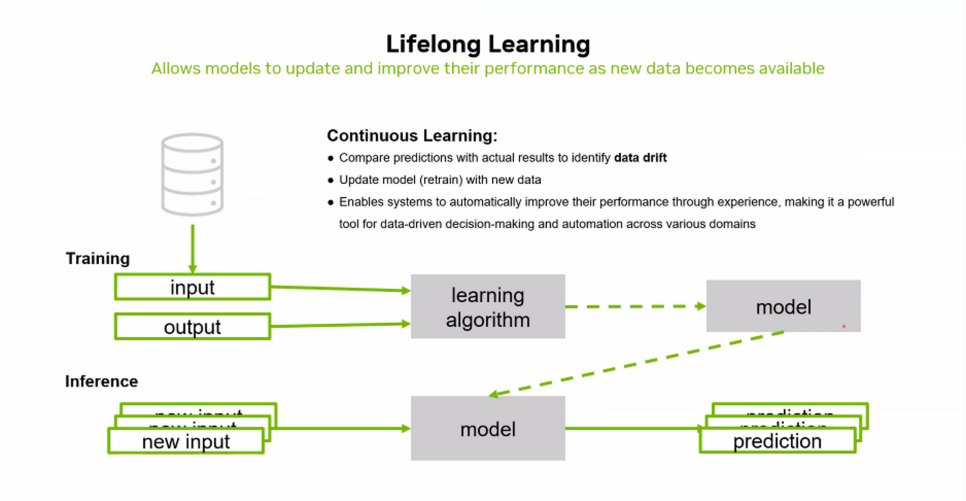
Lec 9

31 октября 2024 г. 20:25

1. What is ML?
2. In traditional programming we usually use rules based approach. But it takes a lot of time and hard.
3. Training approach: input + output -> some learning algo -> model. Than we can use this model to get predictions on real data.
4. Main approaches: supervised(labeled data, there is right answer) and unsupervised(there is no ready answer, only input data).
5. Supervised is divided into regression and classification. Regression is used to predict continous value, classification - labeling objects with some class.
6. Unsupervised is divided into clustering and dimensionality reduction. Clustering - splitting objects in groups(clusters) by some metrics.
7. 

8. Cool examples:

Common Use Cases

	Supervised	Unsupervised
Online Services	• Spam / Not Spam • Predict Probability Customer Will Click on Ad	• Group Articles into Categories • Similar Search Results
Financial Services	• Credit Card Fraud	• Anomaly Detection
Telecom	• Predict Customer Churn	
Health Care	• Probability of Readmission • Predict Days of Hospital Stay	• Patient Similarity
Real Estate	• Predict House Price	
Retail	• Prejudice Price • Forecast Sales • Sentiment Analysis	• Similar Customers • Product Similarity • Customer Group • Products Which Are Purchased Together

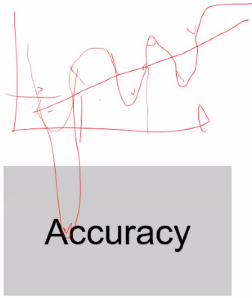
Benefits:

- **Adaptability**: models can adapt to new data, new tasks, and improve over time with retraining
- **Problem solving**: excels at complex problems where patterns are not easily discernible

9.

Model Evaluation

Assessing the quality and performance of a trained machine learning model



- Assessing the performance and quality of a trained machine learning model using different metrics and techniques
- To understand how well the model generalizes to unseen data and makes accurate predictions:
  - Underfit
  - Overfit
  - Data leakage
- Employing appropriate metrics based on the type of problem:
  - Classification: **accuracy**, **precision**, **recall**, **F1-score**, **ROC-AUC**, **confusion matrix**
  - Regression: **R-squared**, **MSE**, **RMSE**, **MA**
- Using techniques likes:
  - Cross-validation
  - Holdout validation
  - Bootstrapping

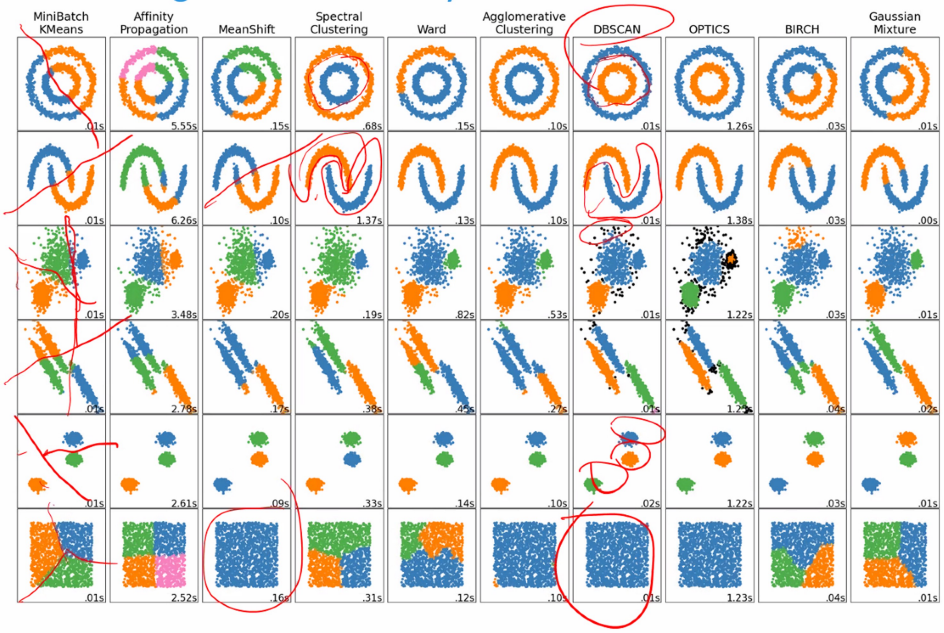
10.

What Kind of Problem do I Need to Solve? How do I Solve it?

The Problem to Solve	The Category of Techniques	Covered in this Course
I want to group items by similarity. I want to find structure (commonalities) in the data	Clustering	K-means clustering Hierarchical clustering DBSCAN
I want to discover relationships between actions or items	Association Rules	Apriori FP Growth
I want to determine the relationship between the outcome and the input variables	Regression	kNN Linear Regression Logistic Regression
I want to assign (known) labels to objects	Classification	kNN Naïve Bayes Decision Trees
I want to find the structure in a temporal process I want to forecast the behavior of a temporal process	Time Series Analysis	ACF, PACF, ARIMA
I want to analyze my text data	Text Analysis	Regular expressions, Document representation (Bag of Words), TF-IDF

11.

Clustering Method Diversity



Partially based on scikit-learn.org materials 5

12. Clustering usually uses distance paradigm. In space N objects are being valued how close they are. Close groups gather into clusters. Also there can be some objects in none of clusters

K-Means

13. Clustering numerical data, so input must be numerical
14. Need to have defined distance metric
15. Defining K - number of clusters.
16. Output is set of centers of clusters(centroids).

Algo itself:

1. Selecting K random centroids. But ideally they should be chosen from real dots from dataset in a big spam of dots(but not necesserly, it's a heuristic and can be adjusted to get more accurate results).
2. Evaluating distance for every dot and assigning it to the closest cluster
3. Recalculating centroid inside of every cluster(recalculated centroid is a mean value of all dots in cluster)
4. Repeat 2 and 3 until centroids stay almost unchangable or until certain number of iterations reached
5. Result: output centroids and output dots distribution among clusters

17. If K is unknown than it is tried to be guessed