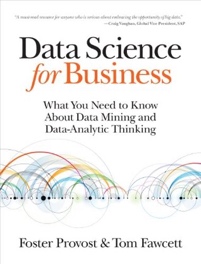
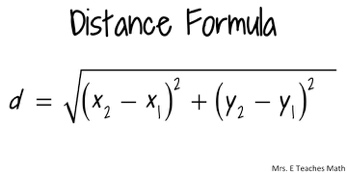
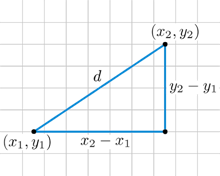
**Topic 5: Machine Learning: Classification & Clustering**

* + 1. **Calculating and interpreting similarity and distance (Ch. 6 & 7)**



1. **Calculate the Euclidean distance**

Pythagorean theorem

1. **Define nearest neighbors and combining function**

The principle behind nearest neighbor methods is to find a predefined number of training samples closest in distance to the new point and predict the label from these.

Combing functions creates a new function by adding two functions.

Voting or Averaging are combined functions.

1. **Explain how combing functions can be used for classification**

There are many general combining algorithms, such as Bagging, Boosting, or Error Correcting Output Coding that significantly improve the classifier such as decision trees, rule learners, or neural networks.

Unfortunately, these combining methods developed do not improve the nearest neighbor classifiers. Instead we can combine multiple KNN classifiers based on different distance functions for something better.

1. **Define weighted voting (scoring) or similarity moderated voting (scoring)**

﻿ To incorporate this concern, nearest-neighbor (NN) methods often use weighted voting or similarity moderated voting such that each neighbor’s contribution is scaled by its similarity.

1. **Calculate contributions using weighted voting for classification**

Consequently, when using weighted scoring the exact value of k is much less critical than with majority voting or unweighted averaging.

Table

Description automatically generated

Table

Description automatically generated

Weight is the inverse of squared distance.

1. **Explain how k in k-NN can be used to address overfitting**

A close up of a map

Description automatically generated

﻿ Smoothing is typically a desirable property for generalization. While it technically depends on the characteristics of the dataset, increasing k should reduce over-fitting but once k is too large the smoothing effect you intuited results in decreased variance which will affect overall performance negatively.

1. **Discuss issues with nearest-neighbor methods with a focus on intelligibility, dimensionality, and computational efficiency.**

Intelligibility

If model intelligibility and justification are critical, nearest-neighbor methods should be avoided.

Dimensionality and domain knowledge

Too many features are not good for nearest neighbors. Curse of dimensionality.

Computational efficiency

One benefit of nearest-neighbor methods is that training is very fast because it usually involves only storing the instances.

1. **Define and discuss the curse of dimensionality.**

Having too many features and not enough data. You need to do feature selection.

* + 1. **Discussing technical details related to similarities and neighbors (Ch. 6)**

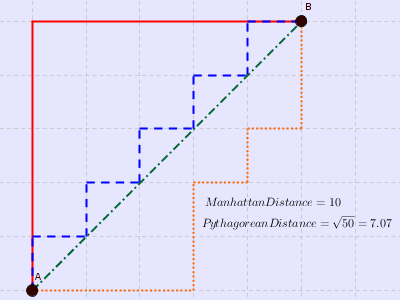
1. **Calculate the Manhattan Distance and the Cosine distance**

Manhattan distance

Definition: The distance between two points measured along axes at right angles.

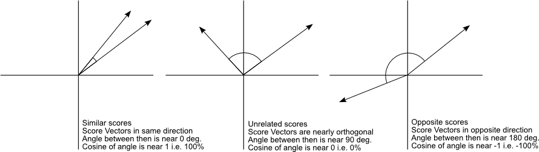
In a plane with p1 at (x1, y1) and p2 at (x2, y2), it is **|x1 - x2| + |y1 - y2|.**

Manhattan uses absolute values. Think stairs to get to distance.

Cosine similarity

It is the measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. The **cosine** of 0° is 1, and it is less than 1 for any angle in the interval (0, π) radians.



A black and red text

Description automatically generated

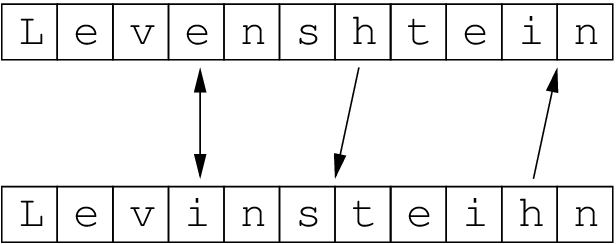
1. **Define the Jaccard Distance**



Note: Picard sharing. Holding two things which share characteristics.

Given two objects, X and Y, the Jaccard distance is the proportion of all the characteristics that either has that are shared by the two.

1. **Define edit distance or Levenshtein metric**



Note: How many licks to get to the center of a Tootsie Pop?

A close up of a logo

Description automatically generated

How many EDITS to change one word into another? 🡪 insert, delete, substitute

Informally, the Levenshtein distance between two words is the minimum number of single-character edits (i.e. insertions, deletions or substitutions) required to change one word into the other.

1. **Define clustering, hierarchical clustering, and dendrogram**

Clustering:

Fundamentally, classification is about predicting a label and regression is about predicting a quantity.

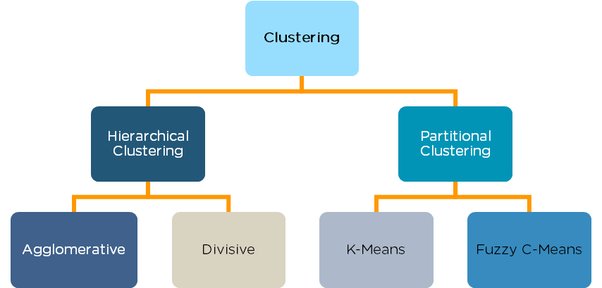
**﻿**

This is unsupervised segmentation.

﻿Modern retailers such as Amazon and Netflix use similarity to provide recommendations of similar products or from similar people.

﻿Also, can be used in fields such as medicine and law.

Clustering involves the grouping of similar objects into a set known as cluster.



Hierarchical clustering, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters.

1. **Describe how a dendrogram can help decide the number of clusters**

**﻿**

**A screenshot of a cell phone

Description automatically generated**

* ﻿﻿ **Agglomerative**: This is a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
* **Divisive**: This is a "top-down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

1. **List the advantages of hierarchical clustering**

Advantages:

Hierarchical clustering outputs a hierarchy, (i.e., a structure that is more informative than the unstructured set of flat clusters returned by k-means). Therefore, it is easier to decide on the number of clusters by looking at the dendrogram. Easy to implement

An advantage of hierarchical clustering is that it allows the data analyst to see the groupings — the “landscape” of data similarity — before deciding on the number of clusters to extract.

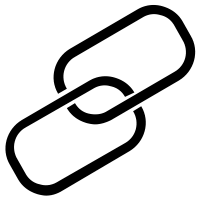
Disadvantages:

It is not possible to undo the previous step: once the instances have been assigned to a cluster, they can no longer be moved around. Note: “Think steps of the diagram. Can’t undo.”

Hierarchical clustering is generally slower, as it needs to know the distances between all pairs of clusters on each iteration, which at the start is all pairs of data points

* Time complexity: not suitable for large datasets
* Initial seeds have a strong impact on the final results
* The order of the data has an impact on the final results
* Very sensitive to outliers

1. **Define linkage functions**



﻿For hierarchical clustering, we need a distance function between clusters, considering individual instances to be the smallest clusters. This is sometimes called the linkage function.

So, for example, the linkage function could be “the Euclidean distance between the closest points in each of the clusters,” which would apply to any two clusters.

1. **Describe how distance measure can be used to decide the number of clusters in the dendrogram**

﻿ ﻿A very common proxy for the similarity of two entities is the distance between them in the instance space defined by their feature vector representation.

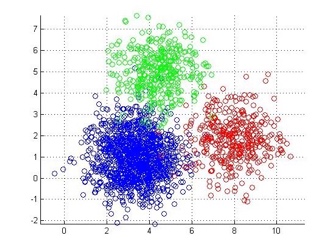
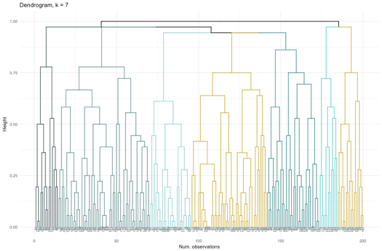
1. **Define “cluster center” or centroid and k-means clustering**

A cluster centroid for a particular cluster is the coordinate-wise mean of all of the vectors in the training data that have been deemed to be in that cluster.

This is a bit circular, since the vectors that are in that cluster are those that are closest to the centroid.

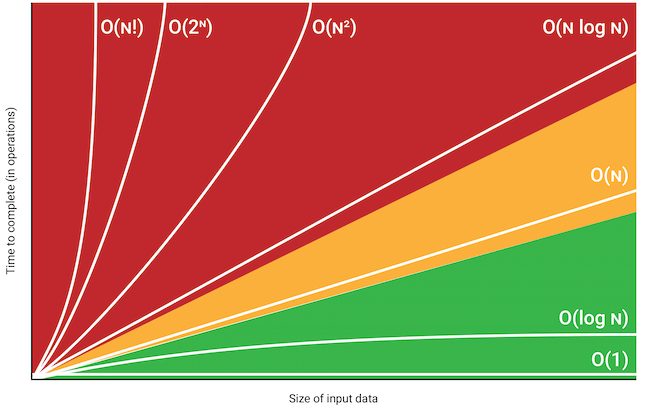
1. **Compare and contrast K-means clustering VS. hierarchical clustering (dendrogram).**

Hierarchical clustering can't handle big data well, but K-Means clustering can.

 VS. 

Hierarchical not good with a lot of data. K-means, however, is good with a lot of data.

This is because the time complexity of K Means is linear i.e. O(n) while that of hierarchical clustering is quadratic (i.e. O(n2)).

A drawing of a person

Description automatically generated greedy algorithm

1. **Describe k-means clustering**

*k*-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. *k*-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

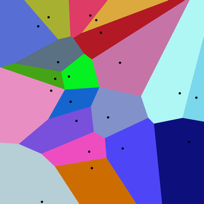
K-Means Advantages:1) If a lot of data, then K-Means most of the times computationally faster than hierarchical clustering, if we keep k small.  
2) K-Means produce tighter clusters than hierarchical clustering, especially if the clusters are globular.

K-Means Disadvantages:1) Difficult to predict K-Value.   
2) With global cluster, it doesn’t work well.  
3) Different INITIAL partitions can result in different final clusters.  
4) Not good with clusters (in the original data) of Different size and Different density

5) Attributes have to be numeric. Not categories. To get averages for the centroids.  
6) Outliers are bad. Get rid of them with kk plus plus.

Unsupervised.

This results in a partitioning of the data space into Voronoi cells.



1. **Define the reason for running k-means clustering many times**

You start by dropping K points of data randomly into the dataset. These will be your starting centroids.

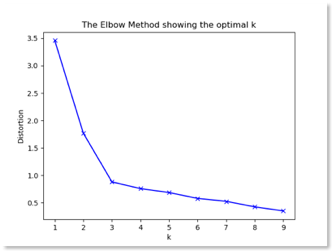
You run k-means clustering many times because you have the multiple iterations of distances to relocate the centroids to optimal positions.

In order to find the optimal centroids of the clusters

1. **Define clusters distortion**

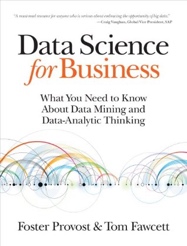
The k-means algorithm tries to minimize distortion, which is defined as the sum of the squared distances between each observation vector and its dominating centroid.

1. {\displaystyle d\_{K}={\frac {1}{p}}\min \_{c\_{1}\ldots c\_{K}}{E[(X-c\_{X})^{T}\Gamma ^{-1}(X-c\_{X})]}}**Describe the method for selecting k in the k-means algorithm**

Elbow method is used for selecting K means.

* + 1. **Evaluating classifiers** Provost (Ch. 7)



1. **Define Accuracy AND Error Rate**

Accuracy Rate

**Everything True** / Everything

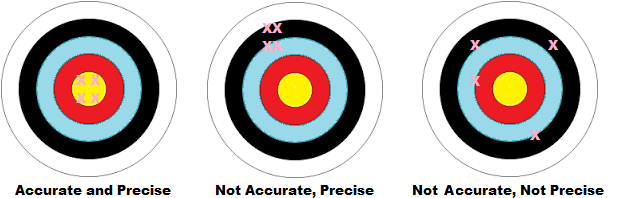
(TP + TN) / (TP + TN + FP + FN)

**Error Rate or Misclassification Rate**

**Everything False** / Everything

(FP+FN)/(TP + TN + FP + FN)

Error rate = 1 - Accuracy



Accurate is how close to target value.

Precise How tightly they are grouped.

Accuracy can be a misleading metric for imbalanced data sets.

1. **Describe a confusion matrix**



A **confusion matrix** is a table that is often used to describe the performance of a classification model (or “classifier”) on a set of test data for which the true values are known or supervised. In unsupervised learning it is usually called a “matching matrix.”

1. **Define false positives and false negatives**

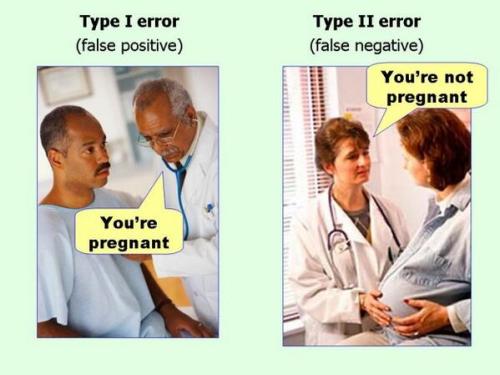
Type 1 is a False Positive. This is your level of significance. Alpha.

Innocence: Reject when True. Alpha Images, Stock Photos & Vectors | Shutterstock



Type 2 is a False Negative. This is your beta. 

Innocence: Accept when False



1. **Identify false positives and false negatives from a confusion matrix**



Positive and negatives will be on axis.

1. **Describe the problems with unbalanced data**

A close up of text on a white background

Description automatically generated

Accuracy will break down for unbalanced data.

Classification accuracy is the most-used metric for evaluating classification models.

The reason for its wide use is because it is easy to calculate, easy to interpret, and is a single number to summarize the model’s capability.

Example:

Predict if a scientific publication will lead to a Nobel Prize.

Always say “no” and you are 99.99% accurate.

You would prefer classifier A even though higher accuracy would be with B classifier.

Give more weight to the false negatives.

1. **Describe the problems with unequal costs and benefits.**



Accuracy makes no distinction between false positive and false negative errors. Tumors are more costly.

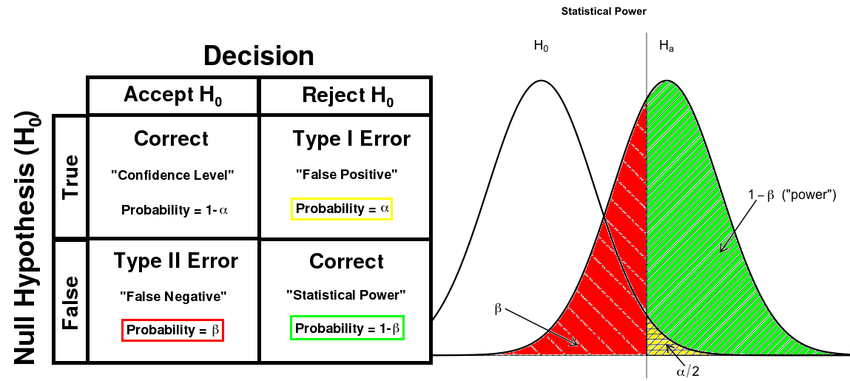
By counting them together, it makes the tacit assumptions that both are equally important. Not so. Type II error is much worse than type I error for cancer patients.

A picture containing person, man, indoor, sitting

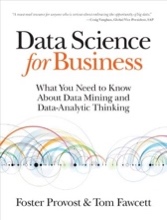
Description automatically generatedA picture containing screen, clock

Description automatically generated

Benjamin Franklin



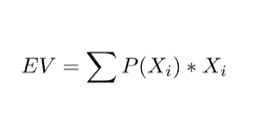
* + 1. **Describing a key analytical framework and calculating expected values (Ch. 7)**



1. **Calculate expected value and expected benefit.**

Straight forward formula.

In statistics and probability analysis, the expected value is calculated by multiplying each of the possible outcomes by the likelihood each outcome will occur and then summing all of those values.

﻿

1. **Describe how expected value (EV) can be used to frame classifier use.**

Essentially, the EV is the long-term average value of the variable.

A picture containing logo

Description automatically generated

1. **Describe how expected value (EV) can be used to frame classifier evaluation.**

You simply measure the number of correct decisions your classifier makes, divide by the total number of test examples, and the result is the accuracy of your classifier. It's that simple.

Table

Description automatically generated

Accuracy of Classifier = Correct /ALL

1. **Define class priors**

Prior probability, in Bayesian statistical inference, is the probability of an event before new data is collected.

This is the best rational assessment of the probability of an outcome based on the current knowledge before an experiment is performed. Think of it as the Base Rate.

1. **Calculate expected profit using priors.**

﻿Expected Profit = matrix multiplication of probability and cost-benefit matrix

*E[X]* = .14 \* [.67 \* 305 + .33 \* 0] +.86 \* [.16 \* -15 + .84 \* 0]

1. **Describe the two pitfalls common to formulating cost-benefit analysis.**
2. Potential Inaccuracies in Identifying and Quantifying Costs and Benefits
3. Subjectivity for Intangible Costs and Benefits. Goodwill.
4. What is the % Interest Rate?
5. **Define precision and recall**

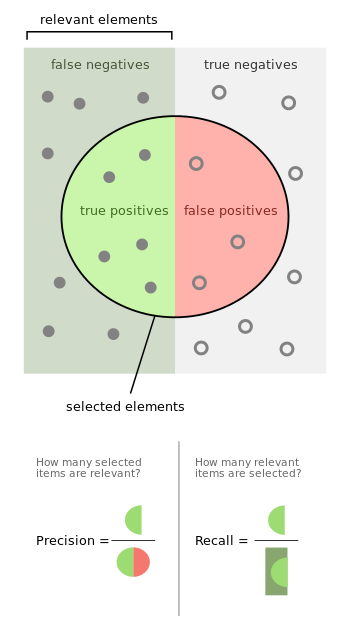
 Note: Total recall

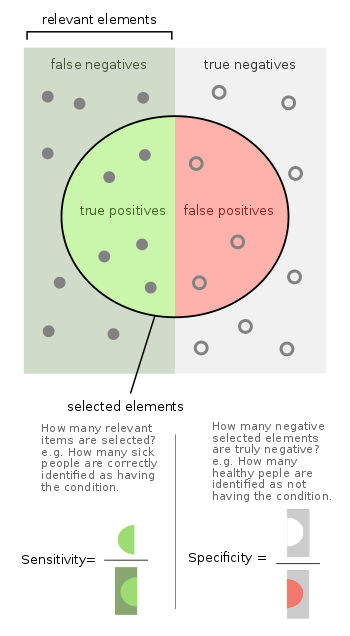
Note: P-spot

Precision = True positive / True positives + False positives

High Precision indicates data labeled as positive is indeed positive (small number of FP)

Recall = True positive / True positives + False negatives

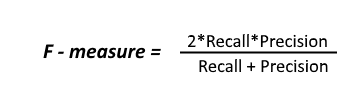




**High recall, low precision:** This means that most of the positive examples are correctly recognized (low FN) but there are a lot of false positives.

**Low recall, high precision:** This shows that we miss a lot of positive examples (high FN) but those we predict as positive are indeed positive (low FP)

**F-measure:**  
Since we have two measures (Precision and Recall) it helps to have a measurement that represents both of them. We calculate an F-measure which uses Harmonic Mean in place of Arithmetic Mean as it punishes the extreme values more.

The F-Measure will always be nearer to the smaller value of Precision or Recall.  


1. **Calculate the value of the F-MEASURE or F1 score or F score**

The F measure (F1 score or F score) is a measure of a test's ACCURACY and is defined as the weighted HARMONIC mean of the precision and recall of the test.

The range for F1 Score is (0 – 1). With 1 being the best. o



{\displaystyle \left({\frac {1^{-1}+4^{-1}+4^{-1}}{3}}\right)^{-1}={\frac {3}{{\frac {1}{1}}+{\frac {1}{4}}+{\frac {1}{4}}}}={\frac {3}{1.5}}=2\,.}It tells you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances).

The F-Measure will always be nearer to the smaller value of Precision or Recall.