

If the second condition is weakened to a non-strict inequality – i.e., $Dis(x, y)$ may be zero even if $x \neq y$ – the function Dis is called a *pseudo-metric*.

Exemplar

- Exemplar refers to a point that is used to summarise or represent a larger group of data
- Exemplars are chosen because they are considered representative of the patterns or characteristics found in the group they represent.
- In distance-based learning, exemplars can be selected based on their proximity to other points, with the aim of minimizing the distance between the exemplar and each point in its cluster.

Exemplar

- E.g., the arithmetic mean of a set of data can be used as an exemplar
- It can be shown that the arithmetic mean is the unique point that minimizes the sum of squared Euclidean distances to those data points
- If one looks for a point that minimizes the sum of Euclidean distances, then that is geometric median

Means and distances

- In certain situations, it makes sense to restrict an **exemplar to be one of the given data points**. In that case, we speak of a **medoid**, to distinguish it from a **centroid which is an exemplar that doesn't have to occur in the data**.
- Finding a medoid requires us to calculate, for each data point, the total distance to all other data points, in order to choose the point that minimises it. Regardless of the distance metric used, this is an $O(n^2)$ operation for n points.
- So, for medoids there is no computational reason to prefer one distance metric over another.
- There may be more than one medoid.

Distance-based models

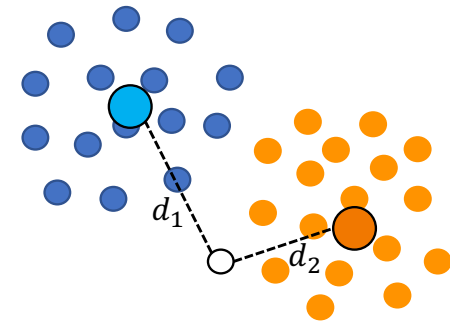
To summarise, the main ingredients of distance-based models are:

- distance metrics, which can be Euclidean, Manhattan, Minkowski or Mahalanobis, among many others;
- exemplars: centroids that find a centre of mass according to a chosen distance metric, or medoids that find the most centrally located data point.

Nearest Centroid Classifier

Nearest Centroid Classifier

- This is a classifier based on minimum distance principle, where the class exemplars are just the centroids (or means)



- Training: for training sample pairs $\{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ where x_i is the feature vector for sample i and y_i is the class label, class centroids are:

$$\mu_k = \frac{1}{|C_k|} \sum_{j \in C_k} x_j$$

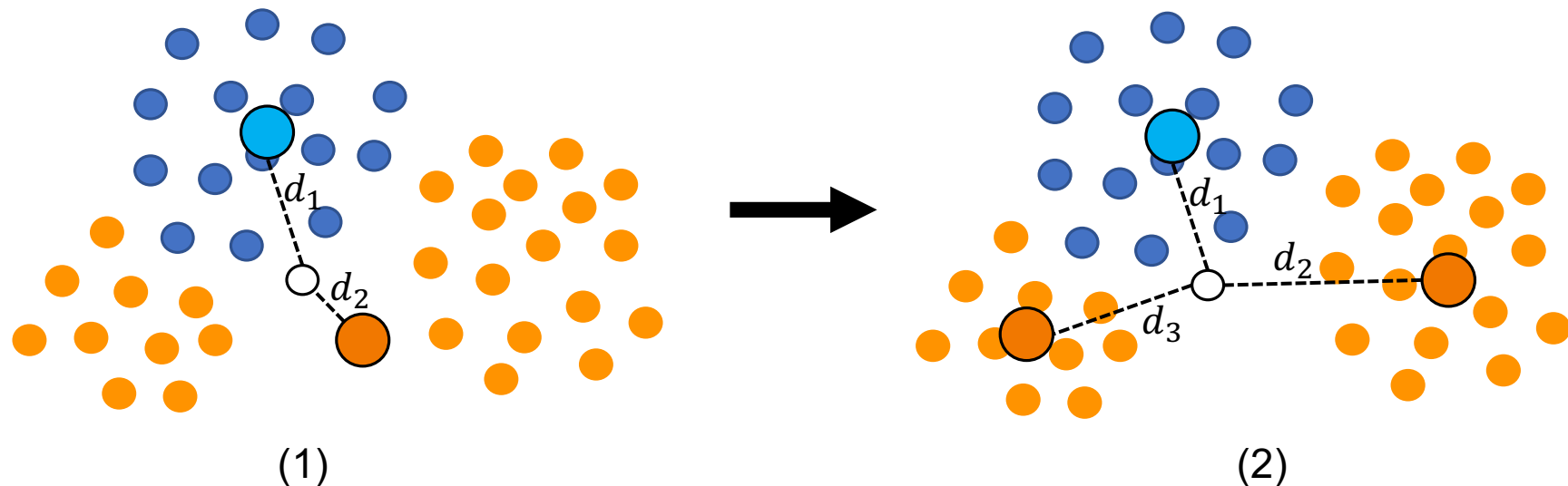
- Test: a new unknown object with feature vector x is classified as class i if it is much closer to the mean vector of class k than to any other class mean vector

Basic Linear Classifier & Nearest Centroid Classifier

- The basic linear classifier is distance-based.
- An alternative, distance-based way to classify instances without direct reference to a decision boundary is by the following decision rule: if x is nearest to μ^{\oplus} then classify it as positive, otherwise as negative; or equivalently, classify an instance to the class of the nearest exemplar.
- If we use Euclidean distance as our closeness measure, simple geometry tells us we get exactly the same decision boundary.
- So the basic linear classifier can be interpreted from a distance-based perspective as constructing exemplars that minimise squared Euclidean distance within each class, and then applying a nearest-exemplar decision rule.

Nearest Centroid Classifier

- What happens if a class has more than one mode? (similar to the image)
 1. If there is only one centroid per class, then it will perform poorly
 2. If we can somehow find different modes, we can define one centroid per each mode which helps the classifier



Nearest Centroid Classifier

Advantages:

- Simple
- Fast
- works well when classes are compact and far from each other.

Nearest Centroid Classifier

Disadvantages:

- For complex classes (e.g., Multimodal, non-spherical) may give very poor results
- Can not handle outliers and noisy data well
- Can not handle missing data

Nearest neighbour classification

Nearest neighbour classification

- Related to the simplest form of learning: rote learning or memorisation
 - Training instances are searched for instance that **most closely resembles** new or *query* instance
 - The instances themselves represent the knowledge
 - Called: *instance-based*, *memory-based* learning or *case-based* learning; often a form of *local* learning
- The *similarity* or *distance* function defines “learning”, i.e., how to go beyond simple memorisation

Nearest neighbour classification

- Intuitive idea — instances “close by”, i.e., neighbours or *exemplars*, should be classified similarly
- Instance-based learning is lazy learning
 - Methods: *nearest-neighbour*, *k-nearest-neighbour*, ...
- Ideas also important for *unsupervised* methods, e.g., clustering (later lectures)

Nearest Neighbour

Stores all training examples $\langle x_j, f(x_j) \rangle$.

Nearest neighbour:

- Given query instance x_q , first locate nearest training example x_n , then estimate $\hat{f}(x_q) \leftarrow f(x_n)$

k -Nearest neighbour:

- Given x_q , take vote among its k nearest neighbours (if discrete-valued target function) (see next slide)
- take mean of f values of k nearest neighbours (if real-valued)

$$\hat{f}(x_q) \leftarrow \frac{\sum_{j=1}^k f(x_j)}{k}$$

K-Nearest Neighbour (KNN) Algorithm

Training algorithm:

- For each training example $\langle x_j, f(x_j) \rangle$, add the example to the list *training_examples*.

Classification algorithm:

- Given a query instance x_q to be classified,
 - Let x_1, \dots, x_k be the k instances from *training examples* that are *nearest* to x_q by the distance function
 - Return

$$\hat{f}(x_q) \leftarrow \arg \max_{v \in V} \sum_{j=1}^k \delta(v, f(x_j))$$

Where $\delta(a, b) = 1$ if $a = b$ and 0 otherwise.

Distance function again

- The distance function defines what is learned.
Instance x_j is described by a feature vector (list of attribute-value pairs)

$$x_j = (x_{j1}, \dots, x_{jd})^T$$

Where x_{jr} denotes the value of the r th attribute/feature of x_j .

- Most **commonly used distance function is *Euclidean distance*** . . .
 - distance between two instances x_i and x_j is defined to be

$$Dis(x_i, x_j) = \sqrt{\sum_{r=1}^d (x_{ir} - x_{jr})^2}$$

Distance function again

Many other distance functions could be used . . .

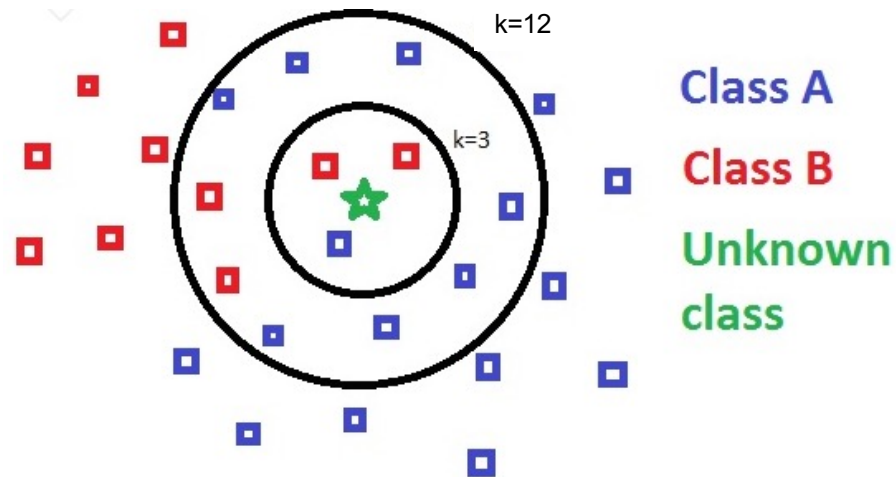
- e.g., *Manhattan* also referred to as *city-block* distance (sum of absolute values of differences between attributes)

$$Dis(x_i, x_j) = \sum_{r=1}^d |x_{ir} - x_{jr}|$$

Vector-based formalization – use norm L_1 , L_2 , ...

KNN Example

- What is the predicted class for the green point given the data for?



<https://towardsdatascience.com/knn-using-scikit-learn-c6bed765be75>

Normalisation and other issues

- Different attributes measured on different scales (for example, one attribute/feature may have a range of $[0,100]$ and another have a range of $[-1,1]$)

So, attributes have to be *normalised* (why ?)

Normalisation and other issues

- Min-max normalisation:

$$x'_{jr} = \frac{x_{jr} - \min(x_{jr})}{\max(x_{jr}) - \min(x_{jr})}$$

where x_{jr} is the actual value of attribute/feature r and x'_{jr} is the normalised value.

- Z-score is another popular normalisation

$$x'_{jr} = \frac{x_{jr} - \mu_r}{\sigma_r}$$

- For nominal (categorical) attributes with no inherent ordering:
Usually simple matching distances is used. E.g., 1 if they are the same and 0 otherwise.

When To Consider Nearest Neighbour

- Instances map to points in \mathbb{R}^d
- Less than 20 attributes per instance
 - or number of attributes can be reduced . . .
- Lots of training data
- No requirement for “explanatory” model to be learned

K-Nearest Neighbour

Advantages:

- Statisticians have used KNN since early 1950s
- Can be very accurate
- Training is very fast
- Can learn complex target functions

K-Nearest Neighbour

Disadvantages:

- Slow at query time: basic algorithm scans entire training data to derive a prediction
- “Curse of dimensionality”
- Assumes all attributes are equally important, so easily fooled by irrelevant attributes
 - Remedy: attribute selection or weights
- Problem of noisy instances:
 - Remedy: remove from data set
 - not easy – how to know which are noisy ?
- Needs homogenous feature type and scale
- Finding the optimal number of neighbors (k) can be challenging

Nearest-neighbour classifier

- 1NN perfectly separates training data, so low bias but high variance
- By increasing the number of neighbours k we increase bias and decrease variance (what happens when $k = m$ (m is the number of observations)?)
- Easily adapted to real-valued targets, and even to structured objects (nearest-neighbour retrieval). It can also output probabilities when $k > 1$

Distance-Weighted KNN

- Might want to weight nearer neighbours more heavily ...
 - Use distance function to construct a weight w_i
 - Replace the final line of the classification algorithm by:

$$\hat{f}(x_q) \leftarrow \arg \max_{v \in V} \sum_{i=1}^k w_i \delta(v, f(x_j))$$

Where,

$$w_i = \frac{1}{Dis(x_q, x_i)^2}$$

$Dis(x_q, x_i)$ is distance between x_q, x_i

Distance-Weighted KNN

For real-valued target functions replace the final line of the algorithm by:

$$\hat{f}(x_q) \leftarrow \frac{\sum_{i=1}^k w_i f(x_i)}{\sum_{i=1}^k w_i}$$

(denominator normalizes contribution of individual weights).

Now we can consider using all the training examples instead of just k :

- using all examples (i.e., when $k = m$ and m is number of training samples) with the rule above is called *Shepard's method*

Evaluation

Lazy learners do not construct an explicit model, so how do we evaluate the output of the learning process ?

- 1-NN – training set error is always zero !
 - each training example is always closest to itself
- k -NN – overfitting may be hard to detect

Solution:

Leave-one-out cross-validation (LOOCV) – leave out each example and predict it given the rest:

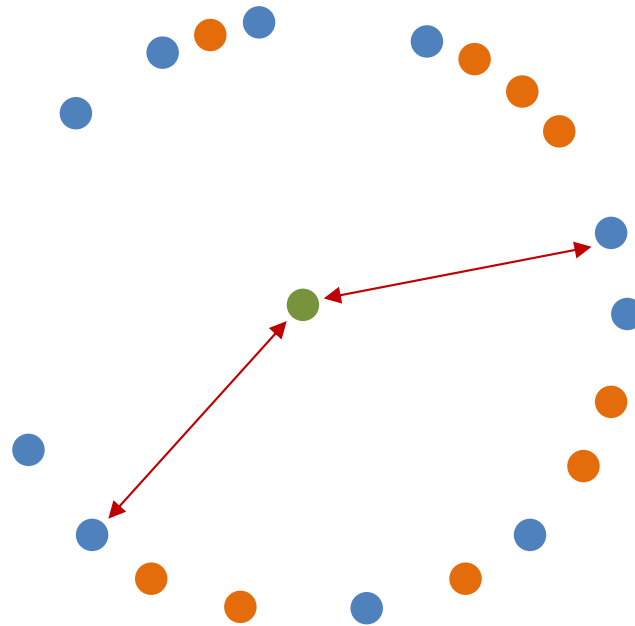
$$(x_1, y_1), (x_2, y_2), \dots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \dots, (x_m, y_m)$$

Error is mean over all predicted examples. Fast – no models to be built !

KNN Computational Time

- KNN uses the training data as exemplars, so using simple search for prediction is $O(n)$!
- There are algorithms to search for neighbours more efficiently with $O(\log n)$ but they do not work very well for above 10 dimensions (more than 10 features/attributes)
- For above 10 dimension, there are some approximate nearest neighbour approaches that can improve computation by orders of magnitude
- In high dimensional space (e.g., above 20 dimensions) even with using such algorithms, the KNN doesn't work well
- In high-dimensional spaces everything is far away from everything and so pairwise distances are uninformative (curse of dimensionality)

When is KNN meaningful?

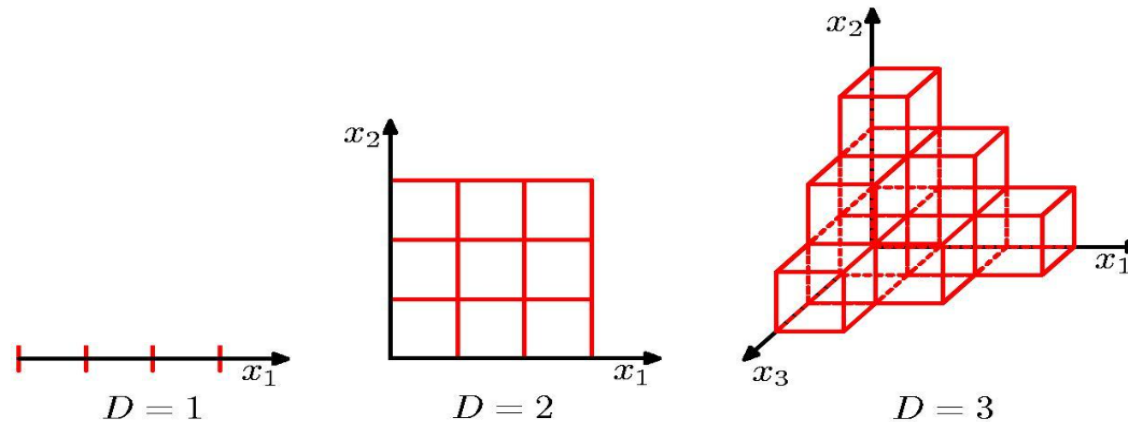


You may think that this is an exceptional example, and this doesn't really happen in practice!!

Curse of Dimensionality

- It can be shown that as dimensions increase the effectiveness of distance metrics decrease and the concept of proximity may not be qualitatively meaningful as all points look equidistant
- This is one symptom of having high dimensional space (curse of dimensionality)
- There are also other problems arising from curse of dimensionality:
 - It becomes polynomially harder to estimate many parameters (e.g., covariances)
 - It becomes more difficult to visualize data
 - Enormous amount of data is needed to train a model

Curse of Dimensionality



- number of “cells” in the instance space grows exponentially in the number of features
- with exponentially many cells we would need exponentially many data points to ensure that each cell is sufficiently populated to make nearest-neighbour predictions reliably

Curse of Dimensionality

- Bellman (1960) coined this term in the context of dynamic programming
- Imagine instances described by 20 attributes, but only 2 are relevant to target function — “similar” examples will appear “distant”.
- *Curse of dimensionality*: nearest neighbour is easily misleading when dealing with high-dimensional x in terms of the number of features – problem of irrelevant attributes

Curse of Dimensionality

One approach:

- Stretch j th axis by weight z_j , where z_1, \dots, z_d chosen to minimize prediction error
- Use cross-validation to automatically choose weights z_1, \dots, z_d
- Note, setting z_j to zero eliminates this dimension altogether

Curse of Dimensionality

Some ideas to address this for instance-based (nearest-neighbour) learning

- Euclidean distance with weights on attributes

$$Dis(x_q, x_i) = \sqrt{\sum_{r=1}^d z_r (x_{qr} - x_{ir})^2}$$

- updating of weights based on nearest neighbour classification error
 - class correct/incorrect: weight increased/decreased
 - can be useful if not all features used in classification

See Moore and Lee (1994) “Efficient Algorithms for Minimizing Cross Validation Error”

Instance-based (nearest-neighbour) learning

Recap – Practical problems of NN scheme:

- Slow (but fast *k-dimensional tree*-based approaches exist)
 - Remedy: removing irrelevant data
- Noise (but KNN copes quite well with noise)
 - Remedy: removing noisy instances
- All attributes deemed equally important
 - Remedy: attribute weighting (or simply selection)

Dealing with noise

Use larger values of k (why ?) How to find the “right” k ?

- One way: cross-validation-based k -NN classifier (but slow)
- Different approach: discarding instances that don't perform well by keeping success records of how well an instance does at prediction (IB3)

IB3 (Instance-Based Learning 3): store classification performance information with each instance & only use in prediction if above a threshold

KNN Example

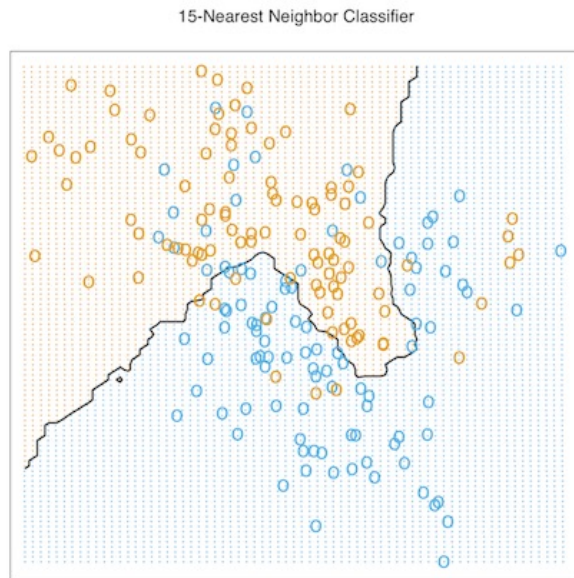


FIGURE 2.2. The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable (BLUE = 0, ORANGE = 1) and then fit by 15-nearest-neighbor averaging as in (2.8). The predicted class is hence chosen by majority vote amongst the 15-nearest neighbors.

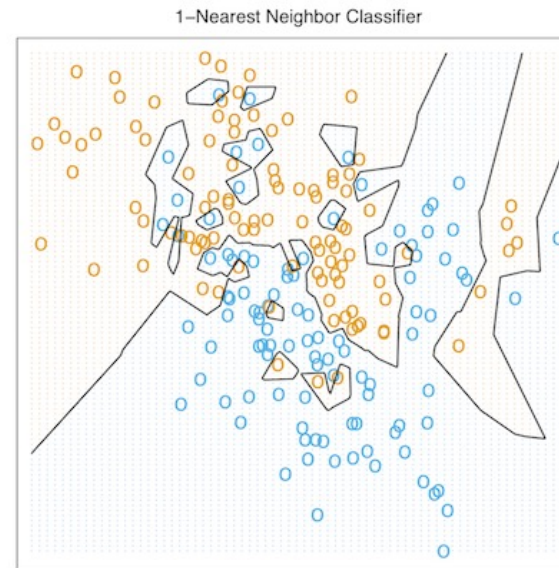


FIGURE 2.3. The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable (BLUE = 0, ORANGE = 1), and then predicted by 1-nearest-neighbor classification.

KNN Example

- Automated MS-lesion segmentation by KNN
- They have used some manually labeled image as the training set
- They used 4 features: Intensity and voxel locations (x,y,z coordinates)

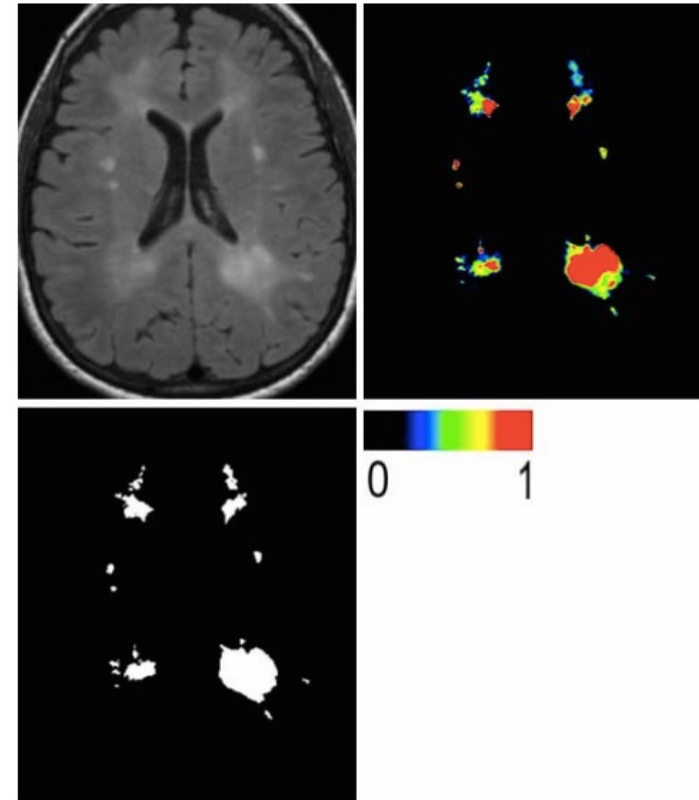


Figure 1 MS-lesion segmentation results. Top left: FLAIR image; top right: probabilistic segmentation, showing probability of lesion per voxel (see color bar); down left: binary segmentation, derived from probabilistic segmentation with threshold 0.4.

- Ref: Anbeek et. Al, "Automated MS-lesion segmentation by K-nearest neighbor classification", MIDAS journal 2008

Summary

- General frameworks for classification
- Logistic Regression models the probability of a class.
- Classification viewed in terms of distance in feature space
- Distance-based learning
- Nearest neighbour classifiers

Acknowledgements

- Material derived from slides for the book “Elements of Statistical Learning (2nd Ed.)” by T. Hastie, R. Tibshirani & J. Friedman. Springer (2009) <http://statweb.stanford.edu/~tibs/ElemStatLearn/>
- Material derived from slides for the book “Machine Learning: A Probabilistic Perspective” by P. Murphy MIT Press (2012) <http://www.cs.ubc.ca/~murphyk/MLbook>
- Material derived from slides for the book “Machine Learning” by P. Flach Cambridge University Press (2012) <http://cs.bris.ac.uk/~flach/mlbook>
- Material derived from slides for the book “Bayesian Reasoning and Machine Learning” by D. Barber Cambridge University Press (2012) <http://www.cs.ucl.ac.uk/staff/d.barber/brml>
- Material derived from slides for the book “Machine Learning” by T. Mitchell McGraw-Hill (1997) <http://www-2.cs.cmu.edu/~tom/mlbook.html>
- Material derived from slides for the course “Machine Learning” by A. Srinivasan BITS Pilani, Goa, India (2016)