# 11 - Prática: Predição e a Base de Aprendizado de Máquina (II)

## Descrição da Atividade

## 3. Predictive Models

### 1. [Activity] Linear Regression

- Fit a line to a data set of observation.
- Use this line to predict unobserved values.
- You use it to predict points .
- Y = ax+b.
- Sometimes called “maximum likelihood estimation”.
- How do we measure how well our line fits our data?
  - R-Squared.
  - Ranges from 0 to 1.

### 2. [Activity] Polynomial Regression

- Not all relationships are linear.
- Second Order Polynomial = y = axˆ2 + bx+ c.
- Third Order Polynomial = y = axˆ3+bxˆ2 + cx + d.
- Higher orders produce more complex curves.

### 3. [Activity] Multiple Regression, and Predicting Car Prices

- More than one variable influences, example: predicting a price for a car there is many attributes (body, style, brand...).
- End up with coefficients for each factor.

- Price = a + b1 mileage + b2age + b3doors.

- Can still measure fit with r-squared.

### 4. Multi-Level Models

- Some effects happen at various levels, example: your health depends on a hierarchy of the health of your cells, organs, your family...your wealth depends on your own work, what your parents did.
- You must identify the factors that affect the outcome you're trying to predict at each level
- Doing this is hard.

---

# 4. Machine Learning with Python

## 4.1 Supervised vs. Unsupervised Learning, and TrainTest

- UL: I don't tell what the "right" set is for any object.
  - Example: Cluster movies based on their properties.
- SL: The data the algorithm learns from comes with the "correct"answers.
  - Example: Predicting car prices based on historical sales data.
- ESL: If you have enough training data, you can split it into two parts: training set and test set.
  - Need to ensure both sets are large enough.
  - Must be selected randomly.
  - Great way to guard against overfitting.
- K-Fold Cross Validation: Sounds complicated, but it's a simple idea.
  - Split your data into K randomly-assigned segments.
  - Reserve one segment as your test data.
  - Train on each of the remaining K-1 segments and measure their performance against the test set.
  - Take the average of the K-1 r-squared scores.

## 4.2 [Activity] Using TrainTest to Prevent Overfitting a Polynomial Regression

- Create a Data Set.
- Split Data Set into Training(80%) and Test (20%).
- Trying fitting an 8th-degree polynomial to this data (which is almost certainly overfitting).
- Plot Train and Test.
- R-Squared to see if the model is great or not.
    - Test: 0.30.
    - Training: 0.64.
    - too bad.

## 4.3 Bayesian Methods Concepts

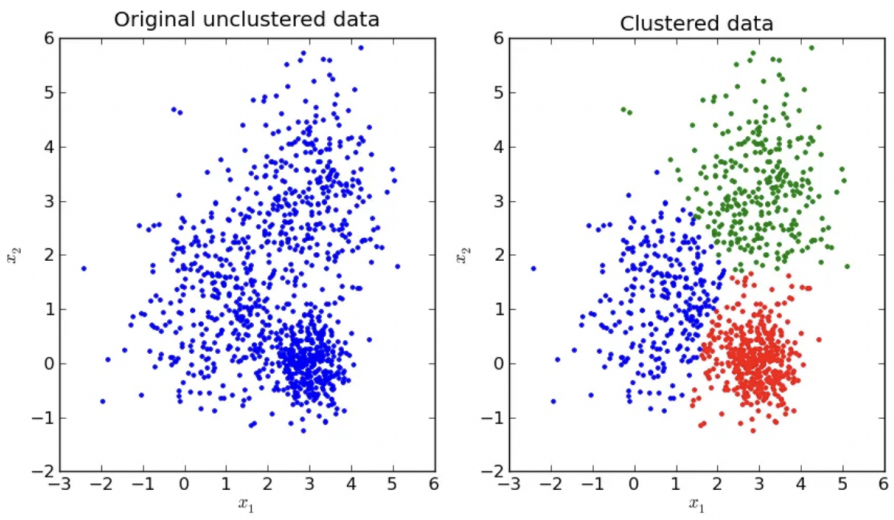
- P(A|B) = P(A) P(B|A) / P(B).
- Example: Spam Classifier.
- In this case is called “Naive Bayes”, because assumes the presence of different words are independent of each other.

## 4.4 [Activity] Implementing a Spam Classifier with Naive Bayes

- Load our training data.
- Split up each message into its list of words, and throw that into a MultinomialNB classifier. (CountVectorizer).
- Test.

## 4.5 K-Means Clustering

- Attempts to split data into K Groups that are closest to K centroids.




1. Randomly pick K Centroids (K-means).
2. Assign each data point to the centroid it's closest too.
3. Recompute the centroids based on the average position of each centroids points.
4. Iterate util points stop changing assignment to centroids.

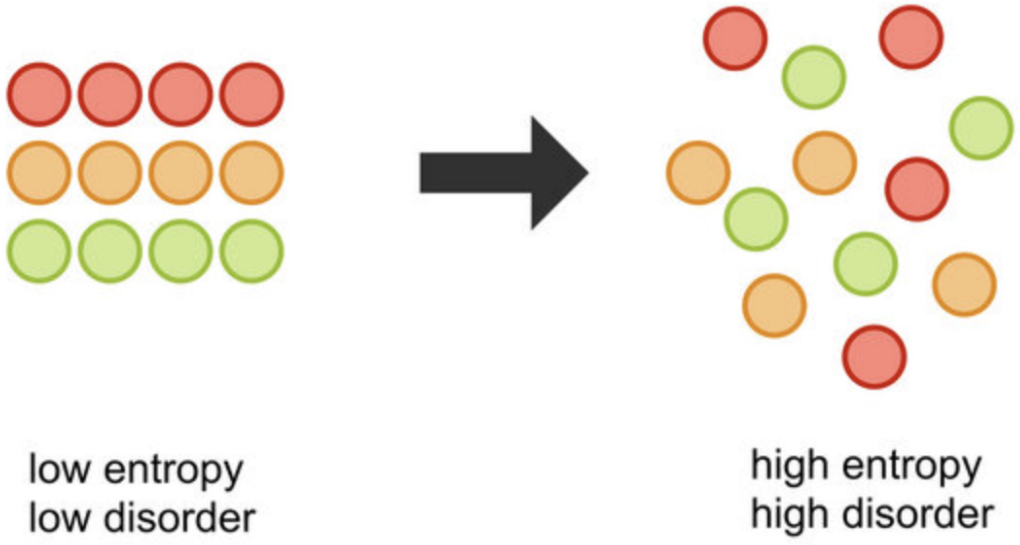
- Choosing K: Try increasing K values until you stop getting large reductions in squared error.
- Avoiding local minima.
- Labeling the clusters.

## 4.6 [Activity] Clustering people based on income and age

- Random data.
- Scale the data/Normalizing.
- Cluster the data.

## 4.7 Measuring Entropy

- A measure of a data set’s disorder.




## 4.8 [Activity] MAC Installing Graphviz

- Install homebrew.
- brew install graphviz.

## 4.10 Decision Trees Concepts

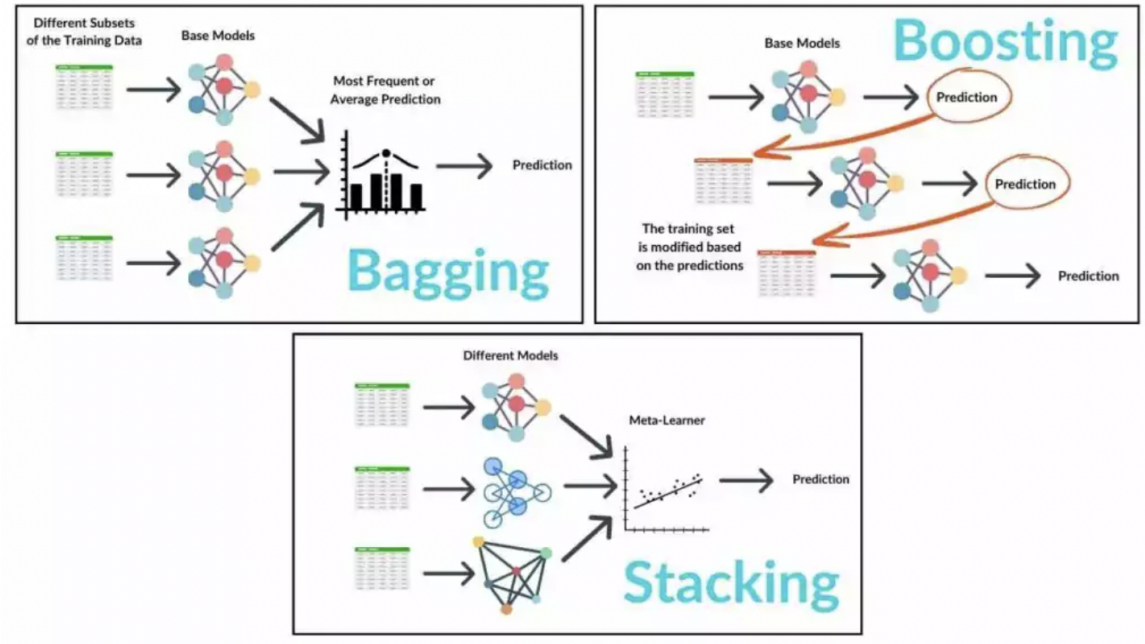
- Construct a flowchart to help you decide.
- Another form of supervised learning.
- Example: a system to filter out resumes based on historical hiring data, then you can train a decision tree and predicted whether a candidate will get hired.
- Are very susceptible to overfitting.
    - To fight this, we can construct several alternate decisions trees and let them “vote” on the final classification.

## 4.12 [Activity] Decision Trees Predicting Hiring Decisions

## 4.13 Ensemble Learning

- We use multiple models to try and solve the same problem, and let them vote on the results.
- Example: Random Forest.

- Bagging.
- Boosting.
- Bucket of models.
- Stacking.

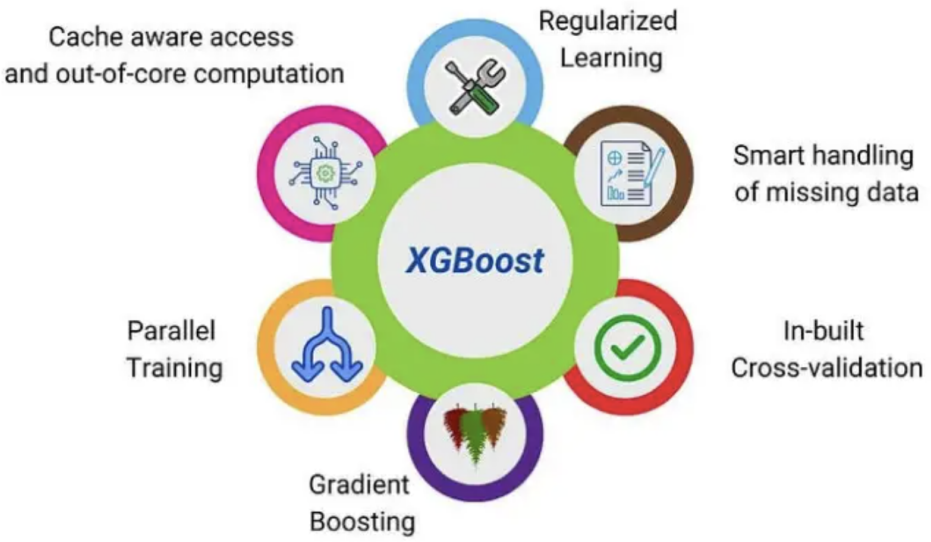



- Advanced Ensemble Learning.
  - Bayes Optimal Classifier.
  - Bayesian Parameter Averaging.
  - Bayesian Modelo Combination.

## 4.14 [Activity] XGBoost

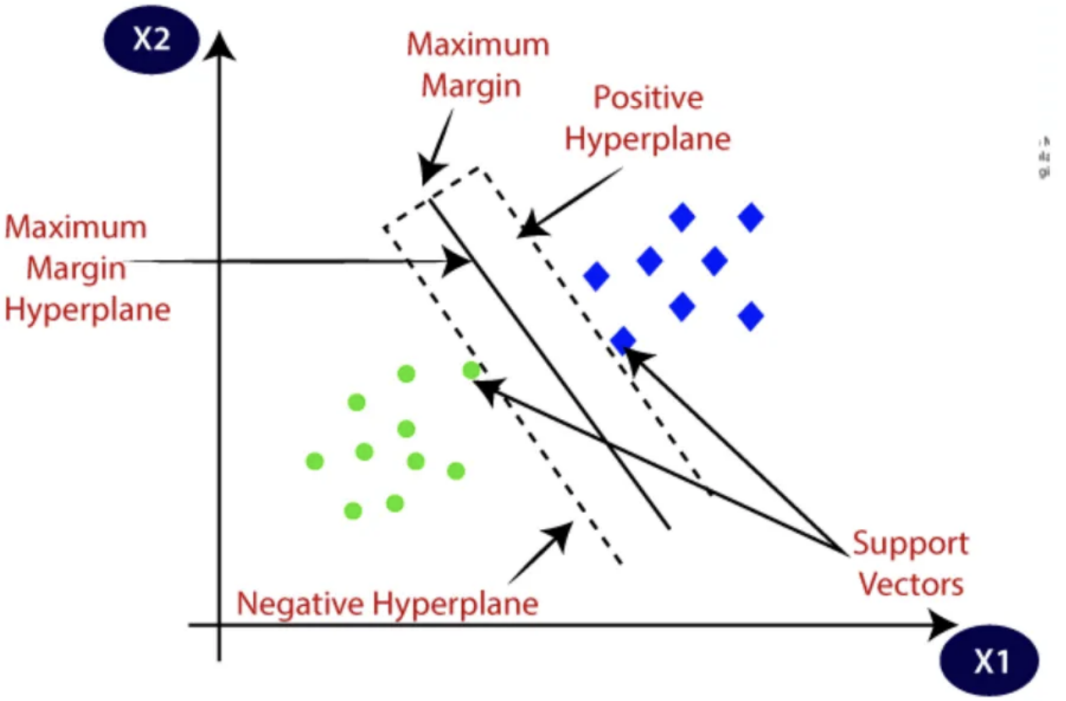
- Each tree boosts attributes that led to mis-classification of previous tree.
- Features.
  - Regularized boosting (prevents overfitting).
  - Can handle missing values automatically.
  - Parallel processing.
  - Can cross-validate at each iteration.
  - Incremental training.

- Can plug in your own optimization objectives.
- Tree pruning.

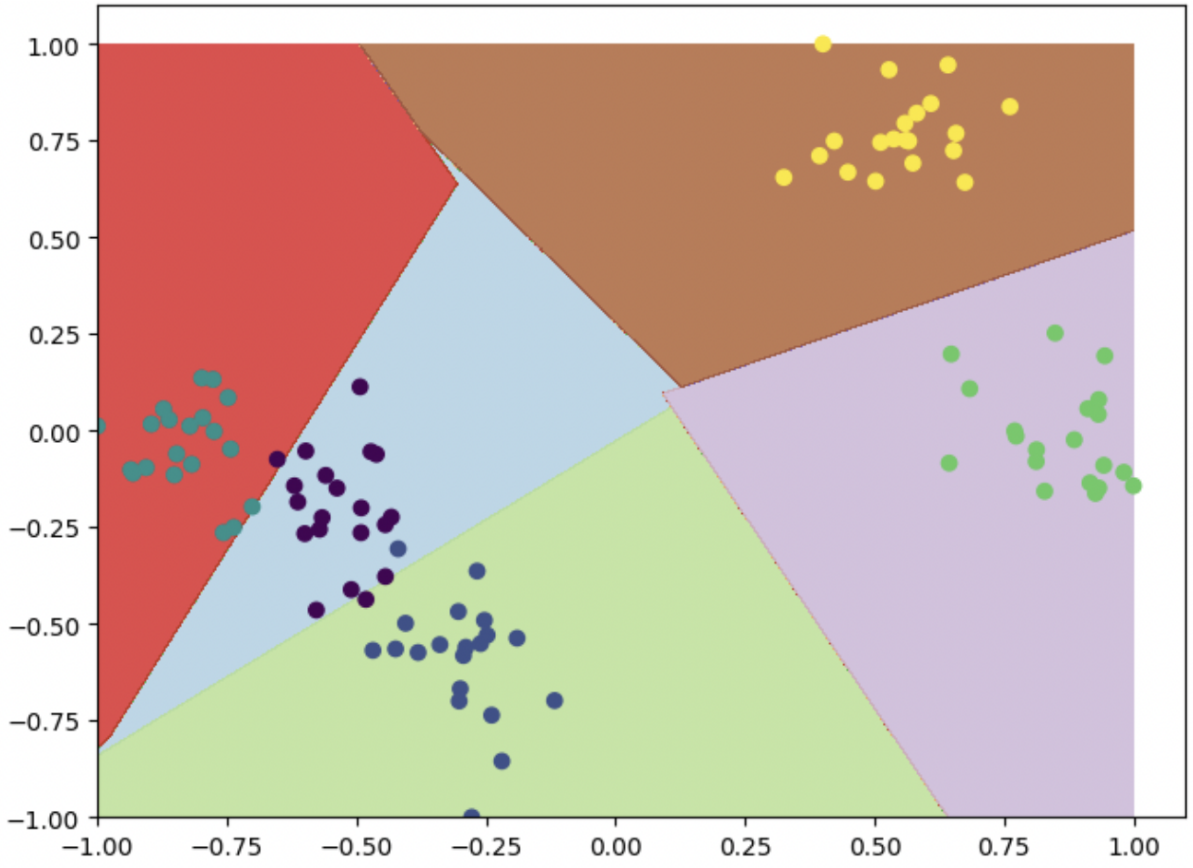



## 4.15 Support Vector Machines (SVM) Overview

- Works well for classifying higher-dimensional data (lots of features).
- Finds higher-dimensional support vectors across which to divide the data.
- Uses Kernel Trick to represent data in higher-dimensional spaces to find hyperplanes that might not be apparent in lower dimensions.




## 4.16 [Activity] Using SVM to cluster people using scikit-learn




# Conclusões

In this section, the activities cover predictive models and machine learning, introducing fundamental techniques for data analysis and forecasting. Linear and polynomial regression were applied to understand how to fit curves to datasets and predict future values, with the R-Squared coefficient used to evaluate accuracy.

Multiple regression demonstrated how various factors influence outcomes, such as predicting car prices. Additionally, supervised and unsupervised machine learning models, like SVM and K-Means, were introduced for efficient data classification and clustering, alongside advanced methods like XGBoost and ensemble learning to improve the accuracy of complex predictions.

# Referências

Curso: Machine Learning, Data Science and Deep Learning with Python. Disponível em <www.udemy.com> Acessado em: 16/10/2024.

Artigo: K-Means Data Clustering. Disponível em <https://towardsdatascience.com/k-means-data-clustering-bce3335d2203> Acessado em: 16/10/2024.

Artigo: Support Vector Machine (SVM) Algorithm. Disponível em <https://medium.com/@sumbatilinda/support-vector-machine-svm-algorithm-064566b5d411> Acessado em: 16/10/2024.

Artigo: What is ensemble learning in machine learning?. Disponível em <https://spotintelligence.com/2023/08/09/ensemble-learning/> Acessado em: 16/10/2024.