early
makers

em
lyon
business
school

Machine Learning

3. Classification

Nicolas Gartner

Classification problems

Binary classification :

- Only two possible choices of labels.

  Examples: A tumor is malignant or not, a student passes or not, a cat is on the picture or not, etc.

- Logistic regression is a very common method used

Multi-class classification :

- Multiple choices of labels.

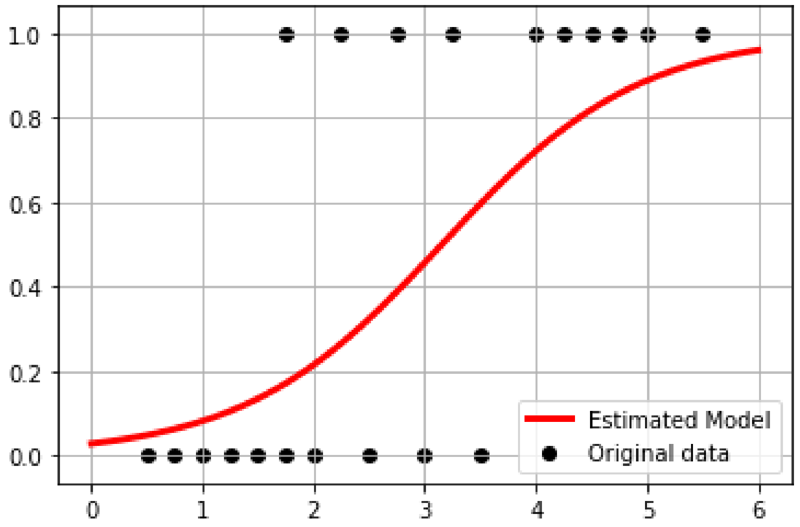
  Examples: divide pictures depending on the animals on them, classify patients according to their supposed disease, classify the nearby elements from cameras to help drones know what is around, etc.

- Decision Trees or Random forests are very common methods for this type of problem

In this class: 4 methods

- Logistic regression
- K-Nearest Neighbours (k-NN)
- Support Vector Machine (SVM)
- Decision Tree

Logistic regression



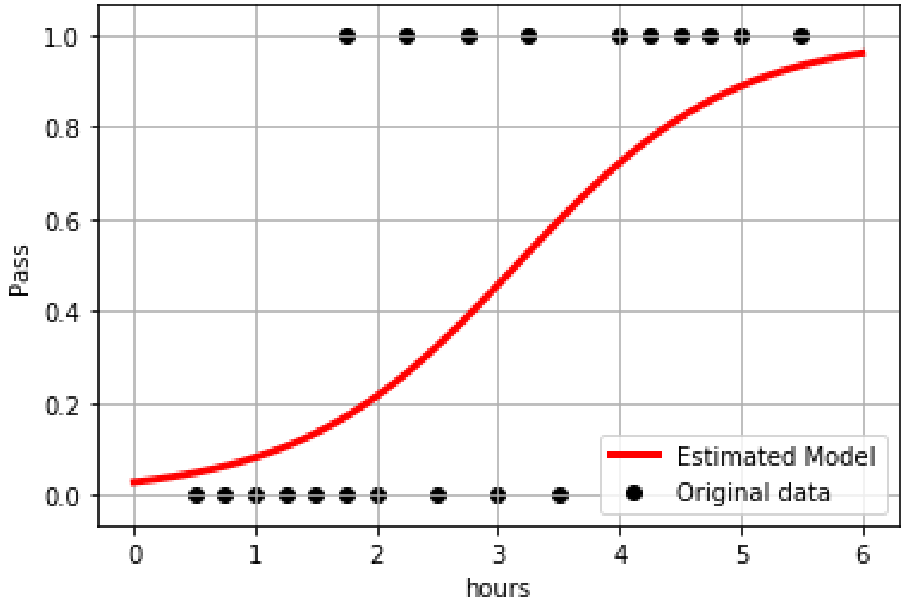
- Is a linear classifier, which is the counterpart of linear regression.
  - A continuous function describes the model
  - Classification decision based on the value of a linear combination of the characteristics
- Is a binary classifier
  - Either 1 or 0, True or False, Good or bad, class A and class B, are predicted by that model
- So called because of the logit function
  - $logit(p) = \log\left(\frac{p}{1-p}\right),$ p the probability of y to be 1
  - Used inside the logistic regression model for the probability analysis of the event happening or not
- The function (in red) that is displayed here is however
  - $f(x) = \frac{1}{1-e^{-(\beta_0+\beta_1 x)}}$
  - With β_0 and β_1 the parameters determined with logistic regression
- Value given by that function gives you the likelihood of having a 1 value.

Logistic regression: example

20 students spend between 0 and 6 hours studying for an exam

Probability of passing the exam ?

Hours	**0.50**	**0.75**	**1.00**	**1.25**	**1.50**	**1.75**	**1.75**	**2.00**	**2.25**	**2.50**	**2.75**	**3.00**	**3.25**	**3.50**	**4.00**	**4.25**	**4.50**	**4.75**	**5.00**	**5.50**
Pass	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1

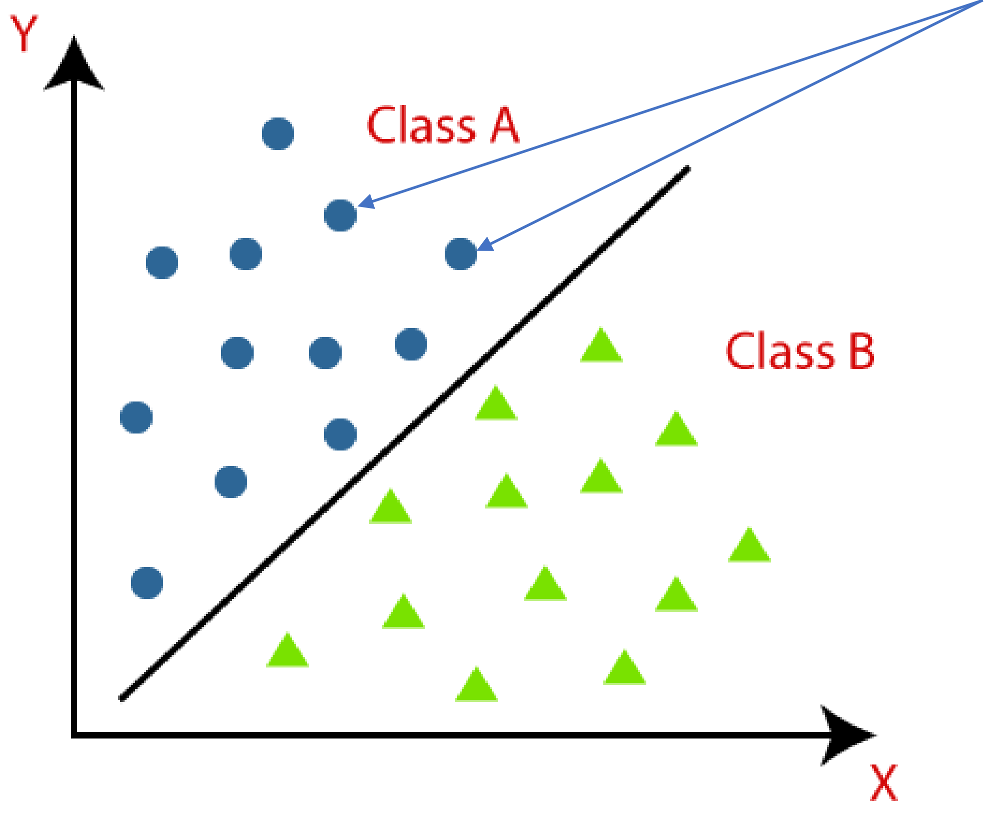


$$f(x) = \frac{1}{1 - e^{-(1{,}12x - 3{,}54)}}$$

- Roughly 3 hours are required to pass the exam with more than 50% chance
- Model predicts pass or fail depending on those probabilities (using .predict from scikit-learn)
- Predictions will not give likelihood, but just 0 or 1

Warning: what all linear classifiers do !

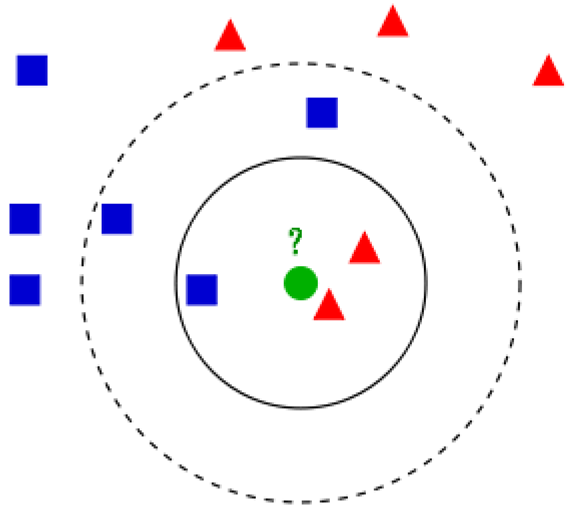
A random linear classifier



- This kind of classifiers draws limits between classes
- Same predictions are given if you are close or far from the limit
- To give accurate results, they often require regularization:
  - Process of adding information in order to solve an ill-posed problem or to prevent overfitting.
  - Modifications are made on data or learning algorithm to reduce its generalization error but not its training error
    - Training error: error made on the computation of the coefficients of the classifier
    - Generalization error: Accuracy measure of an algorithm when predicting outcome values for previously unseen data
- However: Often fastest classification method

K-Nearest Neighbours (k-NN)

- A non-parametric method: no parameter is determined
- Only based on training data
- Consists in finding the majority class of the k closest neighboring data



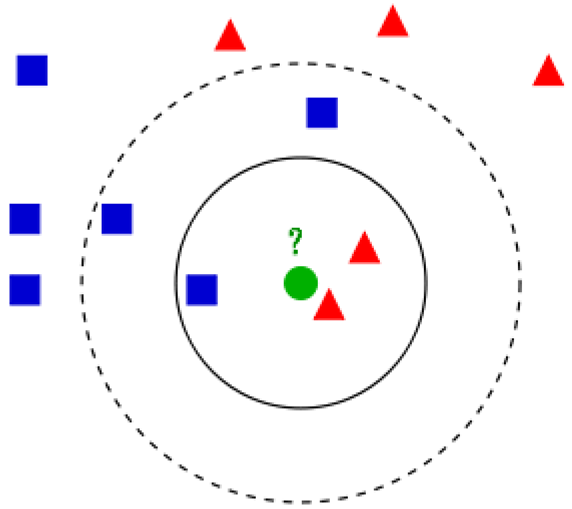
Distance measures:

- For each data you will have different features/metrics that you can plot (2 features -> 2D graph, 3 f. -> 3D graph, …)
- Euclidian distance: length of a line segment between the two points
- Hamming distance: For two strings of equal length, it is the number of positions at which the corresponding symbols are different
  - "**karolin**" and "**kathrin**" is 3.
  - 1011101 and 1001001 is 2.
- Mahalanobis distance: measure of the distance between a point and a distribution

Example:

- Green dot class ?
- Considering k=3 values (red)
- Considering k=5 values (blue)

K-Nearest Neighbours (k-NN)



A good method when:

- You have a lot of data from the different class
- Data from which you want to predict falls inside the domain of your dataset
- Accurate predictions are required

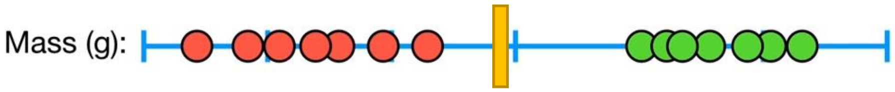
Not a good method when:

- You want to have a human readable model
- You want to get a lot of predictions (every prediction is computationally expensive as no model is stored)

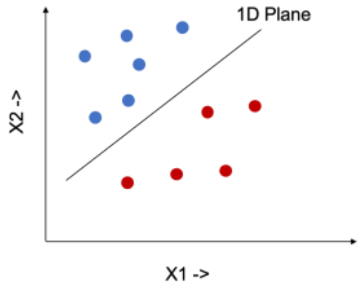
Support Vector Machine (SVM)

- One of the most robust methods for classification:
  - Able to consider misclassification
  - Low variance (standard deviations) in the predictions
- Based on finding a soft-margin or hard-margin hyperplane
  - Hyperplane: A geometrical form that allow to separate classes according to the dimension you are working on

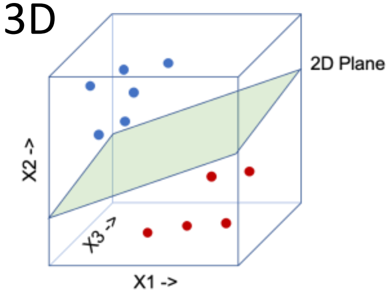
1D



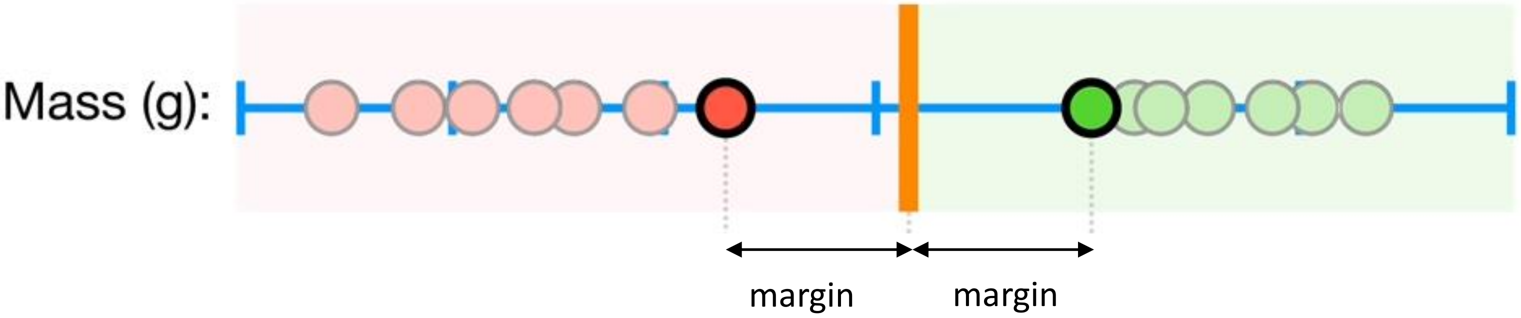
2D



3D

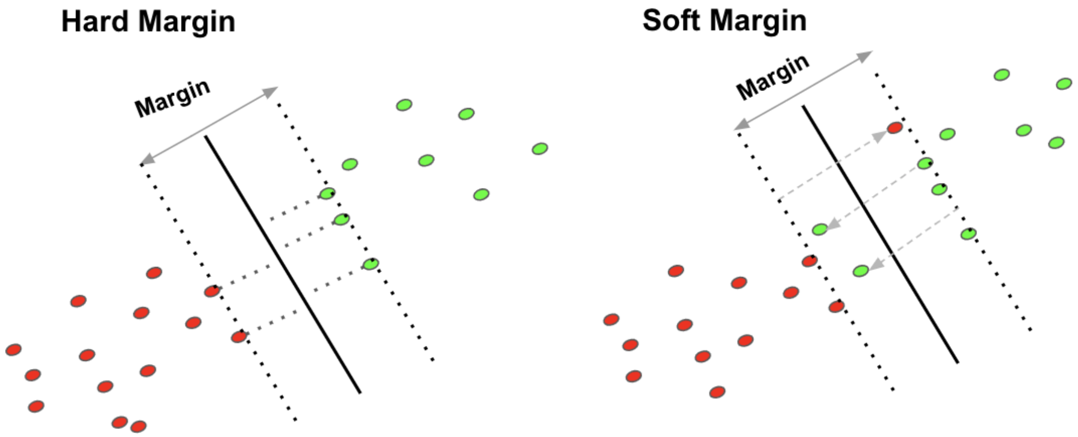


- Margin:

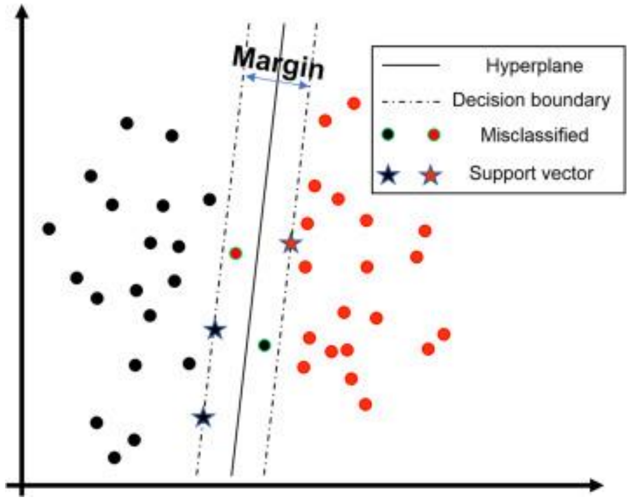


Support Vector Machine (SVM)

- Soft margin vs hard margin:



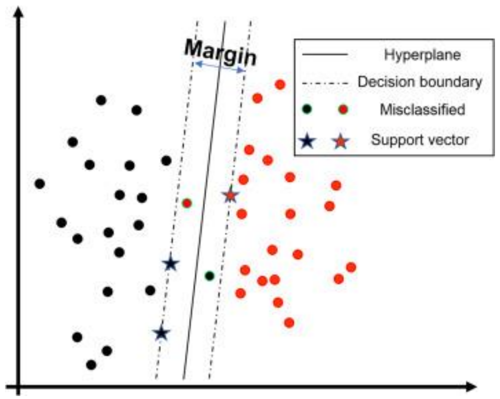
- Support vector classifier:
  - Vector that delimits one domain

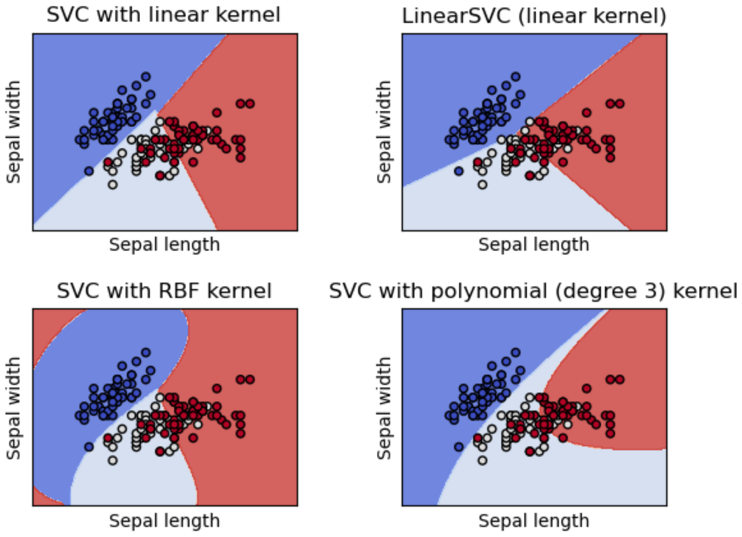


Support Vector Machine (SVM)

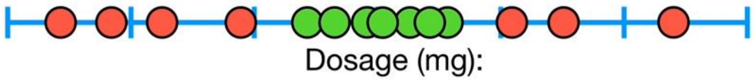
How to find those support vector ?

- Kernel functions:
  - Defines the shape of the delimitations
  - Common functions:
    - Linear function
    - Polynomial function
    - Radial basis kernel (adaptative shape kernel)
  - Can go to higher dimension -> kernel trick

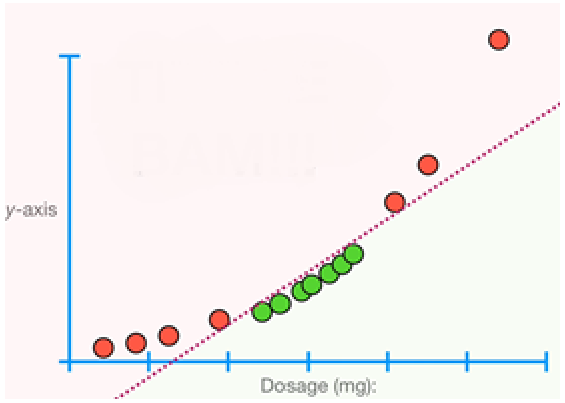




Kernel Trick



Really hard to find where to place the threshold

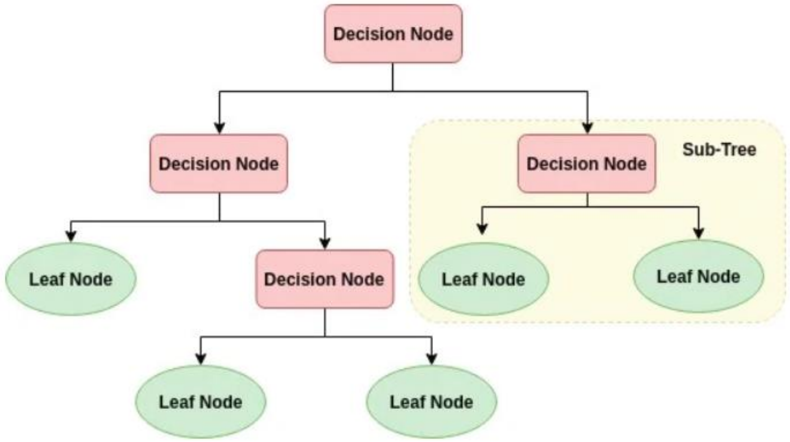


Data brought to higher dimension

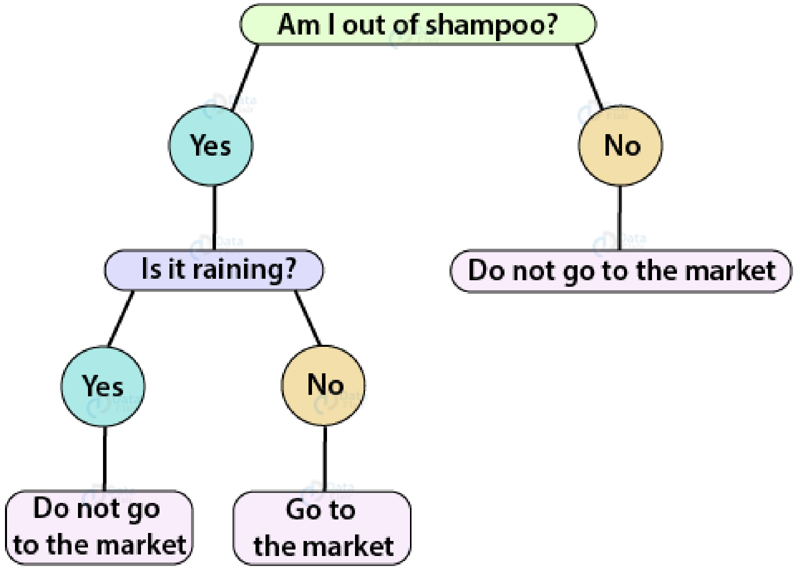
Decision Tree

- Goes from observations about an item:
  - Used for decisions
  - Decisions allows to follow one branch or another
- To the item target value:
  - The class of the data
  - Represented as leaves

- (Can also be used for regression)

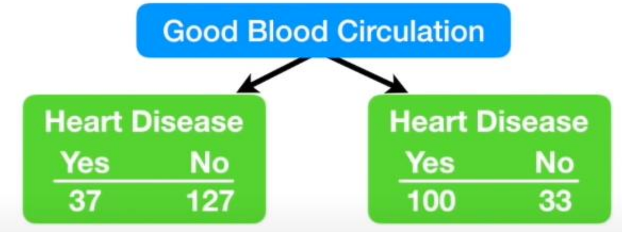


Decision Trees Example



How to build the decision tree ?

- Root node (first node):
    - Requires measuring impurity for all features
    - Most common method is Gini impurity (not to be confused with the coefficient):
        - Measures often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset.
    - Lowest Gini coefficient means least impurity
- Next nodes
    - Looping around all remaining features, find next best Gini coefficient
    - Check if you get a better Gini coefficient without separating
    - If the model gets better with a new separation, add it, else you have a new leaf node
    - Go until getting all leaf nodes are found



Impure data

We don’t have 100% yes and 100% no

A recap of advantages and disadvantages

Classification Model	Advantages	Disadvantages
Logistic regression	Probabilistic approach Statistical significance of features	Not suitable for nonlinear problems Tough to obtain complex relationships
K-Nearest Neighbors	Simple to understand Efficient	Manual choice of ‘k’ the number of neighbors Data storage space
Support vector machine (SVM)	Accurate Not biased by outliers Not sensitive to overfitting Effective in high dimensional spaces.	Only for linear problems High computation cost (when training)
Decision tree classification	Interpretability No need for feature scaling Works on both linear / non – linear problems	Poor results on very small datasets Overfitting can easily occur

Video

Machine Learning Model Evaluation Metrics

https://youtu.be/wpQiEHYkBys

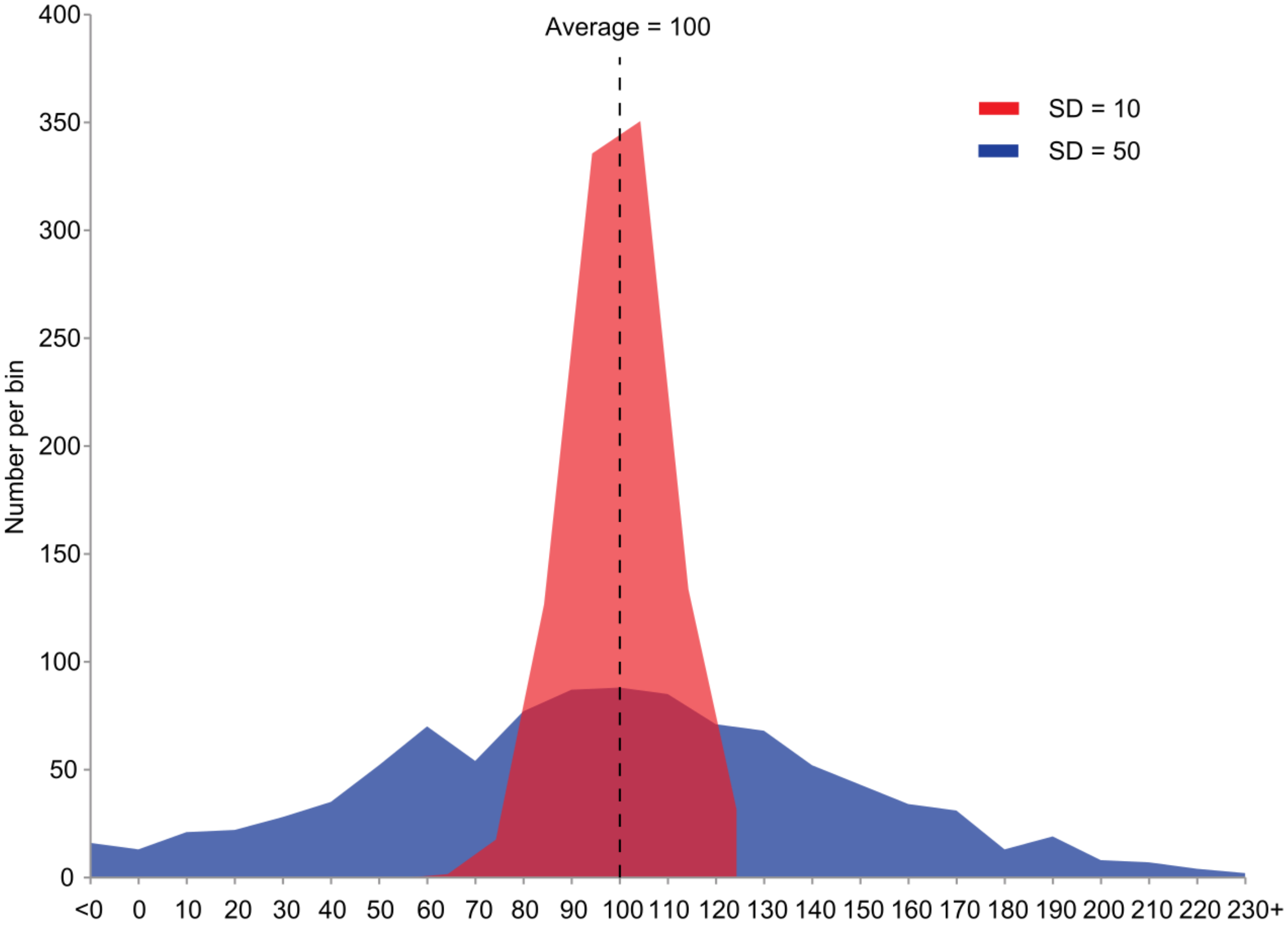
ANACONDACON

MACHINE LEARNING MODEL EVALUATION METRICS

Maria Khalusova // Developer Advocate // JetBrains

Annex

Variance



- A measure of the dispersion of the data
- It is the square of the standard deviation (SD)