Machine Learning

Machine Learning at a Glance

경희대학교 컴퓨터공학과
2019102191 신주영

Machine Learning's definition

- It is a research field that allows computers to have the ability to learn without explicit programming.
- If the performance of a computer program on a dataset T was improved by experience E when the accuracy of a computer program on a dataset T was measured as P, this computer program learned about data T and accuracy P by experience E.
- Training set: Samples used by the system to learn
- Training instance: Each training data

Why using Machine Learning

- Traditional solutions require many manual tuning and rules: one machine learning model can simplify code and perform better than traditional methods.
- Complex problems with no solutions in traditional ways: solutions can be found with the best machine learning techniques.
- Flexible environments: Machine learning systems can adapt to new data.
- Gain insight from complex issues and massive amounts of data

Application example

- Analyze and automatically classify product images on the production line
- Scan the brain to diagnose the tumor
- To classify news articles automatically
- To automatically summarize long documents
- Chatbot
- Predicting the company's earnings for the coming year
- To detect fraudulent credit card transactions
- Reduce high-dimensional datasets to low-dimensional(2D, 3D)

Types of Machine Learning Systems

- Training under human supervision? (Supervised, Unsupervised, Semi-supervised, reinforcement Learning)
- Do you learn progressively in real time? (Online learning, Batch Learning)
- Compare known and new data points or create predictive models by finding patterns in the training data set (Case-based learning and model-based learning)

Supervised Learning

- In supervised learning, a label is included in the training data.
- Classification is a typical supervised learning task.
- Regression task is to predict the target value using features.

Supervised Learning algorithms

- K-Means algorithm
- Linear regression
- Logistic regression
- Support Vector Machine(SVM)
- Decision tree and random forest
- Neural network

Unsupervised Learning

- Unsupervised learning is literally unlabeled in training data.
- The system must learn without any help.

Clustering	Outlier detection, Dimensionality reduction	Visualization, Dimensionality reduction	Association rule learning
K-Means	One-class SVM	Principal component analysis (PCA)	Apriori
DBSCAN	Isolation forest	Kernel PCA	Ecalt
Hierarchial cluster analysis (HCA)		Locally linear embedding(LLE)	
		t-SNE	

Unsupervised Learning

- Clustering refers to the process of dividing objects into several clusters (subgroups) when they are given.
- Visualization creates a 2D or 3D representation capable of schematizing large, unlabeled, high-dimensional data.
- Dimension reduction simplifies data without losing too much information.
- Outlier detection is one of the data analysis techniques based on data mining, which detects outliers, "observed values suspected of being generated by different methods from other observations within a data" (Hawkins, 1980).
- Novelty detection aims to detect new samples that look different from all the samples in the training set.
- Association rule learning finds relationships between characteristics in large amounts of data.

Semi-supervised learning

- Semi-supervised learning deals with labeled data only in part.
- For example, if you add a label to that cluster after clustering, other data points are assigned labels.

Reinforcement learning

- In reinforcement learning, the agent observes the environment and executes the action and receives a reward(penalty) as a result. And we execute a strategy called policy to get the greatest reward.

Batch learning

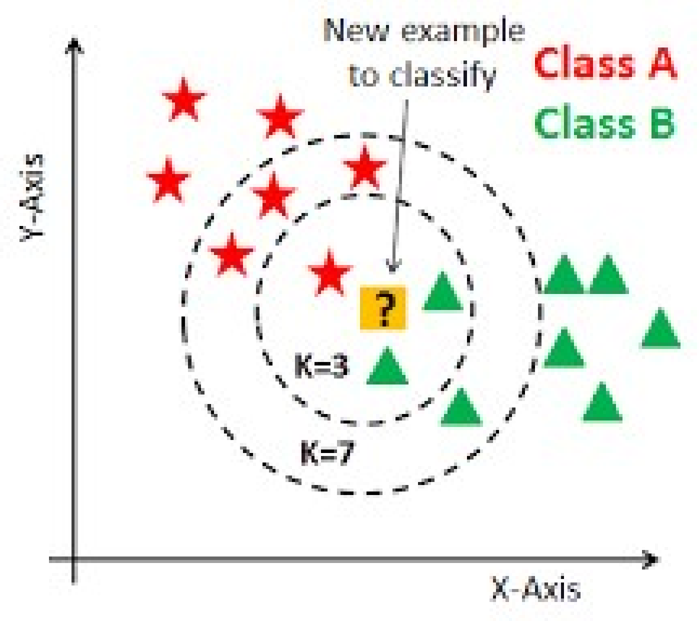
- In batch learning, the system cannot learn progressively.
- In order for batch learning systems to learn about new data, new versions of the system must be trained from scratch using the full data.

Online learning

- Online learning trains the system by injecting data one at a time or in small batches called mini-batch.
- The data that has been learned is no longer needed, so you can throw it away.
- incremental learning algorithms can also be used for systems that learn very large datasets that cannot fit into the main memory of a single computer.

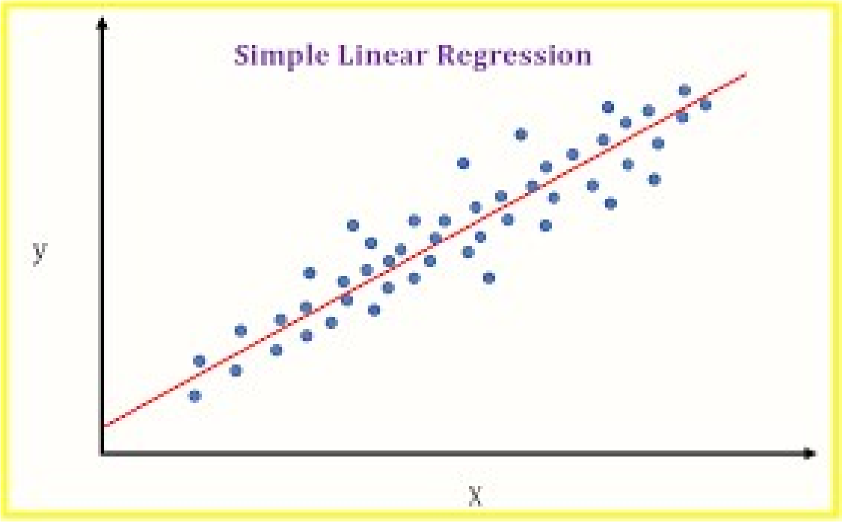
Case-based learning

- Measure the similarity between the surrounding samples.
- A representative example is the kNN algorithm.



Model-based learning

- A model of a given sample is made and used for prediction.



Model-based learning

procedure

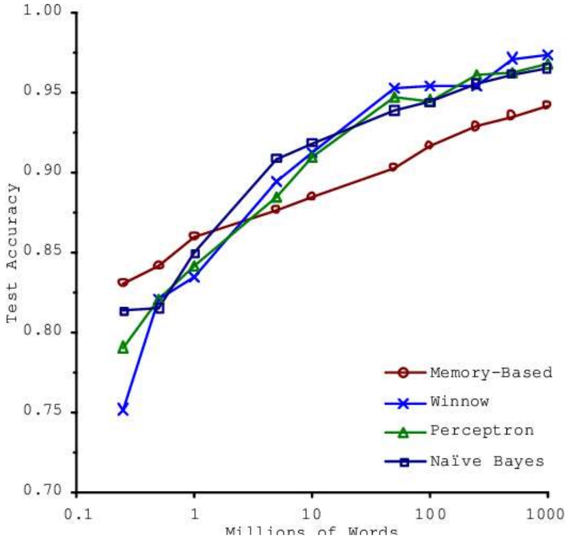
1. Analyze data
2. Select model
3. Train model using train data
4. Predict result and hope this model will generalized well

Key challenges in Machine Learning

- Not enough training data
- Unrepresentative training data
- Low quality data
- Irrelevant characteristics
- Over-fitting
- Under-fitting

Not enough training data

- Machine learning algorithms need a lot of data to work well.
- A very simple problem also requires thousands of data.
- Complex problems such as image and speech recognition may require millions of data.



Unrepresentative training data

- If the sample is small, sampling noise (Unrepresentative data by chance) is generated.
- Even very large samples may not be representative if the sampling method is incorrect.

Low quality data

- If the training data is full of errors, outliers, and noise, it is difficult to find patterns with built-in machine learning systems and does not work well.
- So, training data purification is important.
  - If you are certain that some samples are outliers, it is recommended that you ignore them or correct them manually.
  - If some samples are missing a few attributes, it is necessary to decide whether to ignore all of these attributes, ignore this sample, fill in the missing values, or train the model separately from the model with this attribute.

Irrelavant training data

- Feature selection: Choose the most useful characteristic for training among the characteristics you have.
- Feature extraction: Combining properties creates more useful properties. Representatively, dimension reduction algorithms can be helpful.

Over-fitting

- The accuracy of the training data is good, but it is called overfitting because an error may occur in an actual test.
- Constraining the model to reduce the risk of overfitting is called regularization.

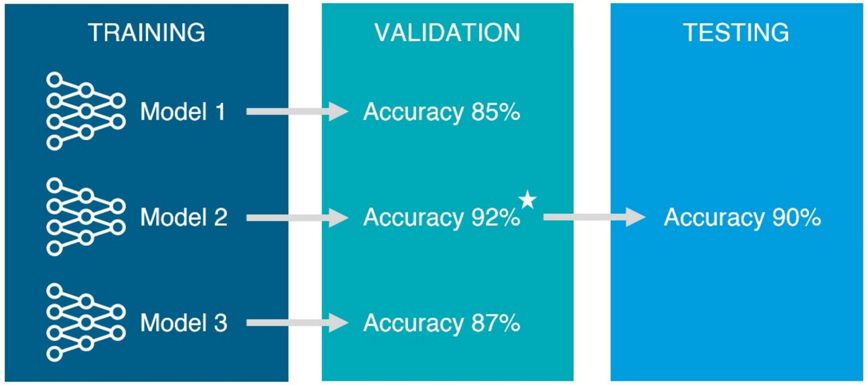
Under-fitting

- Under-fitting occurs when the model is too simple to learn the inherent structure of the data.

- Key techniques to address this issue
  - We select a robust model with more model parameters.
  - It provides better properties for learning algorithms.
  - Reduce the constraints of the model.

Test and Validation

- Holdout Validation: We evaluate several candidate models by tearing off part of the training set and choose the best one.
  - Training data: dataset to train
  - Test data: dataset to test
- Cross Validation: For each validation set, models trained from the rest of the data are evaluated on that validation set.



Practice problem

1. How can I define machine learning

- It is a research field that allows computers to have the ability to learn without explicit programming.

2. Name four types of problems that machine learning can help you with

- one machine learning model can simplify code and perform better than traditional methods.
- solutions can be found with the best machine learning techniques.
- Machine learning systems can adapt to new data.
- Gain insight from complex issues and massive amounts of data

Practice problem

3. What is labeled training set

- Labeled data is a designation for pieces of data that have been tagged with one or more labels identifying certain properties or characteristics, or classifications or contained objects.

4. What are the two most widely used supervised learning tasks?

- Regression
- classification

경희대학교
KYUNG HEE UNIVERSITY

Practice problem

5. What are the five most widely used unsupervised learning task

- Clustering
- Outlier detection
- Dimensionality reduction
- Visualization
- Association rule learning

6. What kind of machine learning algorithm can be used to walk the robot on multiple terrain without prior information

- Reinforcement learning

Practice problem

7. Which algorithm should I use to divide the customer into groups

- Supervised Learning: SVM
- Unsupervised Learning: K-Means

8. Can you see the problem of spam detection as a problem of supervised or unsupervised learning

- Supervised Learning

Practice problem

9. What is online learning

- Online learning trains the system by injecting data one at a time or in small batches called mini-batch.

10. What is out-of-core learning

- Applying online learning algorithms(mini-batch) to systems that learn very large datasets that can't fit into the main memory of a computer.

Practice problem

11. Which learning algorithm relies on silmilarity measurements to make predictions?

- Case based learning (kNN, etc…)

12. What is the difference between the model parameters and the hyper parameters of the learning algorithm?

- The model parameter is a parameter obtained by learning, and the hyper parameter is a parameter that must be input before learning

Practice problem

13. What are model-based algorithms looking for? What is the most common strategy this algorithm uses for success? How do you make predictions?

- Generalized model
- Regularization
- Enter the characteristics of the new sample into the prediction function of the model using the parameters found by the learning algorithm.

14. What are the main challenges of machine learning?

- Insufficient data
- Poor data quality
- Unrepresentative data
- A meaningless trait
- Over-fitting
- Under-fitting

Practice problem

15. What is the problem if the model performs well in the trainig data but generalizes poorly in the new sample? What are the three possible solutions?

- Over-fitting
- It simplifies by selecting a model with a small number of model parameters, reducing the number of characteristics in training data, or imposing constraints on the model.
- More training data
- Reduce training data's noise

16. What is the test set and why should I use it?

- To prevent over-fitting

Practice problem

17. What is the purpose of the validation set?

- To choose the best model and tune the hyperparameter.

18. What is a training-development set? When do I need it and how do I use it?

- It is to separate part of the training set and create another one.
- When a data mismatch occurs.

Practice problem

19. What is the problem with tuning the hyper parameter using the test set?

- Hyperparameters create models optimized for test sets.