Machine Learning

Introduction to Machine Learning

Dr. Sherif Saad

Learning Objectives

Introduce the students to machine learning concepts

Explain the main three types of learning and ML terminology

Understand the building blocks for successfully designing machine learning systems.

How to apply machine learning algorithms (not really how to create them)

Outlines

- Introduction To Machine Learning
- Roadmap For Building ML System
- Types of Machine Learning Techniques

Machine Learning

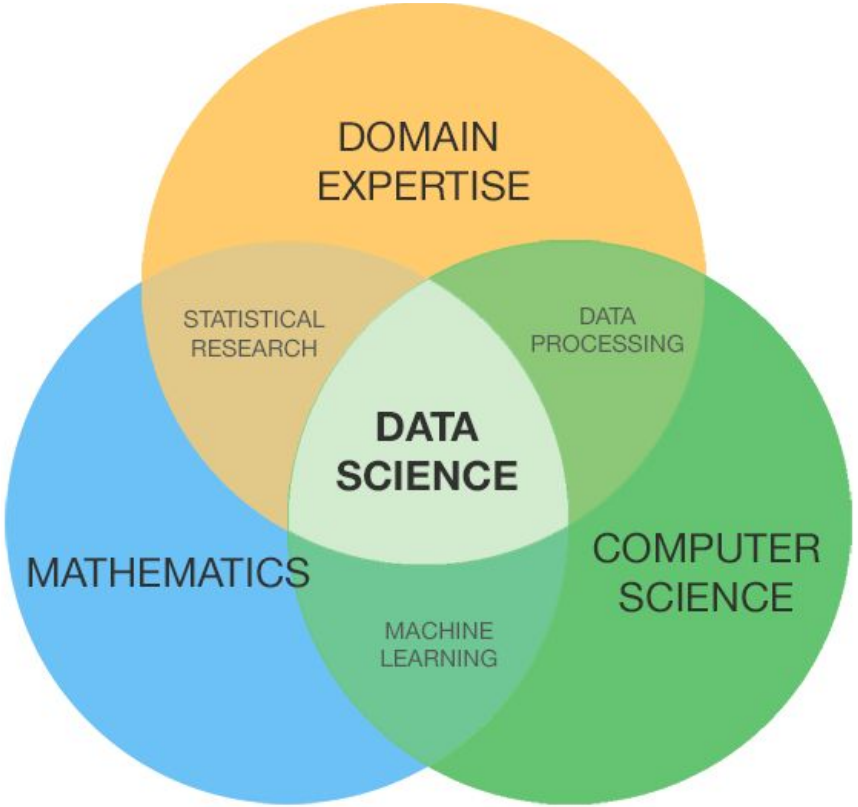
What is machine learning?

- The term was coined by Arthur Samuel in 1950s
- Machine learning refers to a collection of data-driven algorithms that make sense of data.
- ML is a a collection of data driven algorithms where rules and solutions are derived by examining data
- Machine learning is the use of scientific methods to explore data and extract knowledge and insights
- Machine learning is intersection of many fields, primarily, statistics, probability, computer science, and engineering.

Data Science

What is data science?

- The intersection of math and statistics knowledge, computer science, and domain expertise.
- Analyze data sources using different techniques to extract descriptive, predictive and prescriptive insights.



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

Machine Learning

Why learning is Possible?

Data has unique implicit and explicit characteristics.

Patterns are everywhere, without patterns learning is impossible.

What if ML does not exist?

If we do not have ML then humans have to manually derive rules and build models from analyzing large amounts of data.

What are the applications of machine learning?

- **Finance:**
  - Predict Stock Prices
  - Estimate Insurance Payouts
  - Recommend Tax Deductions
- **Healthcare:**
  - Predicting Emergency Room Wait Times
  - Predicting Psychopaths
  - Identify Heart Failure
  - Diagnosing Cancer
- **Cybersecurity:**
  - Fraud and Money Laundry Detection
  - Anomaly Detection
  - Biometrics & Access Control
  - Email Spam
- **Marketing:**
  - Recommendations
  - Search Engines
  - Opinion Spam and Fake News

What are the applications of machine learning?

- **Natural Language Processing:**
  - Text Generation
  - Machine Translation
  - Text Mining
  - Text Classification
- Entertainment Industry
  - Creating an AI Trailer
  - Personalizing Choices
  - Video Games

- **What else, can you list real life example of ML application?**

Roadmap For Building ML System

What are the key steps in data analysis process?

1. Data collection (what are the questions we are trying to answer?)
2. Data Wrangling (Munging): is the process of manually converting or mapping data from one "raw" format into another format that allows for more convenient consumption
3. Data Exploration: Data Visualisation
4. Data Transformation: to prepare it for modeling
5. Data Modelling: Selecting a model that fit the data
6. Model Evaluation: Test and evaluate the learning process

Roadmap For Building ML System

We will focus on the following phases:

- Preprocessing|Transformation
- Learning | Modelling
- Evaluation
- Prediction

Roadmap For Building ML System

Preprocessing

1. Raw data rarely comes in the form and shape that is necessary for the optimal performance of a learning algorithm.
2. Many machine learning algorithms also require that the selected features are on the same scale for optimal performance, which is
3. Often achieved by transforming the features in the range [0-1] or std with zero mean and unit variance
4. Some of the selected features may be highly correlated and therefore redundant to a certain degree.
5. Reducing the dimensionality of our feature space has the advantage that less storage space is required, and the learning algorithm can run much faster.

Roadmap For Building ML System

Learning and Modelling

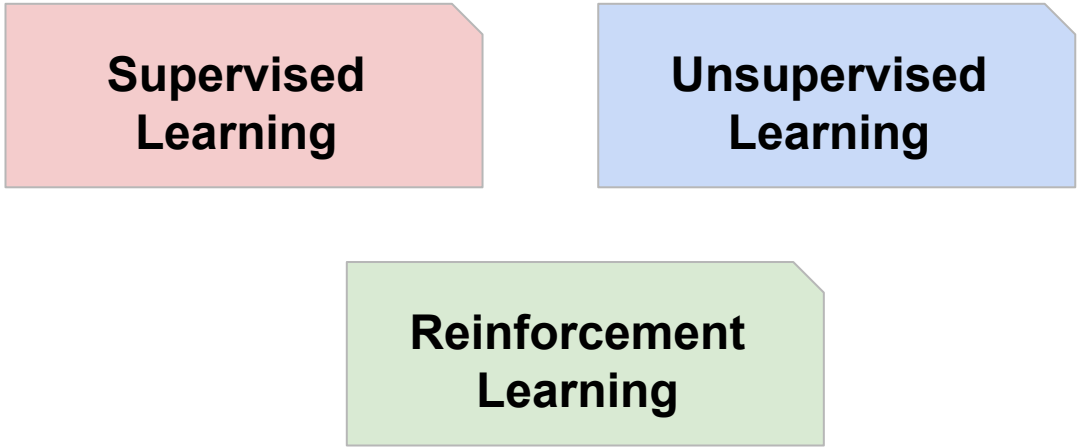
- Each classification algorithm has its inherent biases, and no single classification model enjoys superiority if we don't make any assumptions about the task.
- In practice, it is therefore essential to compare at least a handful of different algorithms in order to train and select the best performing model.

Roadmap For Building ML System

Evaluating Models

- Here we use the unseen data to evaluate the selected model and use some metrics to measure the performance of the selected model

Types of Machine Learning



Types of Machine Learning: Supervised Learning

- Enable us to learn a model from labeled training data that allows us to make predictions about unseen or future data.
- The term supervised refers to a set of samples where the desired output class (labels) are already known.
- There are two main applications of supervised learning, namely, classification and regression.
- Examples: Email Spam Detection, Cancer Diagnosis, Stock Price Prediction

Supervised Learning

We have an input variable x and an output variable y, we use supervised learning algorithm to learn a mapping function from x to y, where Y = F(x)

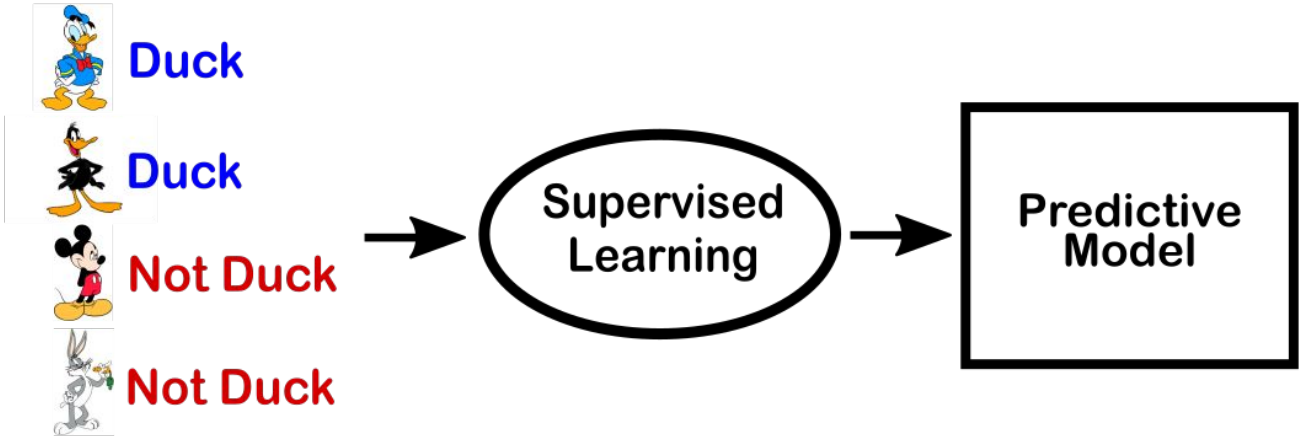


Image source: http://www.java-machine-learning.com/

Supervised Learning

Our objective to learn a mapping function so well that when we have a new input (unseen before) data that our predictive model can predict the out variable Y with high accuracy

Image source: http://www.java-machine-learning.com/

Supervised Learning

As we mentioned before supervised learning problems can be categorized into classification problems and regression problems.

- Classification:
  - Assign categorical, unordered labels to instances.
  - The goal is to predict the categorical class labels of new instances based on past observations.
- Regression:
  - The prediction of continuous outcomes,
  - Regression is to predict a continuous-valued feature associated with a sample, such as weight, temperature, or price

Supervised Learning

Most Common Algorithms

- K-Nearest Neighbors
- Decision Trees
- Random Forest
- Support Vector Machines
- Neural Networks
- Logistic Regression
- Linear Regression

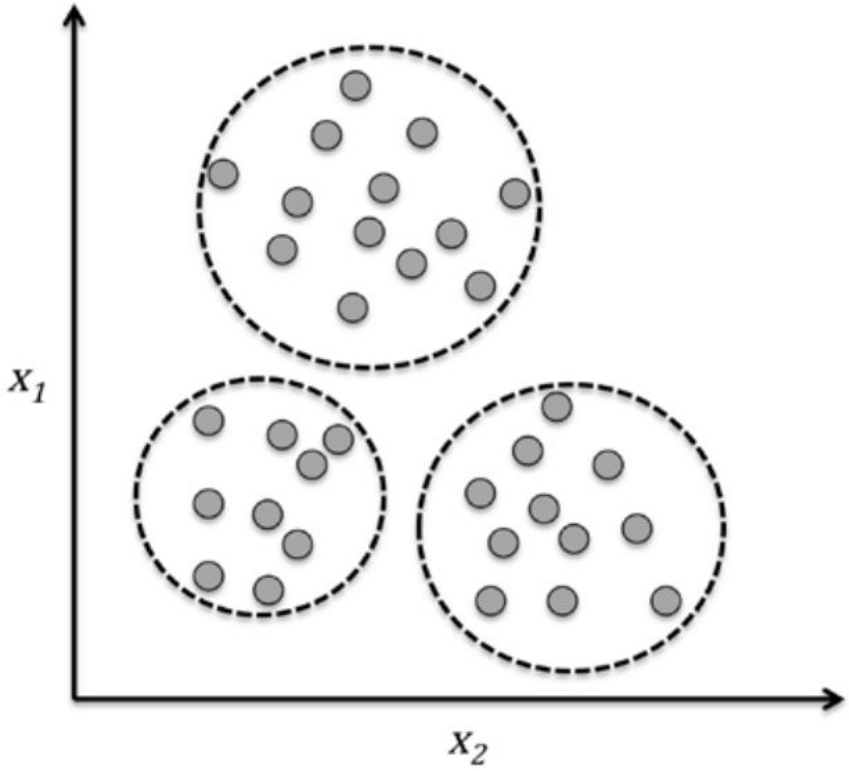
Types of Machine Learning: Unsupervised Learning

- Can draw inferences from datasets consisting of input data without labels
- We are dealing with unlabeled data or data of unknown structure.
- Using unsupervised learning techniques, we are able to explore the structure of our data to extract meaningful information without the guidance of a known outcome variable or reward function.
- There are two main application of unsupervised learning, namely, data clustering and dimensionality reduction

Unsupervised Learning

- What is Clustering?
  - is an exploratory data analysis technique that allows us to organise a pile of information into meaningful subgroups (clusters) without having any prior knowledge of their group memberships.
  - Each cluster that may arise during the analysis defines a group of objects that share a certain degree of similarity but are more dissimilar to objects in other clusters, which is why clustering is also sometimes called "unsupervised classification."

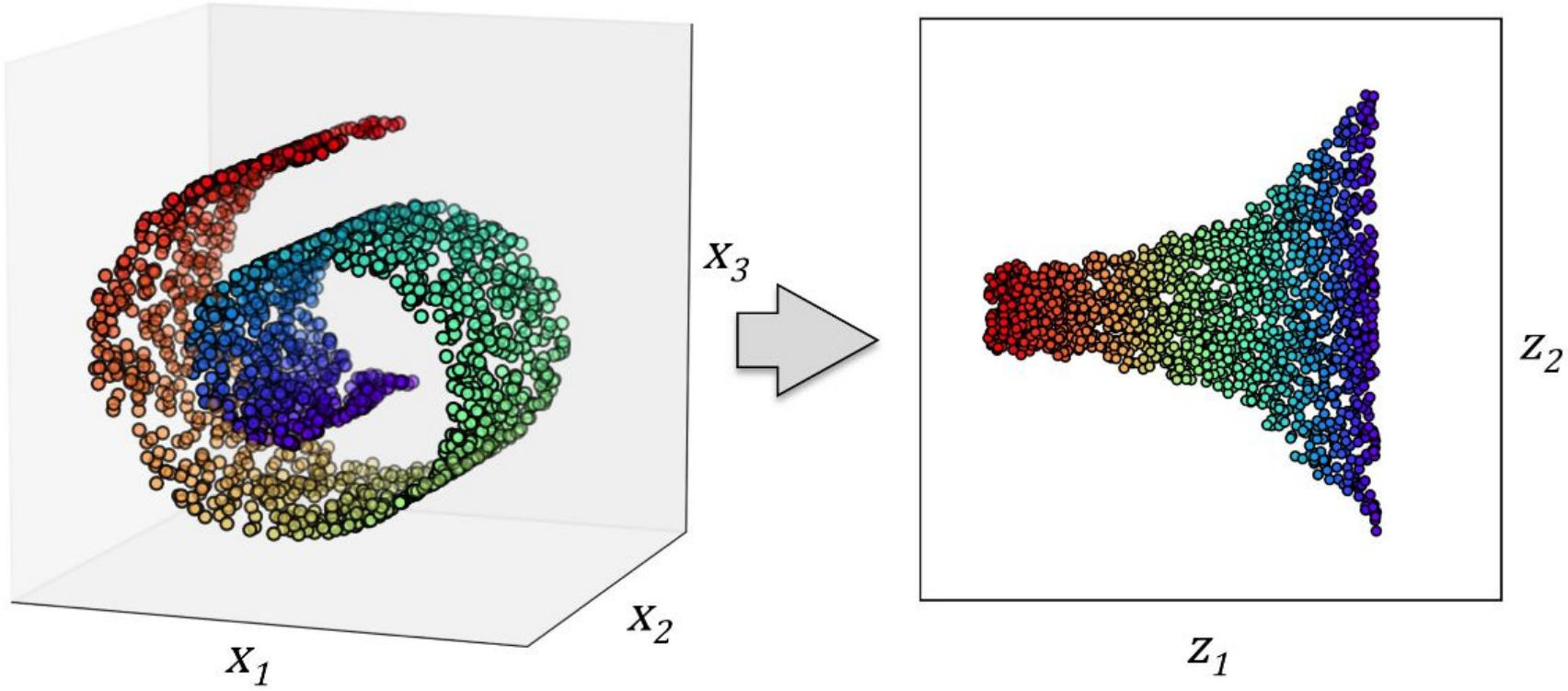
Data Clustering



Unsupervised Learning

- What is Dimensionality Reduction?
  - A commonly used approach in feature preprocessing to remove noise from data, which can also degrade the predictive performance of certain algorithms, and compress the data onto a smaller dimensional subspace while retaining most of the relevant information.
  - Sometimes, dimensionality reduction can also be useful for visualizing data—for example, a high-dimensional feature set can be projected onto one-, two-, or three-dimensional feature spaces in order to visualize it via 3D- or 2D-scatterplots or histograms.

Dimensionality Reduction



Unsupervised Learning

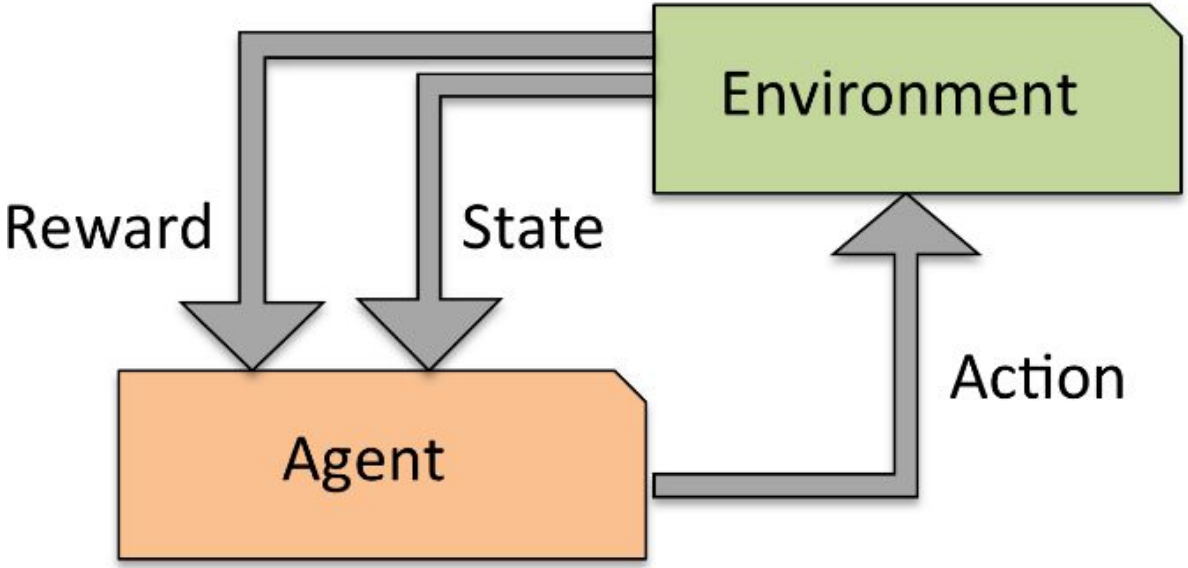
- Most Common Algorithms
  - K-Means
  - Principle Component Analysis
  - Hierarchical Clustering
  - Association Rule Learning
  - Rough Set
  - Neural Network

Types of Machine Learning: Reinforcement Learning

- Is a form of supervised learning with a reward function.
- The goal is to develop a system (agent) that improves its performance based on interactions with the environment.
- Through the interaction with the environment, an agent can then use reinforcement learning to learn a series of actions that maximizes this reward via an exploratory trial-and-error approach or deliberative planning.

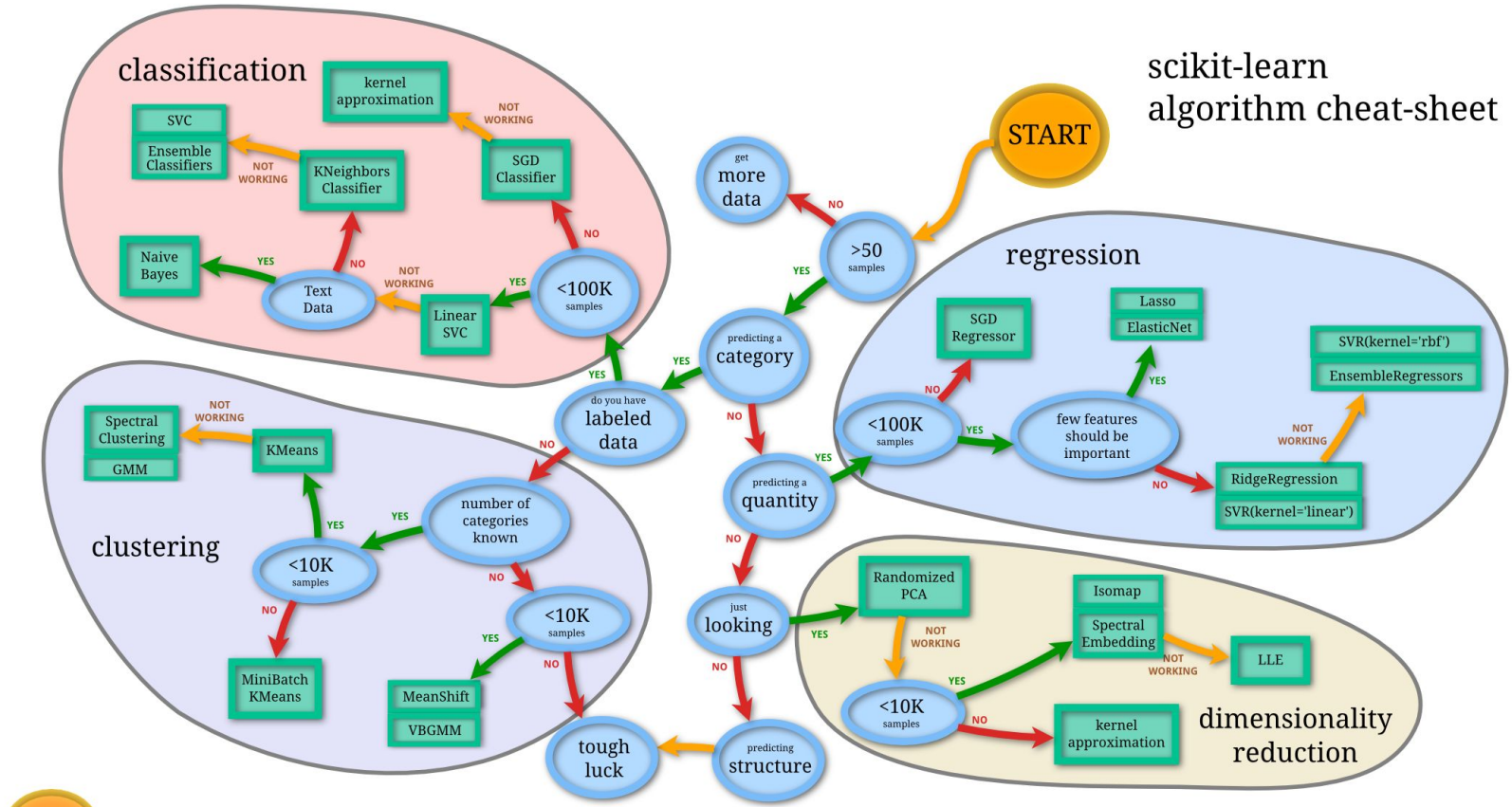
Reinforcement Learning

What is the most common application of reinforcement learning?



Semi-supervised learning

- Falls between supervised and unsupervised learning
- In this case only small amount of the dataset (training dataset) is labeled and a large amount of the data is unlabeled
- Many real world applications or problems are categorized under this category. [Why??]
- The process of building ML system using a semi-supervised learning method is usually require applying both unsupervised learning and supervised learning


scikit-learn
algorithm cheat-sheet
START
>50 samples
get more data
predicting a category
do you have labeled data
predicting a quantity
just looking
predicting structure
tough luck
classification
kernel approximation
SVC
Ensemble Classifiers
KNeighbors Classifier
SGD Classifier
Naive Bayes
Text Data
Linear SVC
<100K samples
clustering
Spectral Clustering
GMM
KMeans
number of categories known
<10K samples
MiniBatch KMeans
MeanShift
VBGMM
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
<100K samples
few features should be important
RidgeRegression
SVR(kernel='linear')
dimensionality reduction
Randomized PCA
Isomap
Spectral Embedding
LLE
<10K samples
kernel approximation
YES
NO
NOT WORKING
Back
scikit learn