CDSF13, CDSF14, CDSF15 and CDSF16

Recap V

DS Academy

TOTVS

2018

Evaluation
Metrics

Mean Absolute Error:

- average of the difference between the *original values* and the *predicted values.*

Mean Squared Error:

- average of the **square** of the difference between the *original values* and the *predicted values*.

$$MeanSquaredError = \frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2$$

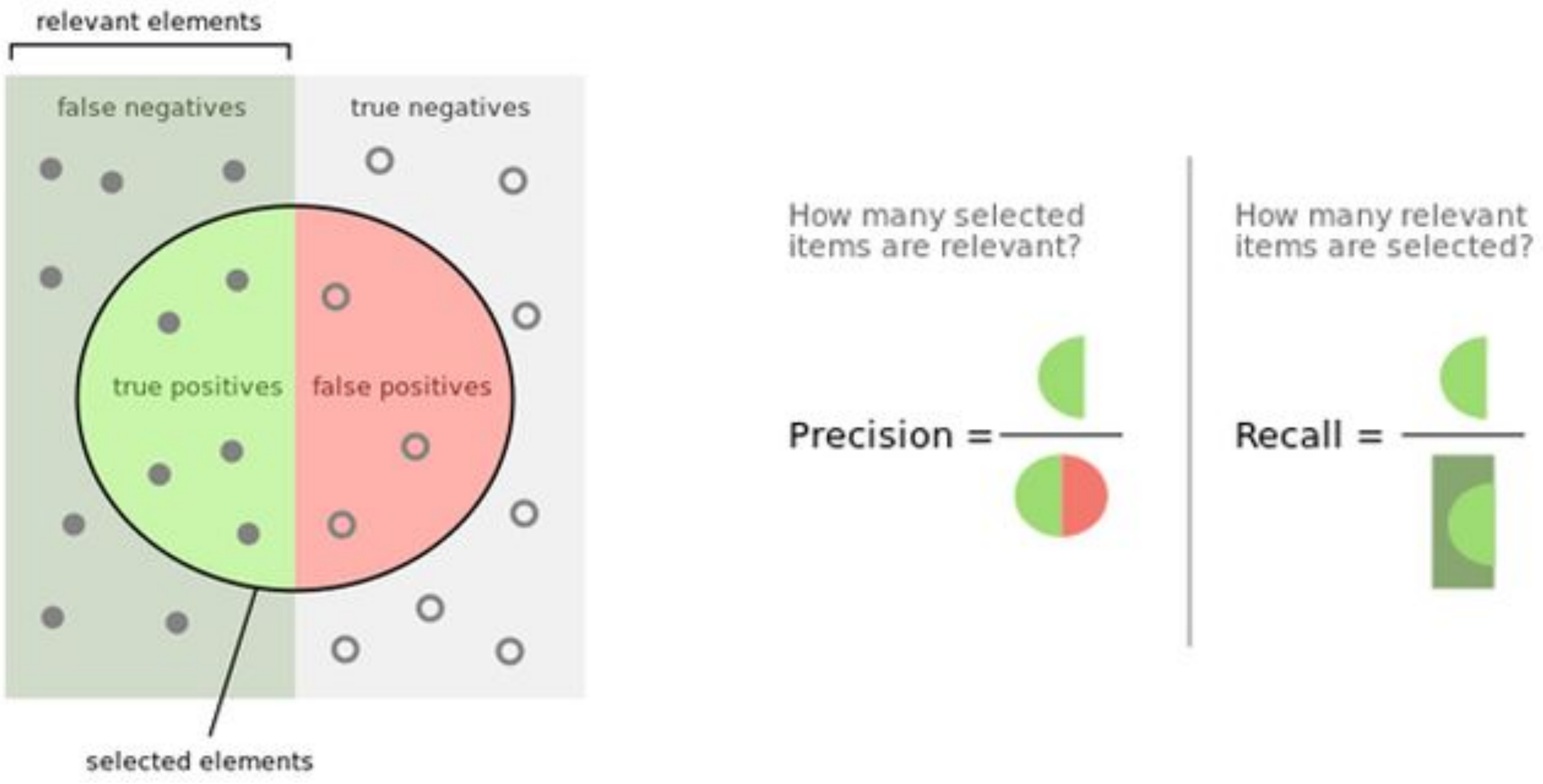
Evaluation Metrics - Classification

This is a list of rates that are often computed from a confusion matrix for a binary classifier:

- **Accuracy:** Overall, how often is the classifier correct?
  - (TP+TN)/total = (100+50)/165 = 0.91
- **True Positive Rate:** When it's actually yes, how often does it predict yes?
  - TP/actual yes = 100/105 = 0.95
  - also known as "Sensitivity" or "Recall"
- **False Positive Rate:** When it's actually no, how often does it predict yes?
  - FP/actual no = 10/60 = 0.17
- **True Negative Rate:** When it's actually no, how often does it predict no?
  - TN/actual no = 50/60 = 0.83
  - equivalent to 1 minus False Positive Rate
  - also known as "Specificity"
- **Precision:** When it predicts yes, how often is it correct?
  - TP/predicted yes = 100/110 = 0.91
- **F1 Score:** Harmonic mean of Precision & Recall
  - F1 = (2*Precision*Recall)/(Precision + Recall) = 0.93

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Evaluation Metrics - Classification



The output of the model is a **probability** (most of the time):

- Values between 0 and 1 for each sample

What is the best **threshold**?

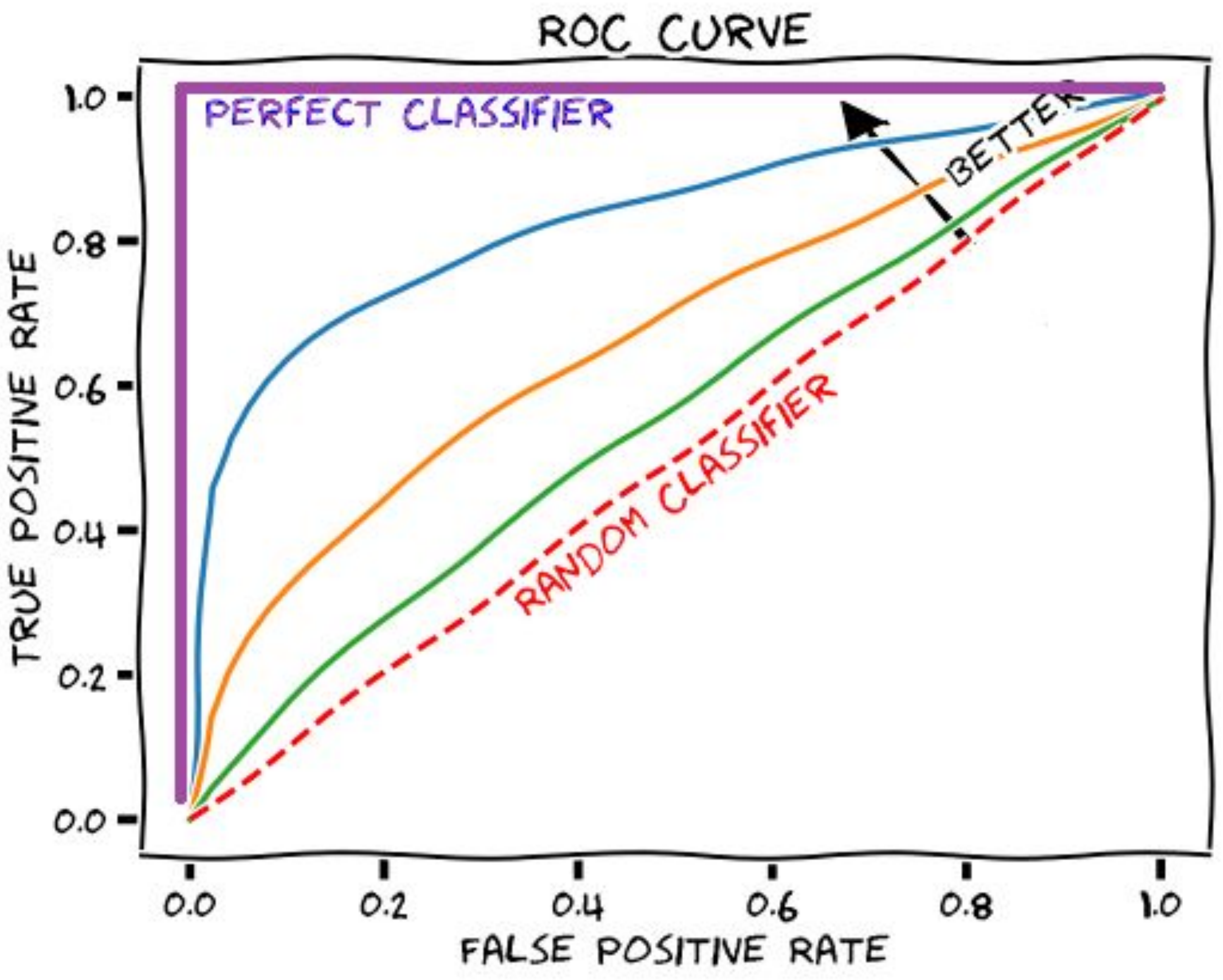
- 0.5, 0.1, 0.9?

It depends!

- How much does it **cost** for a FP or a FN?
- The model is not perfect there will always be a **tradeoff**.

ROC Curve (Receiver Operating Characteristic)

- It is a plot of the true positive rate versus the false positive rate for the predictions of a model for multiple thresholds between 0.0 and 1.0.
- It starts at the lower left-hand corner
  i.e. the point (FPR = 0, TPR = 0)
  - decision threshold of **1**
  - Every example is classified as **negative**
- It ends at the upper right-hand corner
  i.e. the point (FPR = 1, TPR = 1)
  - decision threshold of **0**
  - every example is classified as **positive**

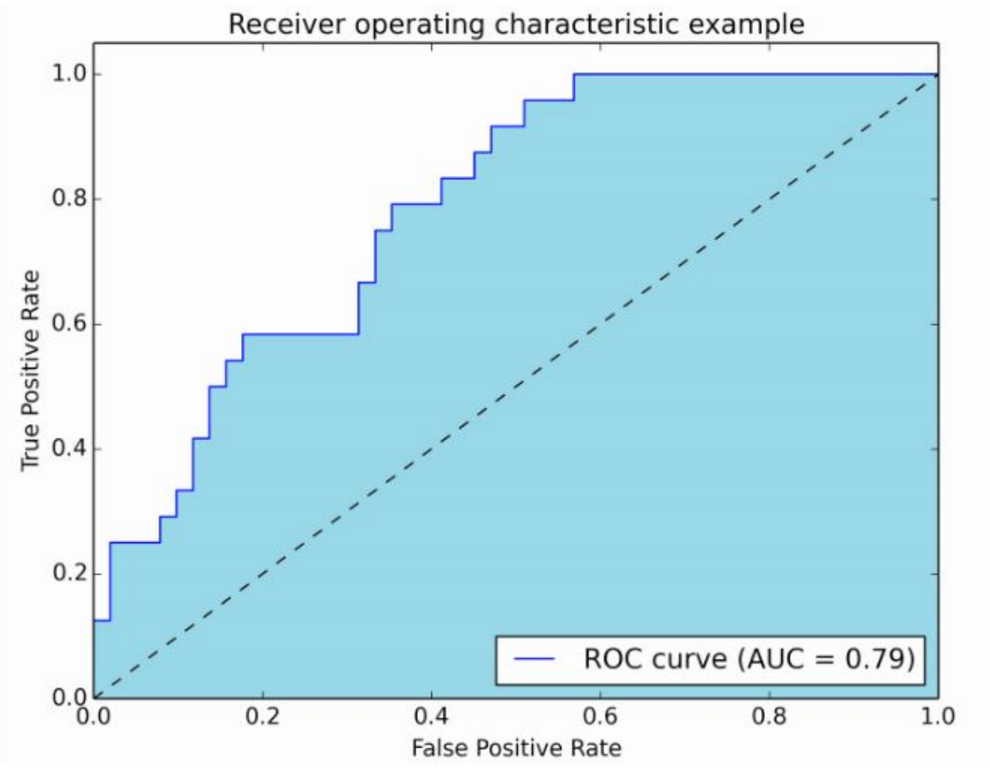


Demonstration
http://www.navan.name/roc/

AUC (area under the ROC curve)

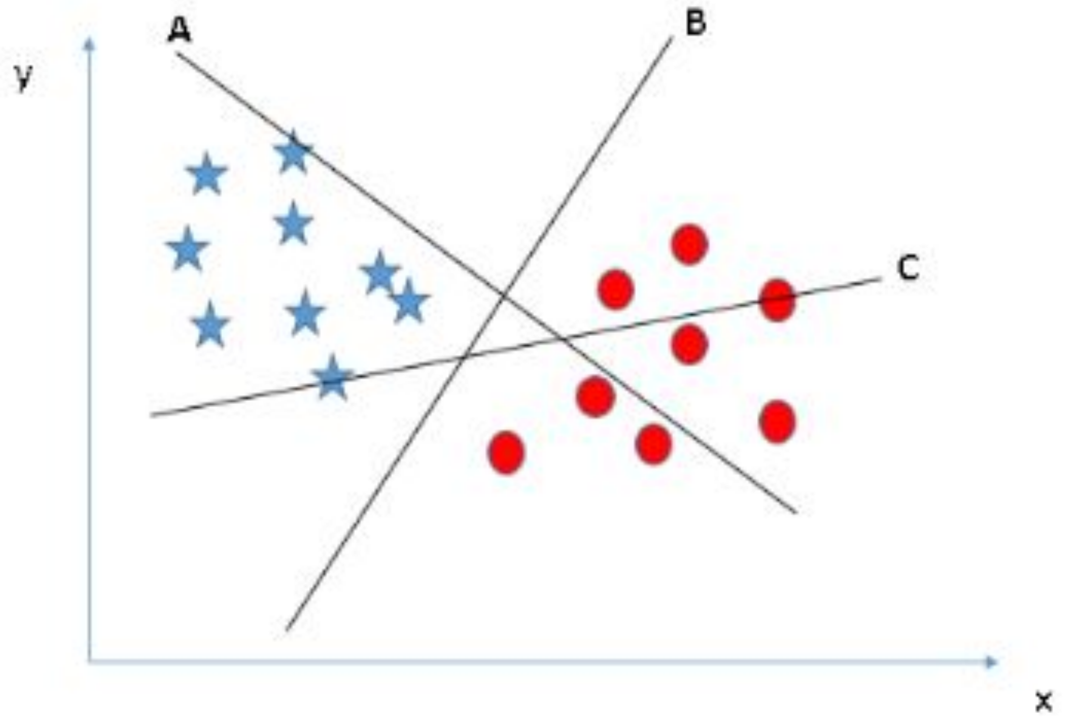
- it is the probability a randomly-chosen positive example is ranked more highly than a randomly-chosen negative example
- AUC is more informative than accuracy for imbalanced data, but it can be "excessively optimistic" about the performance of models for datasets with a much larger number of negative examples than positive examples

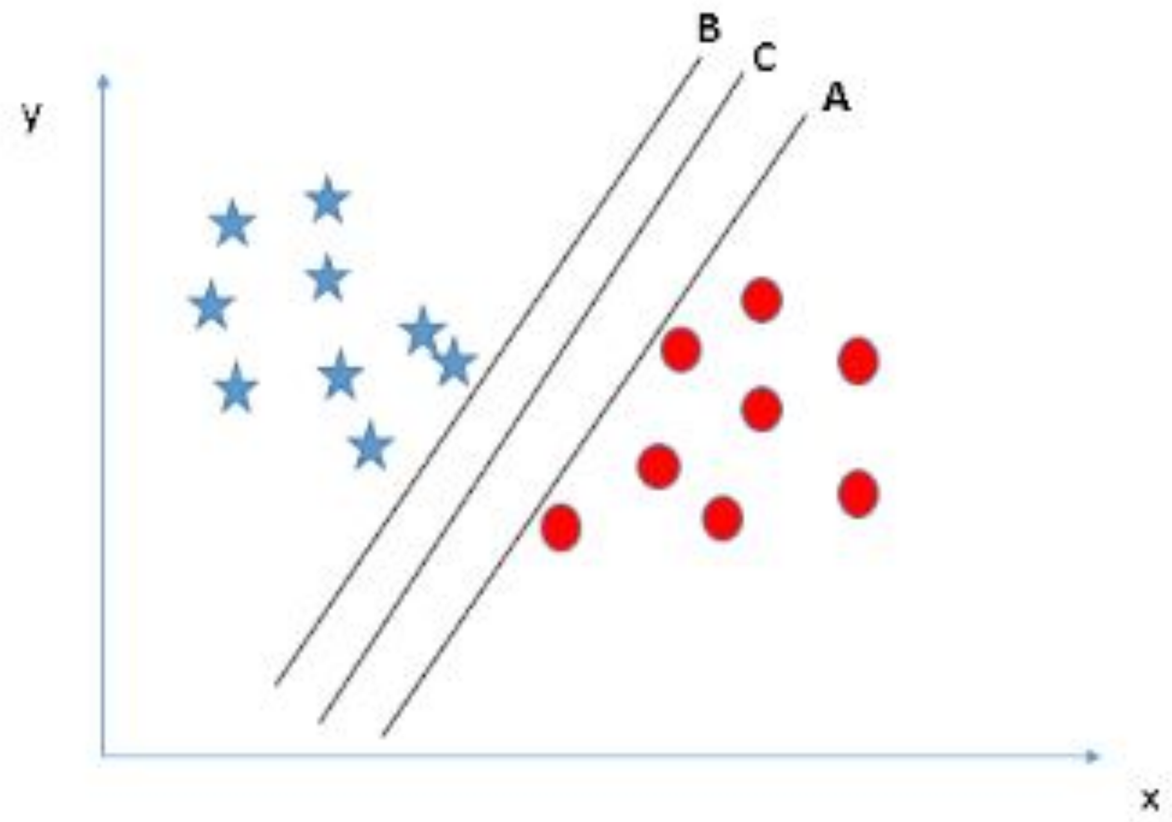
$$FPR = \frac{FP}{(FP + TN)}$$



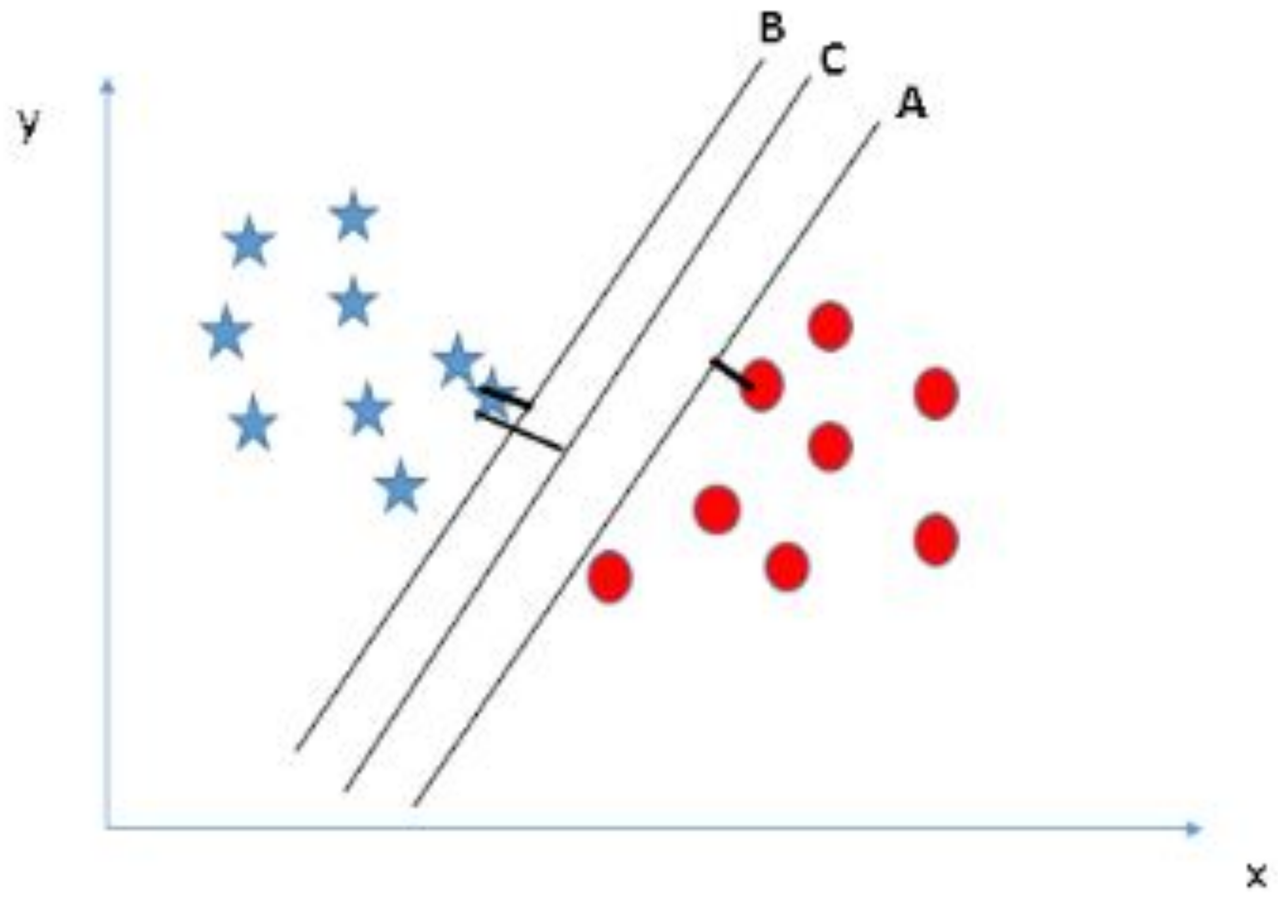
SVM

The objective of the **support vector machine** is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.





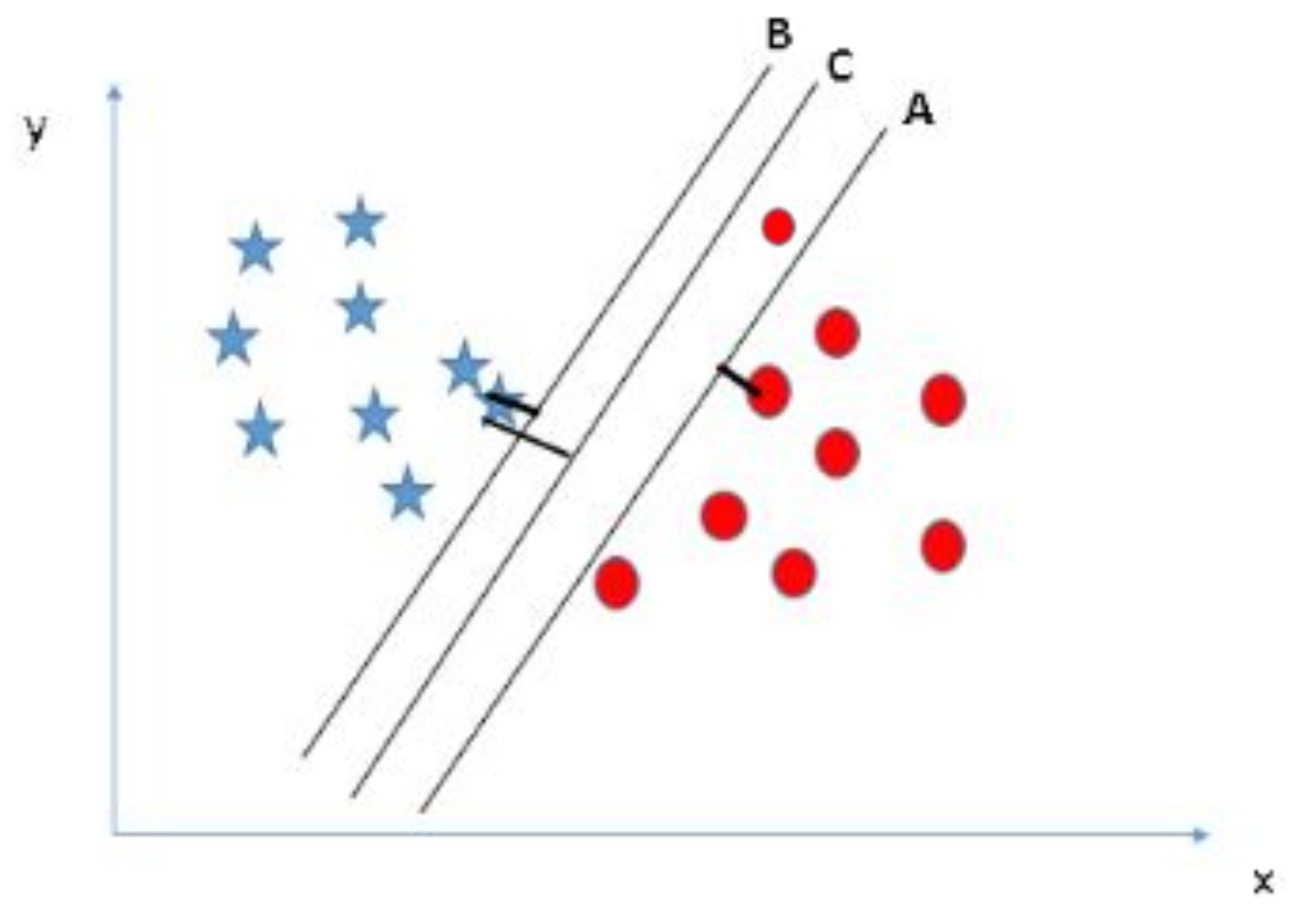
- The goal is to maximize the distances between nearest data point and hyperplane.
- This distance is called as **Margin**.
- If that output is greater than 1, we identify it with one class and if the output is -1, we identify is with another class.



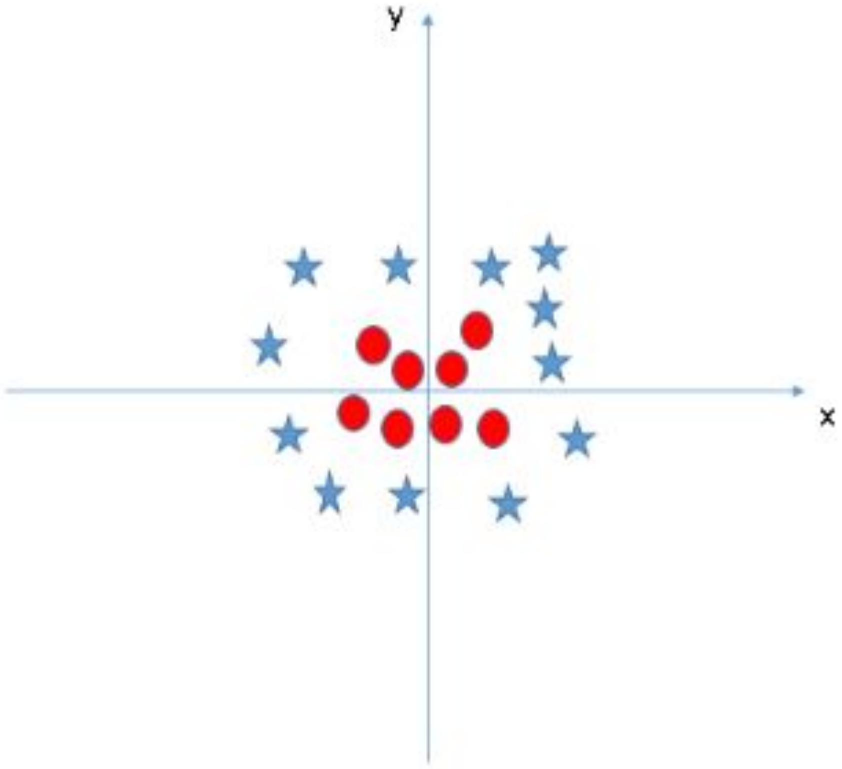
SVM - Support Vector Machine

- The goal is to maximize the distances between nearest data point and hyperplane.
- This distance is called as **Margin**.
- If that output is greater than 1, we identify it with one class and if the output is -1, we identify is with another class.

Another reason for selecting the hyperplane with higher margin is robustness. If we select a hyperplane having low margin then there is high chance of mis-classification.



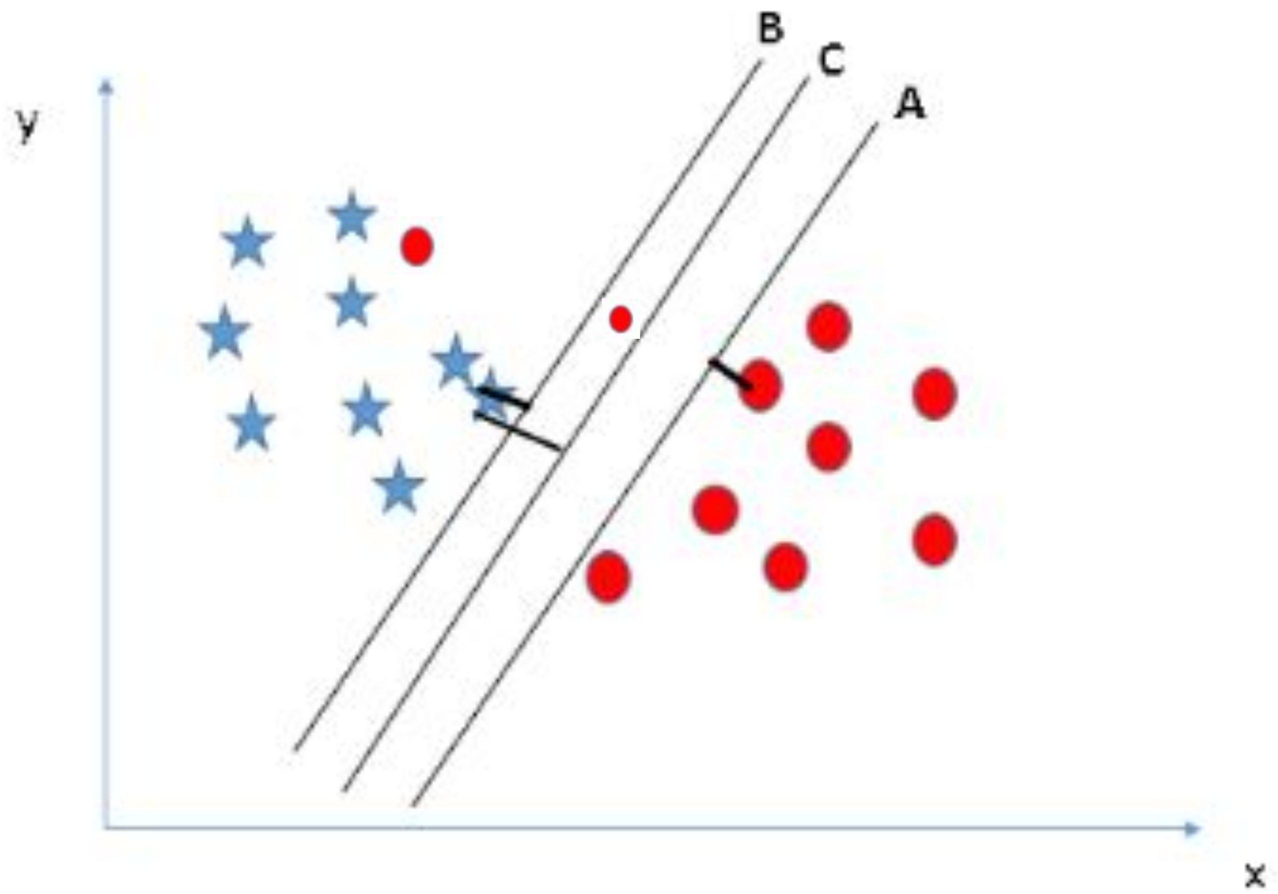
- SVM address non-linearly separable cases by introducing two concepts:
  - **Soft Margin:** try to find a line to separate, but tolerate one or few misclassified values
  - **Kernel Tricks:** try to find a non-linear decision boundary



Soft Margin

Two types of misclassifications are tolerated by SVM under soft margin:

1. The dot is on the wrong side of the decision boundary but on the correct side/ on the margin (shown in left)
2. The dot is on the wrong side of the decision boundary and on the wrong side of the margin (shown in right)

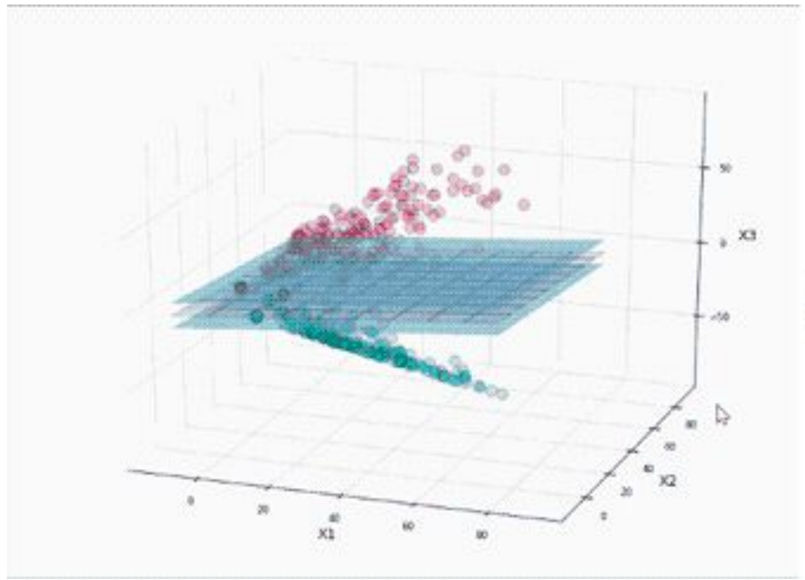


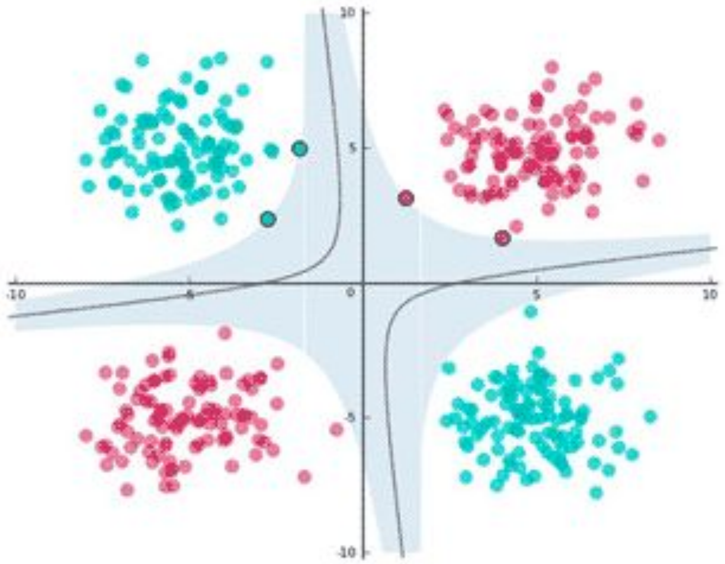
Degree of tolerance

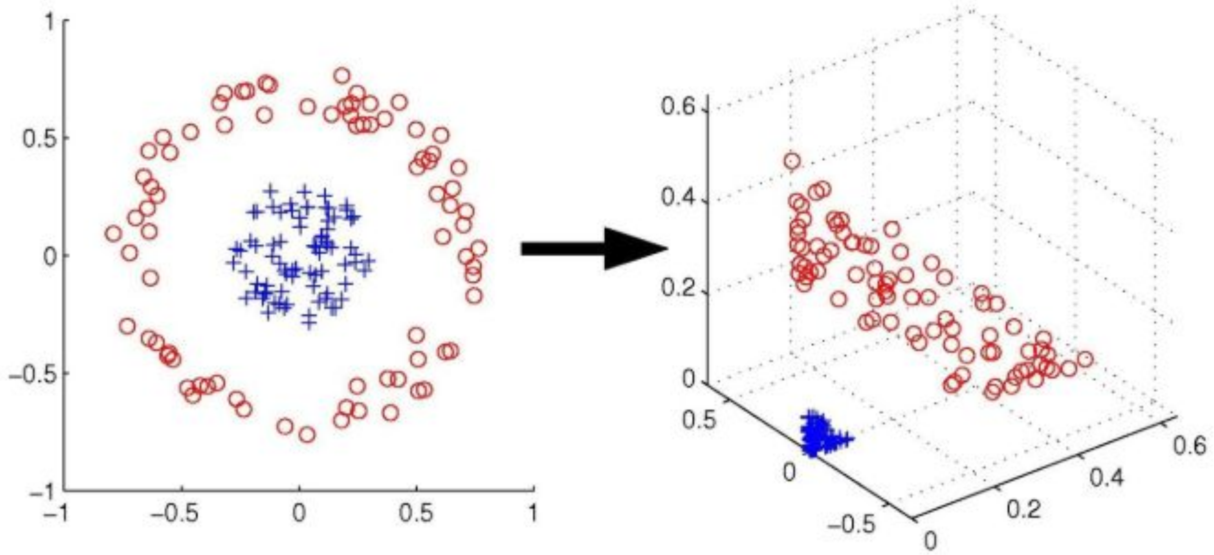
How much tolerance we want to give when finding the decision boundary can be found using cross validation.

Kernel Trick

- We use kernel functions in this case that help transform the data into another dimension that has a clear dividing margin between the two classes.
- Kernel functions help transform non-linear spaces into linear spaces.
  - What kernel should I use?
  - It adds new hyperparameters to the problem







Pros:

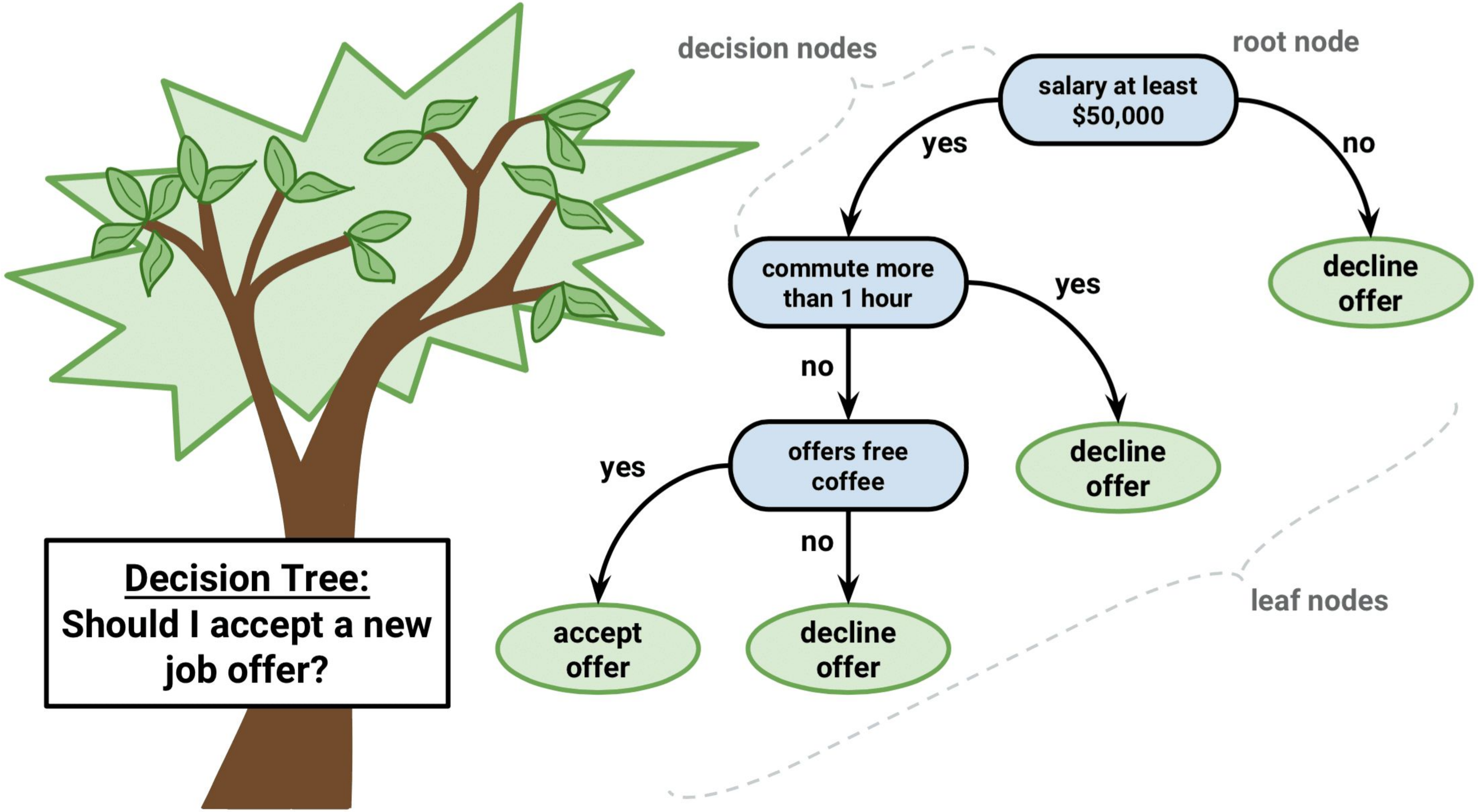
- It is useful for both linearly Separable (hard margin) and Non-linearly Separable (soft margin) data.
- It is effective in high dimensional spaces.
- It is effective in cases where a number of dimensions are greater than the number of samples.
- It uses a subset of training points in the decision function, so it is also memory efficient.

Cons:

- Picking the right kernel and parameters can be computationally intensive.
- It also doesn’t perform very well, when the data set has more noise
- SVM doesn’t directly provide probability estimates.

Decision
Trees

Categorical vs. Regression

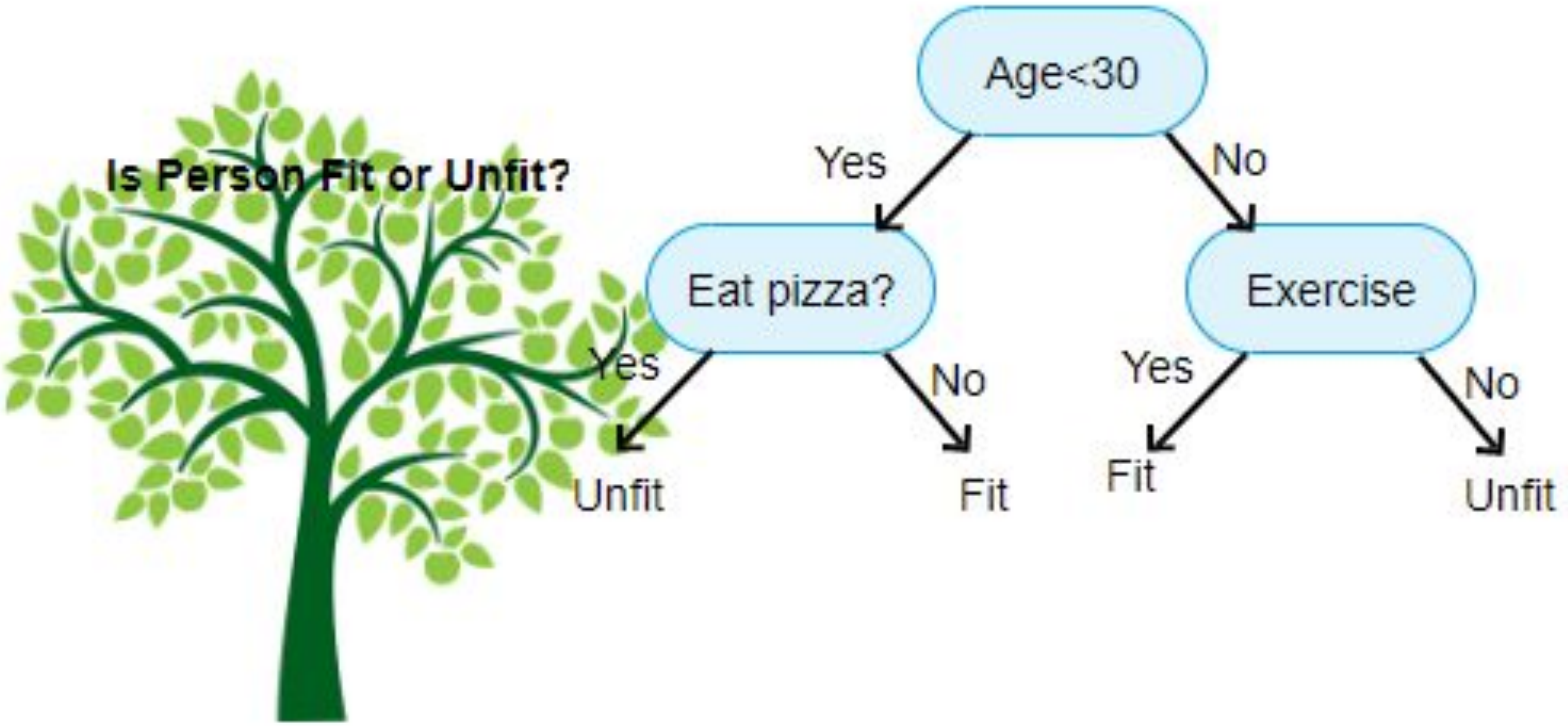


Example of a categorical decision tree

DATA

Age	Eat Pizza	Exercise	Fit
15	No	Yes	Yes
25	Yes	No	No
58	Yes	Yes	Yes
35	No	No	No

How can we define the **best** attribute to split the data?



- Node's purity
  - Gini
  - Entropy
  - Chi-Squared
  - Reduction in Variance

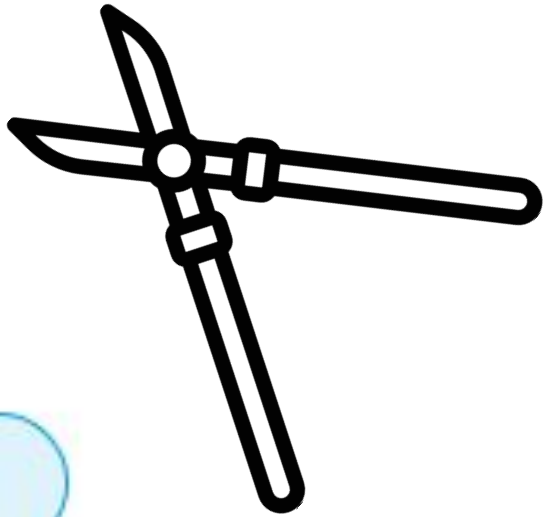
Decision Tree

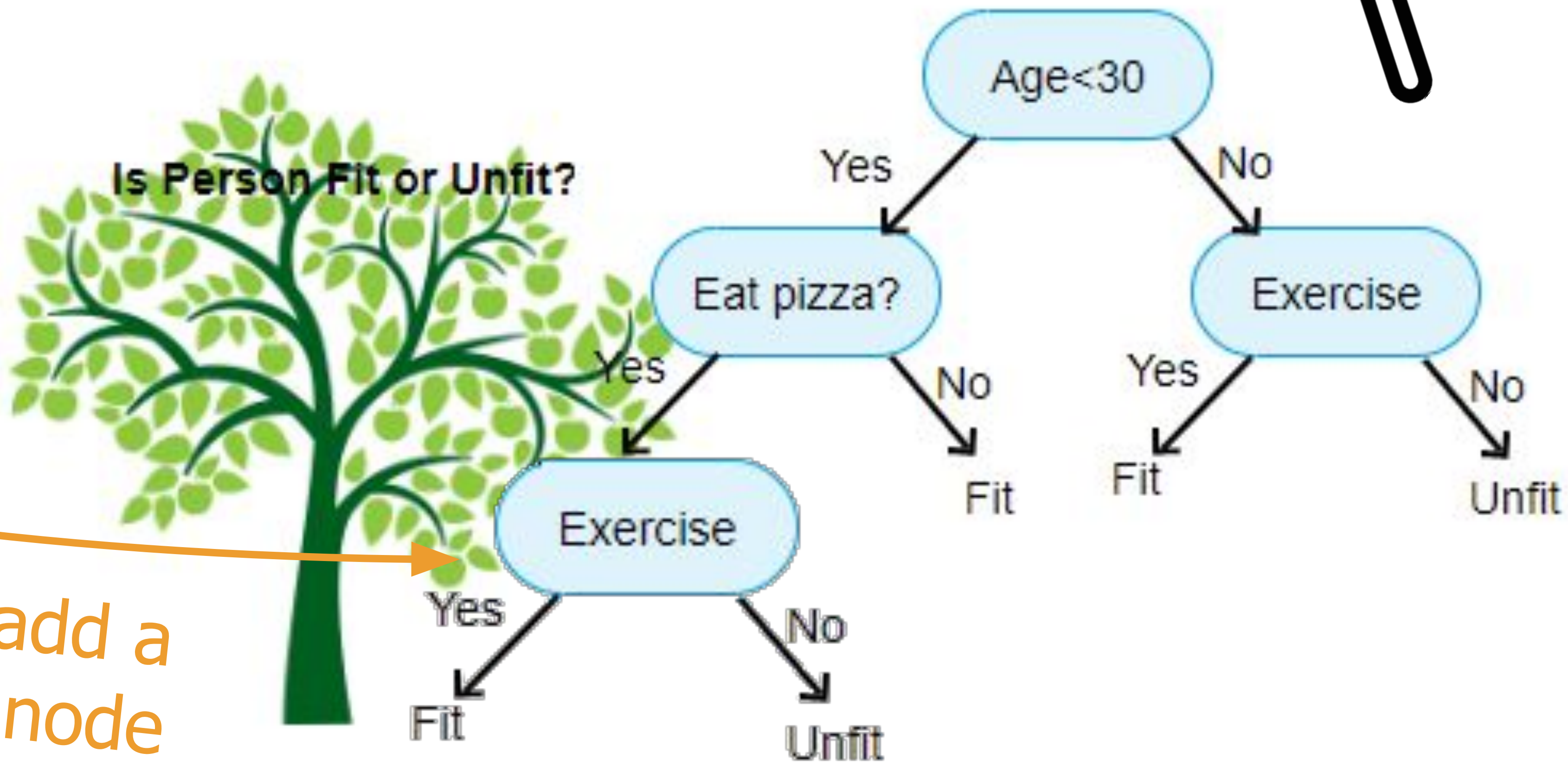
DATA

Age	Eat Pizza	Exercise	Fit
15	No	Yes	Yes
25	Yes	No	No
58	Yes	Yes	Yes
36	No	No	No
29	Yes	Yes	Yes

New data

Without constraints, a tree can overfit by creating leaves for each possibility.





can add a new node

Decision Tree

PROS

- Less effort for data preparation
  - No need for normalization
  - Missing values does not affect
- Can handle numerical and categorical
- Interpretable and easy to explain
- Fast during test time

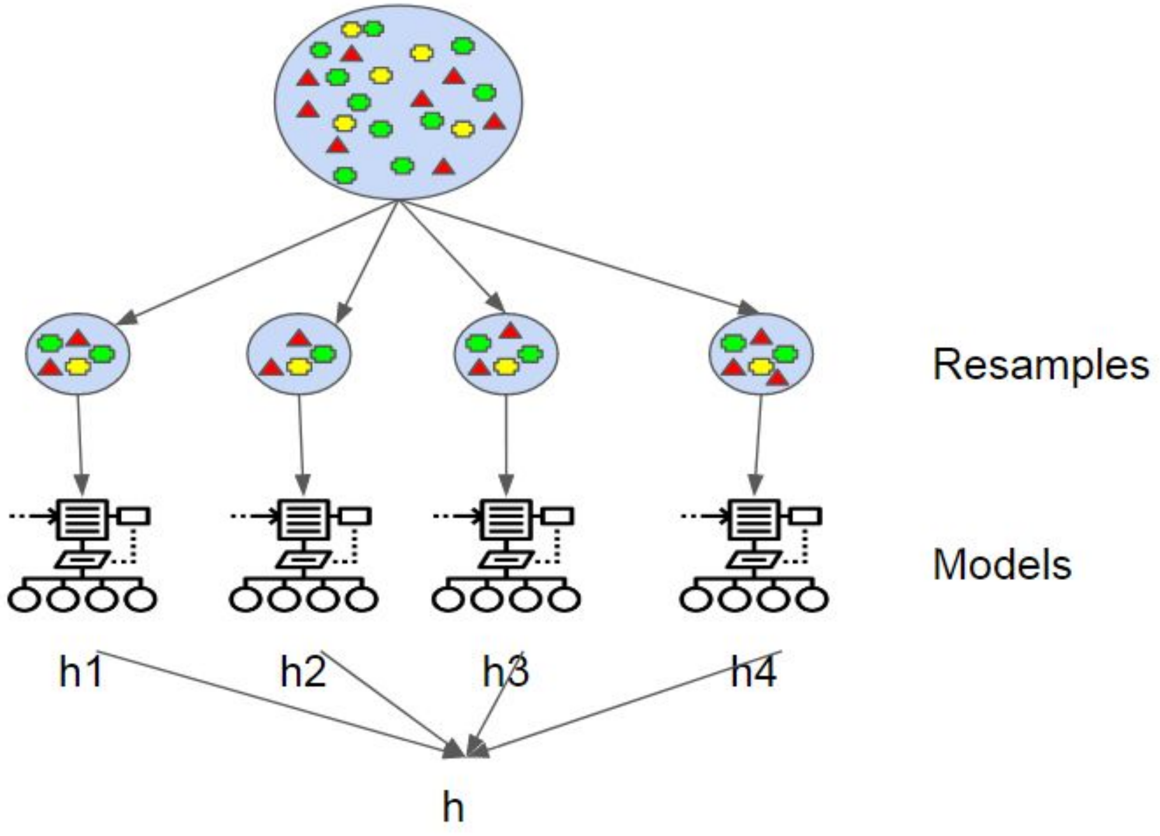
CONS

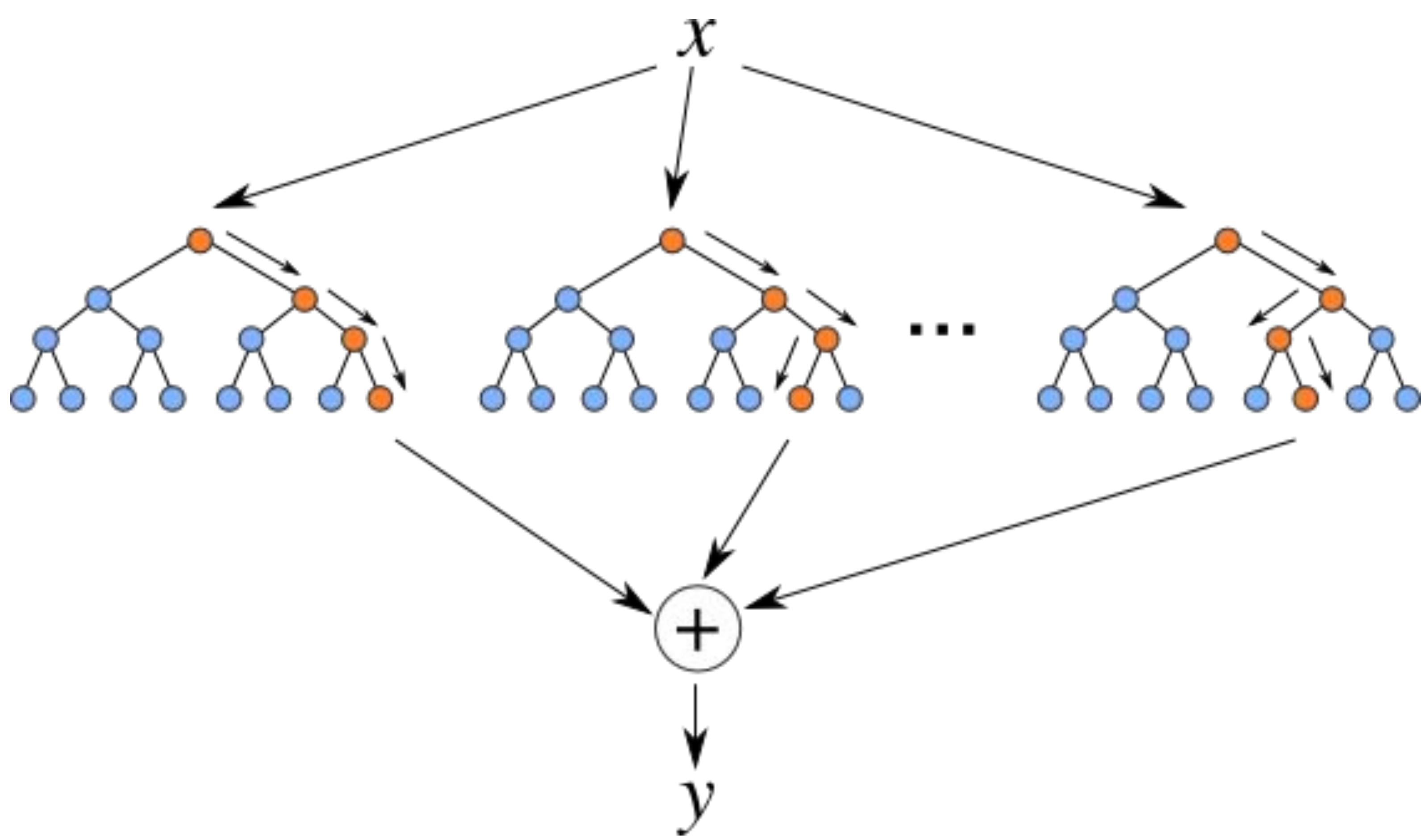
- Greedy (may not find best tree)
- Overfits
- Regression
- Training time

How to reduce error due to variance and bias?

Ensemble + Bagging

- Random features
- Random samples





Useful Resources

Useful Resources

ROC CURVE
https://www.youtube.com/watch?time_continue=737&v=OAl6eAyP-yo
http://www.navan.name/roc/

Kernel list SVM
http://crsouza.com/2010/03/17/kernel-functions-for-machine-learning-applications/

Decision Trees
https://www.youtube.com/watch?v=7VeUPuFGJHk
Random Forest
https://www.youtube.com/watch?v=J4Wdy0Wc_xQ

THANK YOU

Tecnologia + Conhecimento são nosso DNA.

O sucesso do cliente é o nosso sucesso.

Valorizamos gente boa que é boa gente.

#SOMOSTOTVERS

totvs.com
company/totvs
@totvs
fluig.com