Universidade do Minho
Escola de Engenharia
Departamento de Informática

Mestrado Integrado em Engenharia Informática
Mestrado em Engenharia Informática
Aprendizagem e Extração de Conhecimento
2020/2021

Filipe Gonçalves, César Analide, Paulo Novais

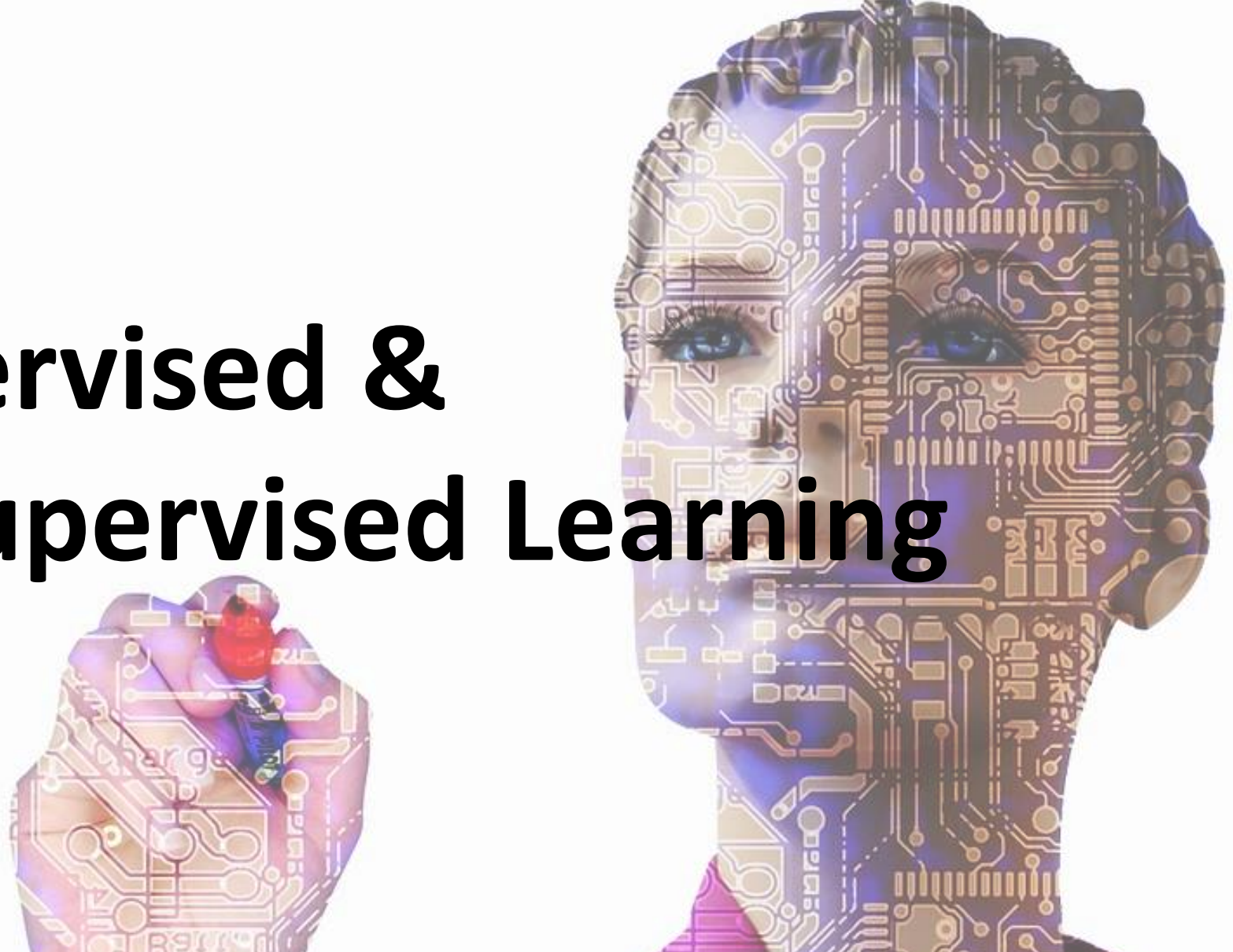
ISLab
Synthetic Intelligence Lab

AEC @ 2020/2021

- Paulo Novais – pjon@di.uminho.pt
- César Analide – analide@di.uminho.pt
- Filipe Gonçalves – fgoncalves@algoritmi.uminho.pt

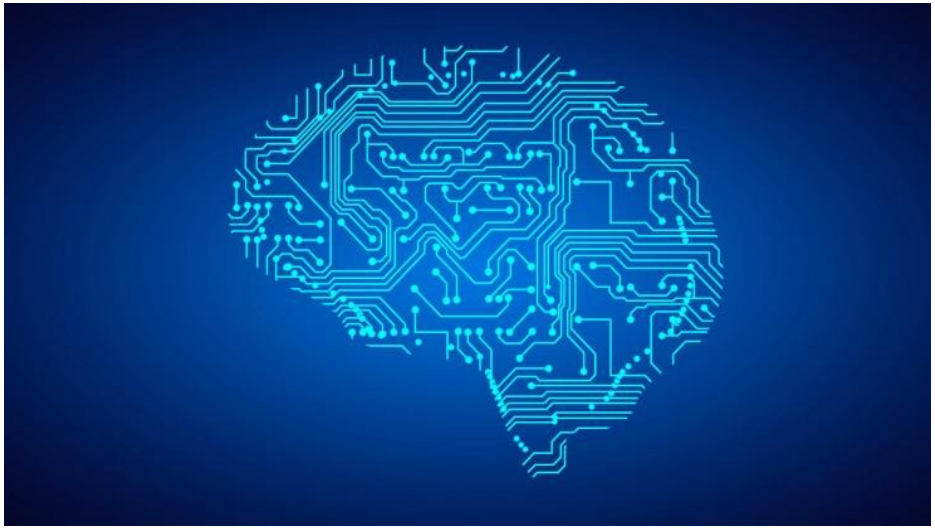
- Departamento de Informática
  Escola de Engenharia
  Universidade do Minho
- ISLab – (Synthetic Intelligence Lab)
- Centro ALGORITMI
  Universidade do Minho

Supervised & Unsupervised Learning



What is machine learning?

- Algorithms that can learn from observational data, and can make predictions based on it
  - Technique based on brain's activity



Unsupervised Learning

- The model is not giving any “answers” to learn from - it must make sense of the data just given the observations themselves.
- Example: group/cluster objects together into 2 different sets. However, no information is provided stating what the “right” set is for any object ahead of time.




Do I want big and small things? Round and square things? Red and blue things? Unsupervised learning could give me any of those results.

Unsupervised Learning

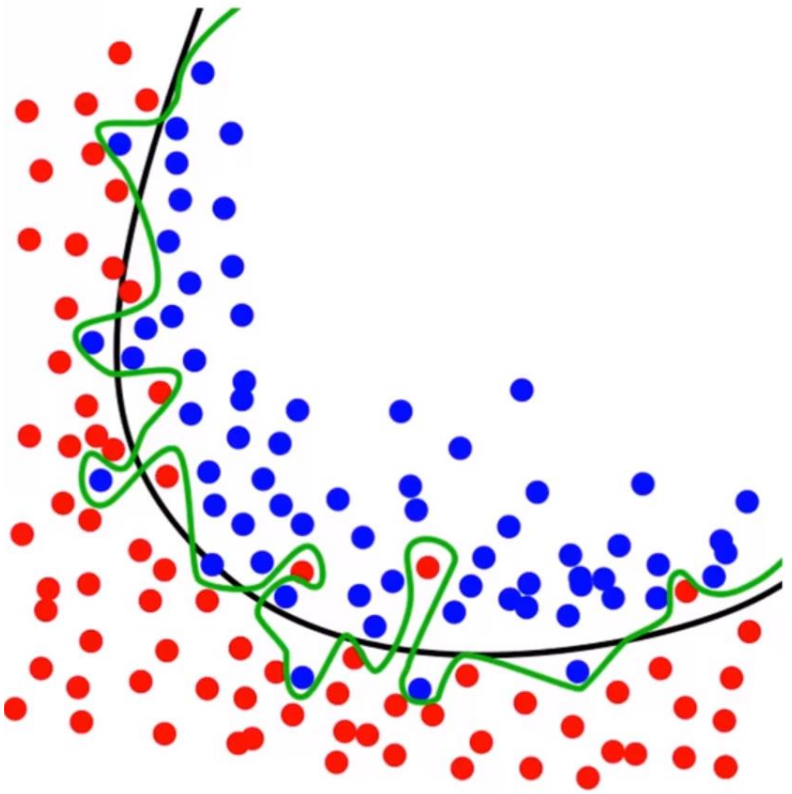
- Objective: looking for relatable/latent variables without the “correct” answers
- Example:
  - Cluster users on a dating site based on their information and behaviour
  - Cluster movies based on their properties (current concepts of genre may be outdated?)
  - Analyze the text of product descriptions to find the terms that carry most meaning for a certain category

Supervised Learning

- Objective: the data the algorithm “learns” from outcomes with the “correct” answers
- The model created is then used to predict the answer for new, unknown values
- Example:
  - Train a model for predicting car prices based on car attributes using historical sales data
  - Model can predict the optimal price for new cars that haven’t been sold before

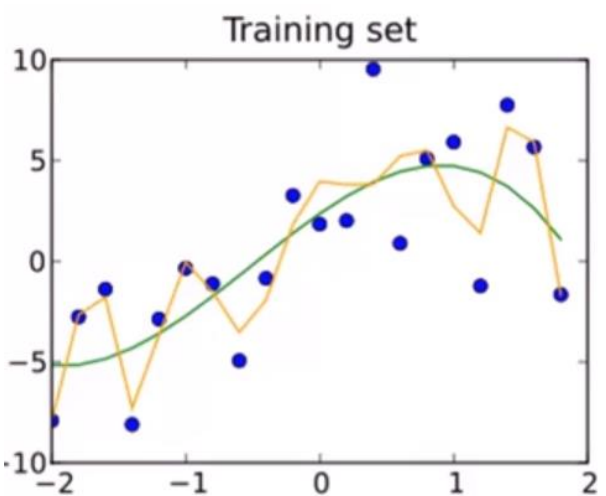
Train / Test in practice

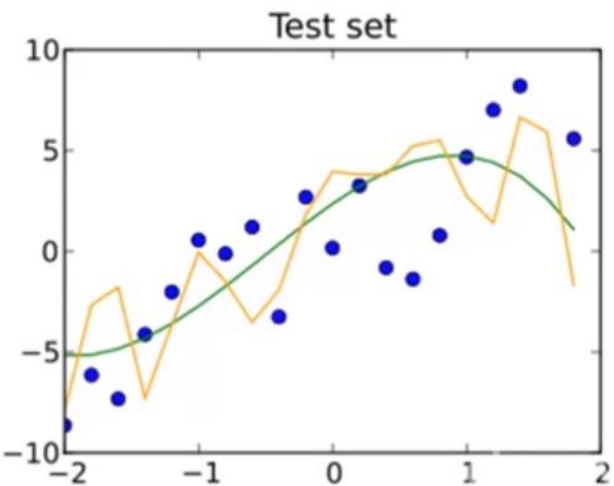
- Need to ensure both sets are large enough to contain representatives of all the variations and outliers in the data you care about
- Datasets must be selected randomly
- Train / Test is a great way to guard agaisnt overfitting



Train / Test is not Infallible

- Maybe your sample sizes are too small
- Or due to random change your train & test cases look remarkably similar
- Overfitting can still happen





K-fold Cross Validation

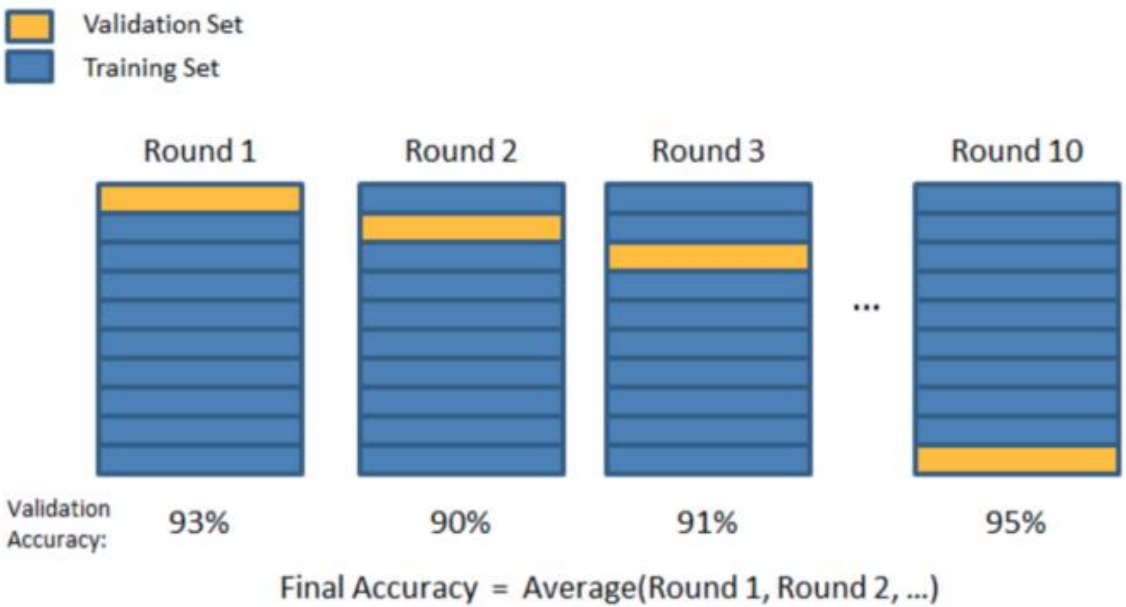
- One way to further protect against overfitting is K-fold cross validation
- How it works?

1) Split your full dataset into K randomly-assigned segments

2) Reserve one segment as your test data

3) Train on each of the remaining K-1 segments and measure their performance against the test set

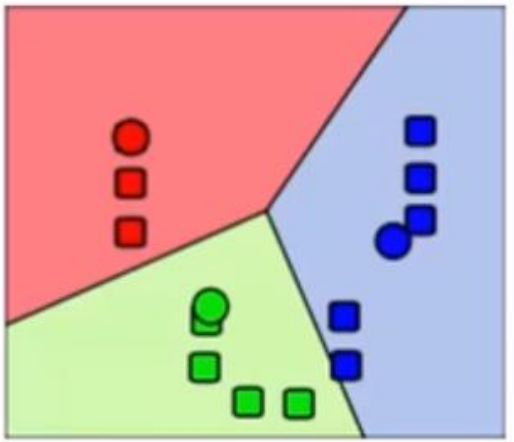
4) Take the average of the K-1 evaluation scores



K-Means Clustering

K-Means Clustering

- Attempts to split data into K groups that are closest to K centroids
- Unsupervised learning – uses only the positions of each data point
- Can uncover interesting groupings of people / objects / behaviours
- Example:
  - Where do millionaires live?
  - Genres of music / movies that naturally fall out of data?
  - Create your own stereotypes from demographic data



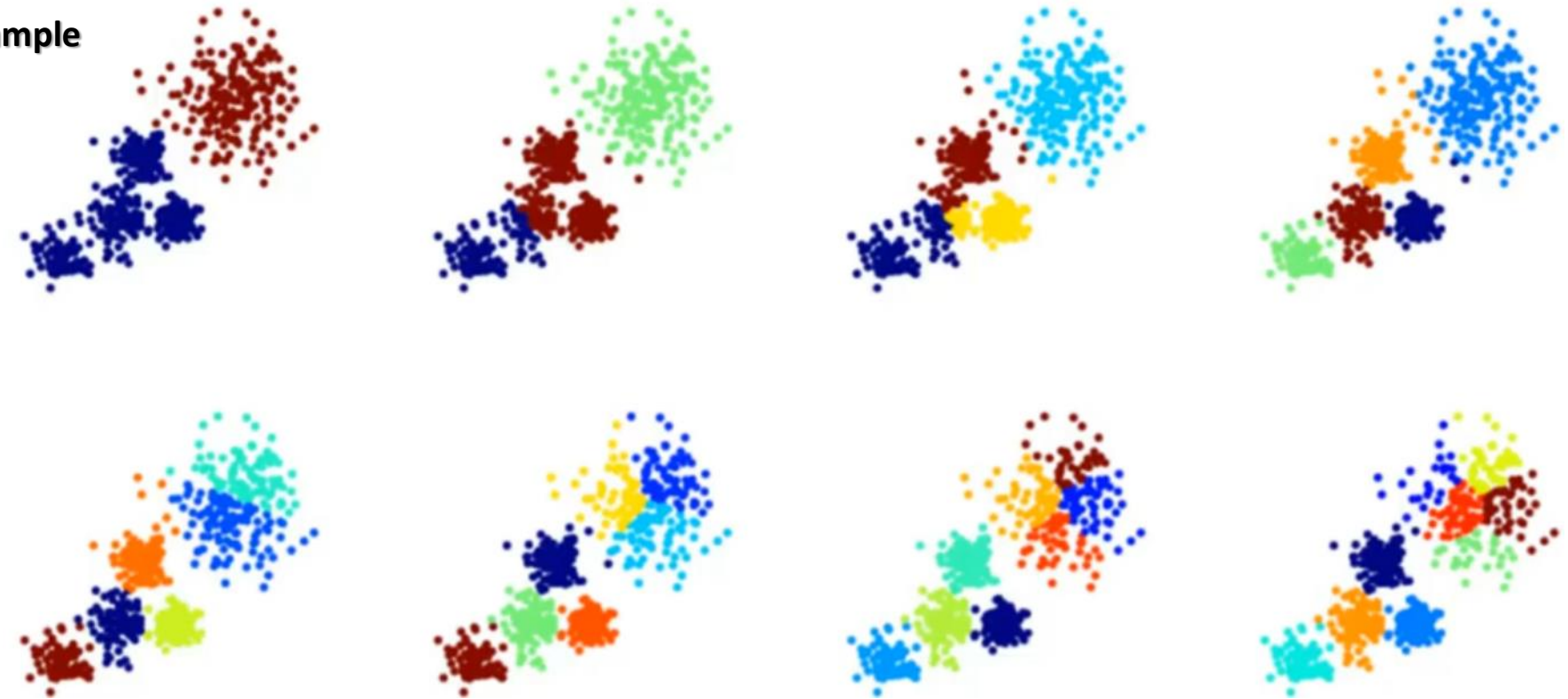
K-Means Clustering

- How it works?

  1) Randomly pick K centroids (k-means)
  2) Assign each data point to the centroid it's closest to
  3) Recompute the centroids based on the average position of each centroid's points
  4) Iterate until points stop changing assignment to centroids
  5) To predict the cluster for new points, find the centroid they're closest to

Example

Example

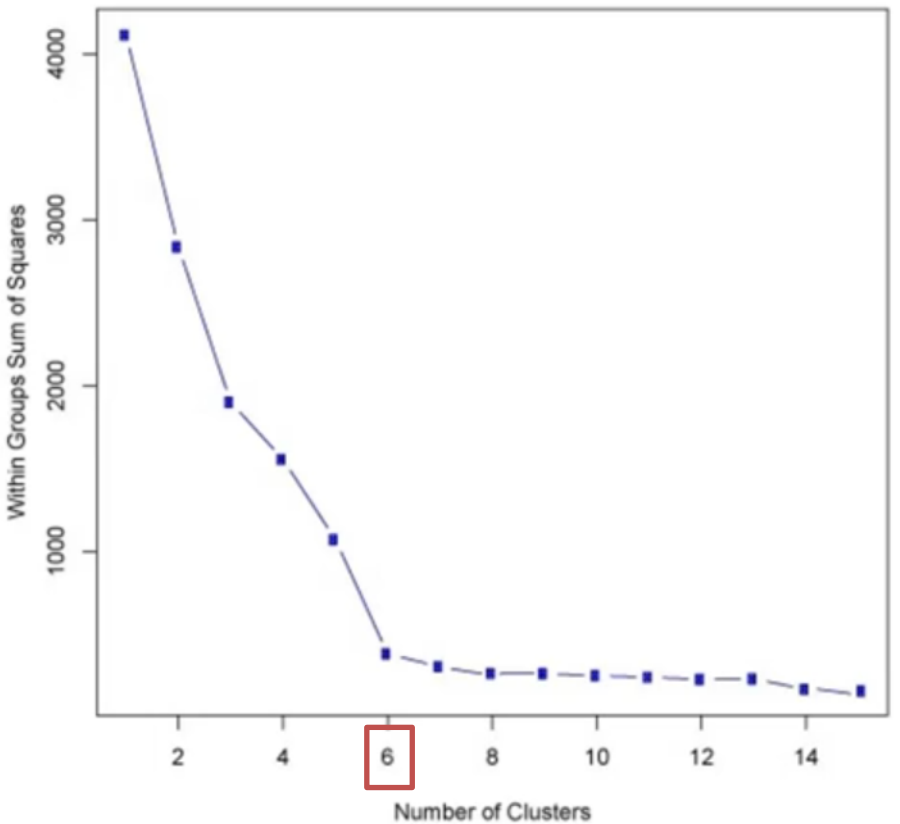


K-Means Clustering

- Choosing a K Value
  - To choose the “best” K value is no easy answer
  - Elbow method provides a way to choose K at which the sum of squared error (SSE) decreases abruptly:

  1) Compute the SSE for some values of K (e.g., 2,4,6,8,etc.)
  2) SSE is defined as the sum of the squared distance between each member of the cluster and its centroid.
  3) By ploting K agaisnt SSE, the error decreases as K gets larger – when the clusters increase, they should be smaller, so distortion is also smaller

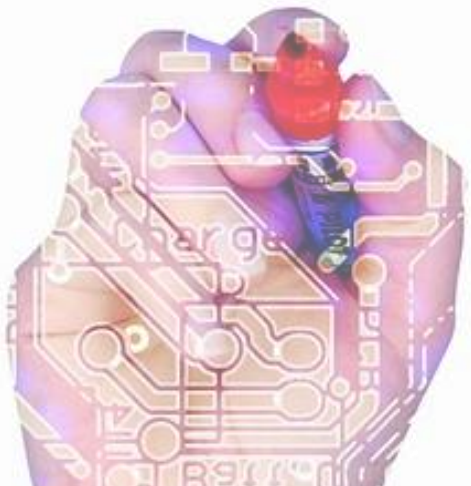
Example – Elbow Effect

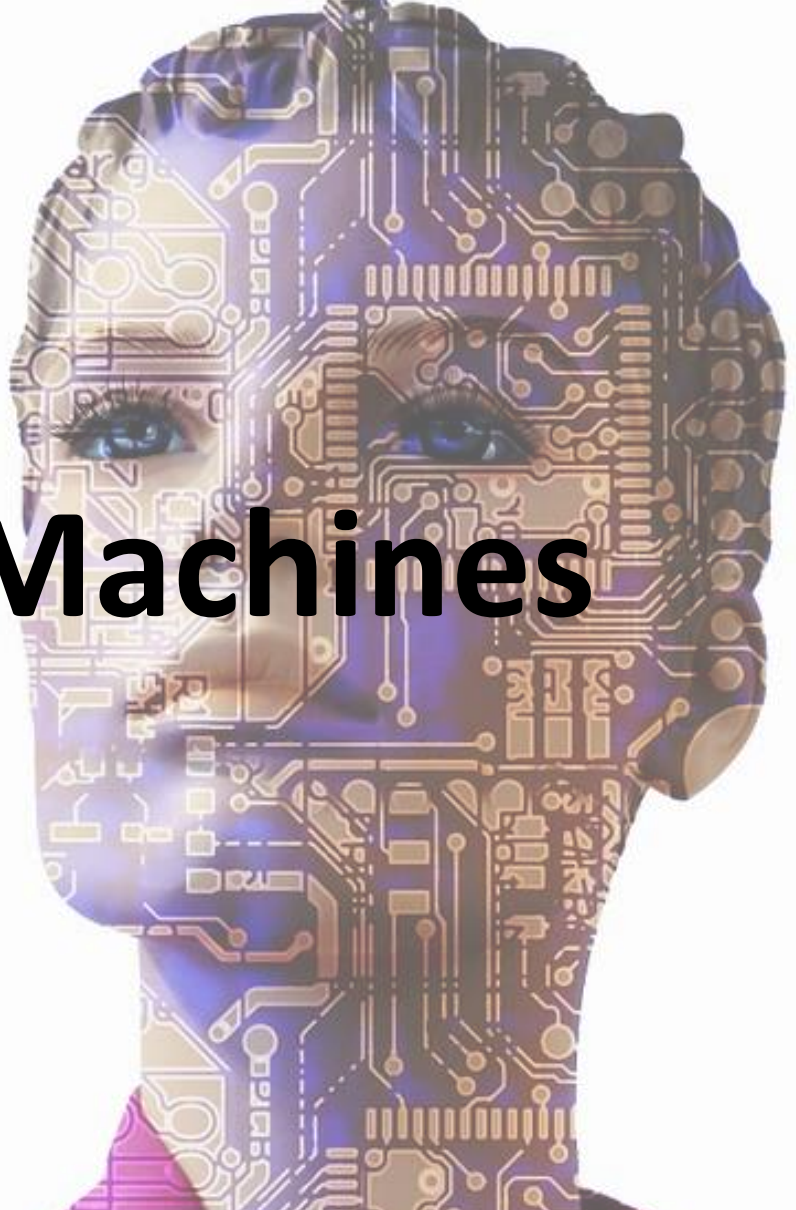


K-Means Clustering – Take in mind!

- Choosing a K Value
  - Increase K values until you stop getting large reductions in squared error – distances from each point to their centroids
- Avoid local mínima
  - Random choice of initial centroids can yield different results
  - Run it a few times to make sure your initial results don't variate too much
- Labeling the clusters
  - K-Means does not attempt to assign any meaning to the clusters it finds
  - It's up to the data scientist to dig into the data and try to determine that

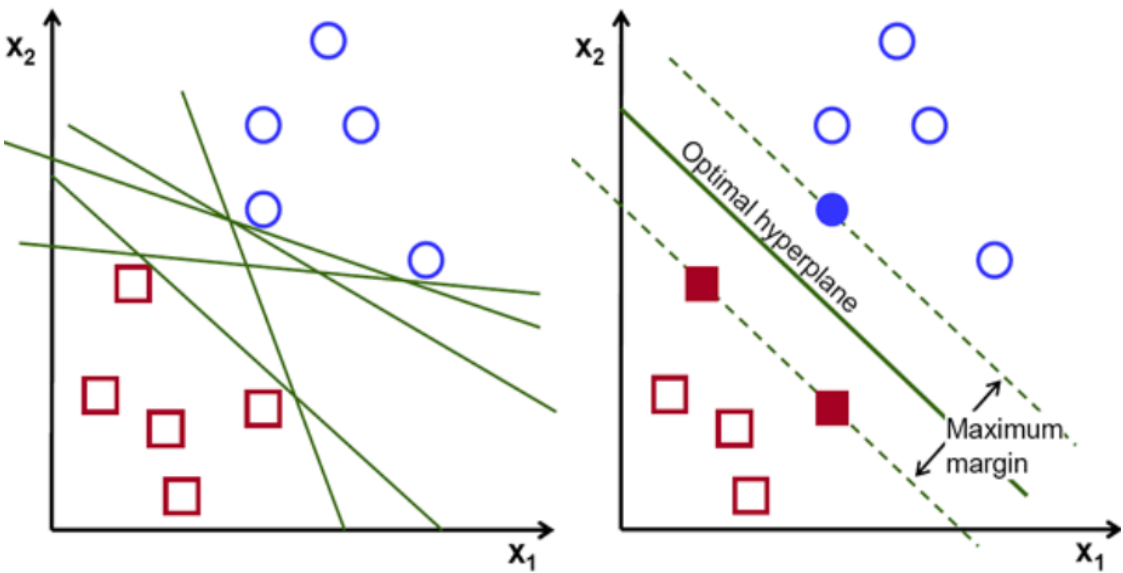
Support Vector Machines





Support Vector Machines (SVM)

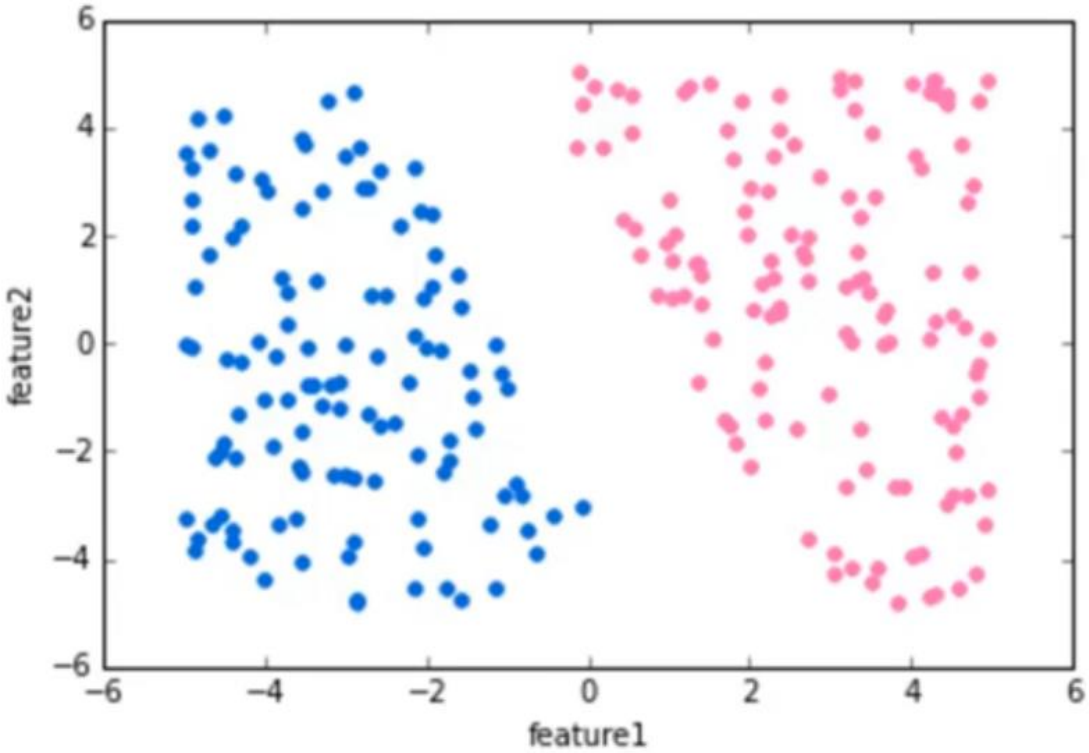
- Works well for classifying higher-dimensional data (datasets with lots of features)
- SVM - supervised learning model
- Finds higher-dimensional support vectors which “divides” the data
- Applys kernels to represent data in higher-dimensional spaces to find hyperplanes that might not be apparent in lower dimensions



Support Vector Machines (SVM)

- How it works?

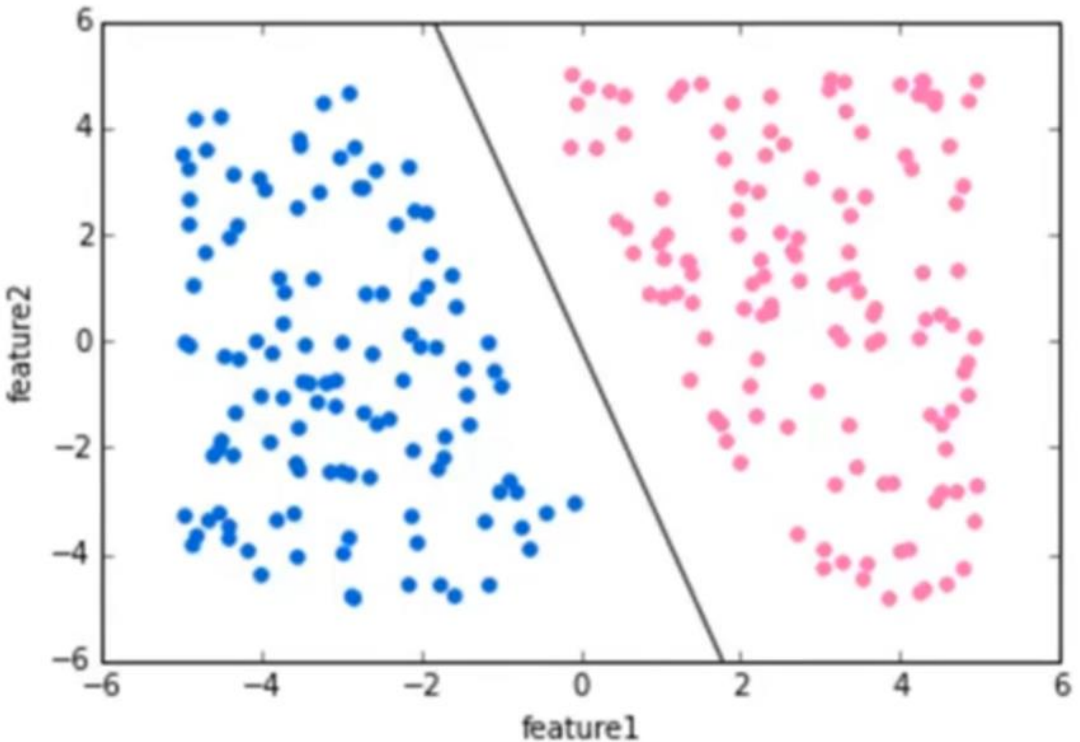
1) Image a labeled training data



Support Vector Machines (SVM)

- How it works?

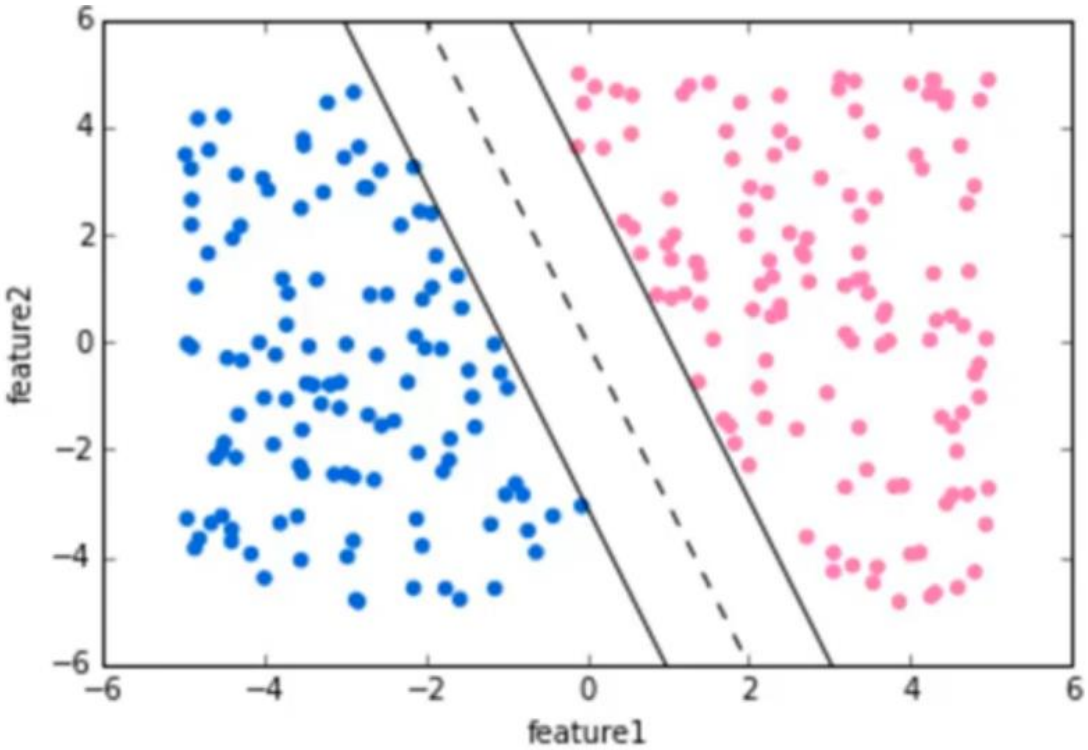
1) Image a labeled training data
2) Draw a separating “hyperplane” between the classes



Support Vector Machines (SVM)

- How it works?

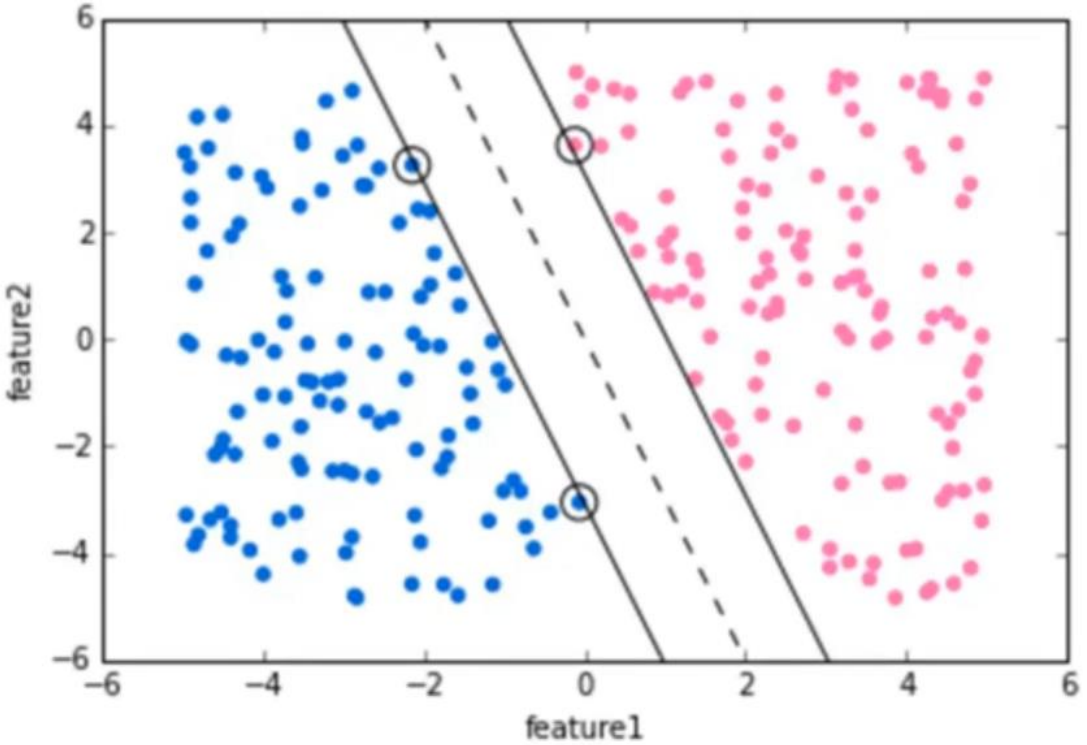
1) Image a labeled training data
2) Draw a separating “hyperplane” between the classes – many options that separate perfectly…
3) Choose a hyperplane that maximizes the margin between classes



Support Vector Machines (SVM)

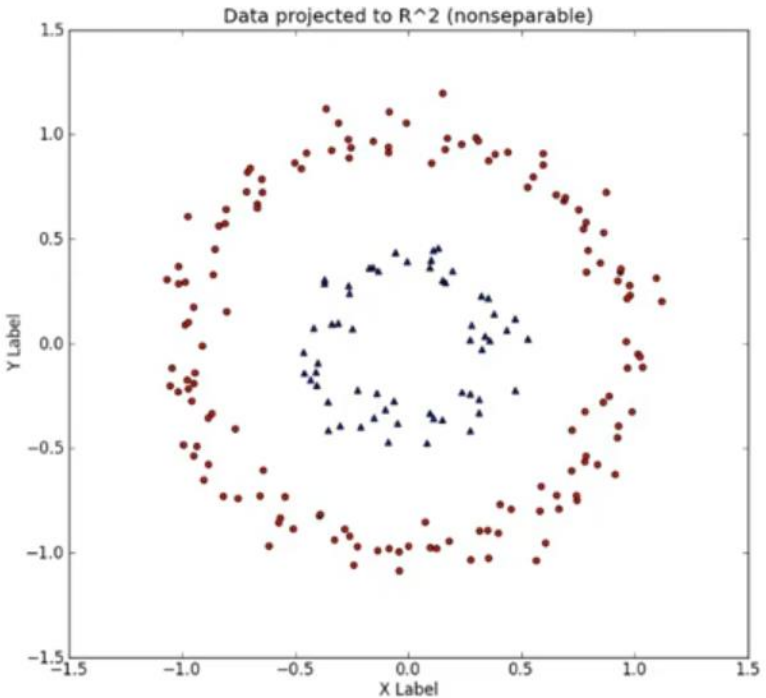
- How it works?

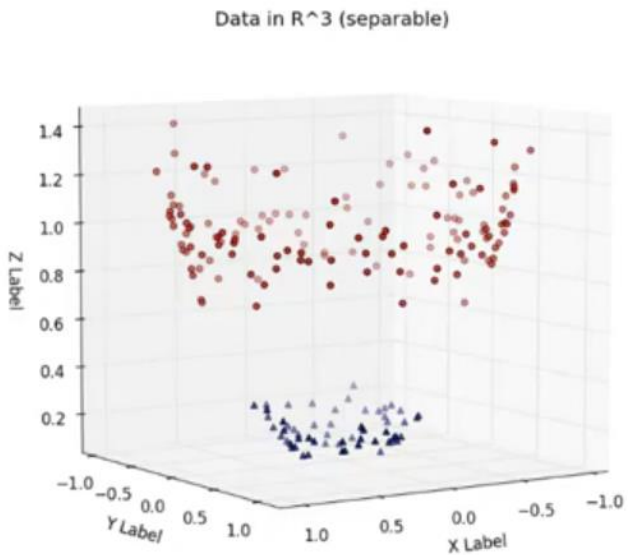
1) Image a labeled training data
2) Draw a separating “hyperplane” between the classes – many options that separate perfectly…
3) Choose a hyperplane that maximizes the margin between classes – vector points that the margin lines touch are known as Support Vectors



Support Vector Machines

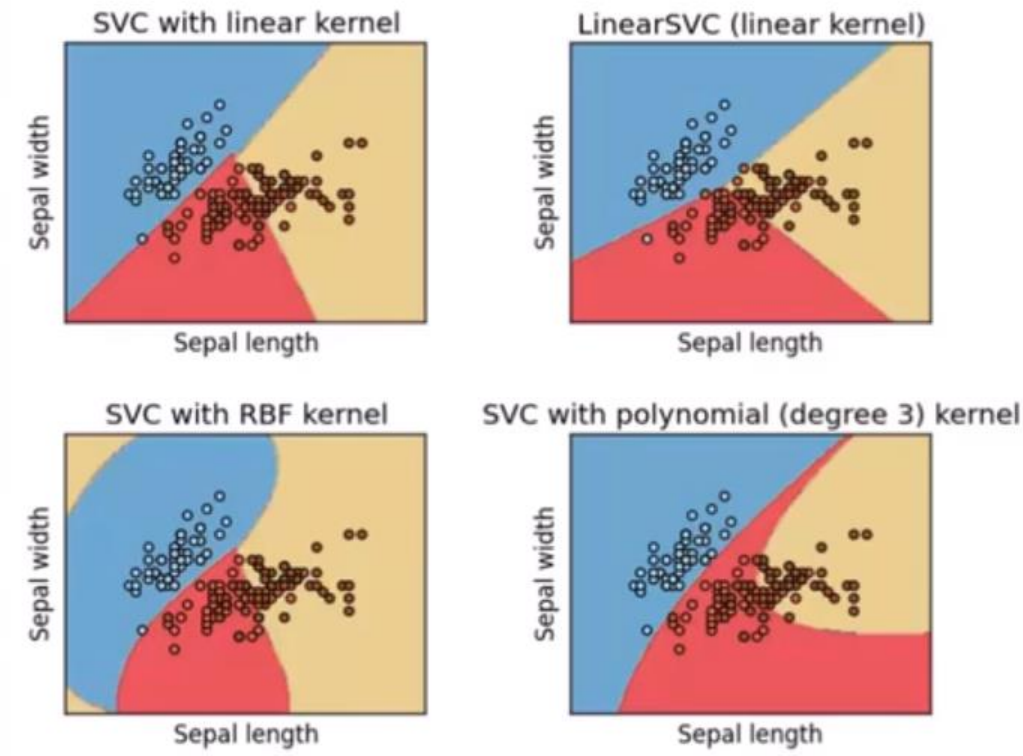
- The idea can be expanded to non-linearly separable data through the “kernel trick”





Support Vector Classification

- Use of SVC to classify data using SVM
- Apply different “kernels” with SVC
- Different kernels provide different results for a given dataset



Universidade do Minho
Escola de Engenharia
Departamento de Informática

Mestrado Integrado em Engenharia Informática
Mestrado em Engenharia Informática
Aprendizagem e Extração de Conhecimento
2020/2021

Filipe Gonçalves, César Analide, Paulo Novais