Machine Learning and Scikit-Learn

CME451 Tutorial 8

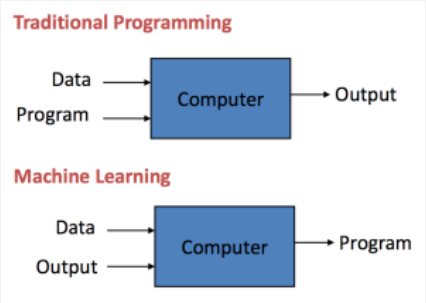
Hao Zhang
(Graduate Teaching Fellow)

Department of Electrical & Computer Engineering
University of Saskatchewan

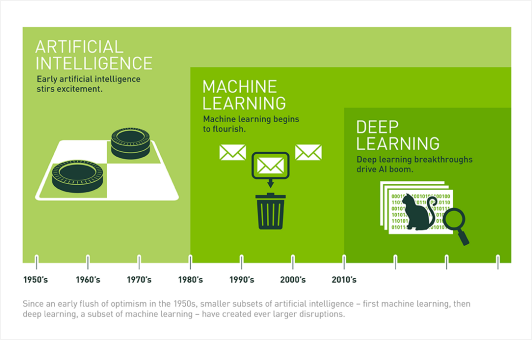
Mar 3, 2017

Overview of Machine Learning

- Machine Learning:
  - Exists everywhere.
  - Gives computers the ability to learn without being explicitly programmed.
  - Consider a set of samples of data (training).
  - Predict the properties of unknown data (testing/deploy).



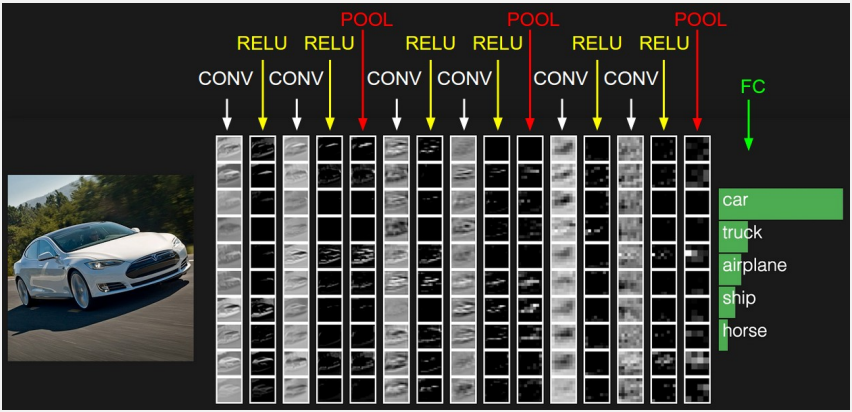
Overview of Machine Learning



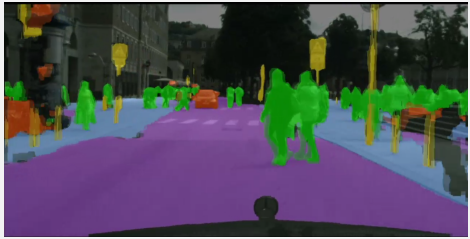
- Artificial Intelligence:
  - Human intelligence exhibited by machines.
- Machine Learning:
  - An approach to achieve artificial intelligence.
- Deep Learning:
  - A technique for implementing machine learning.

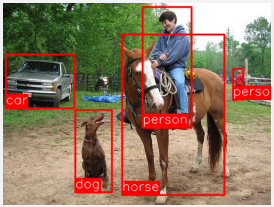
(* https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/)

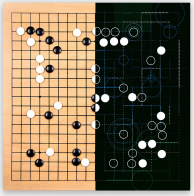
Overview of Machine Learning



Overview of Machine Learning



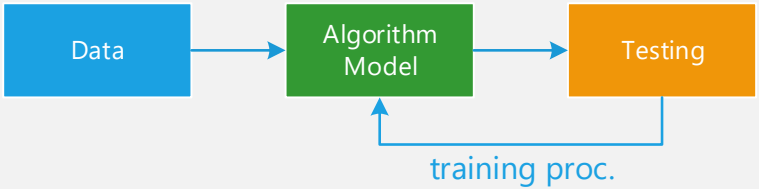




- Deep learning is everywhere:
  - Google DeepMind's AlphaGo program.
  - NVIDIA self-driving cars.
  - Natural images classification.
  - Natural language processing.
  - Biomedical and health informatics.
  - .....

Machine Learning Components

- Components and Process:

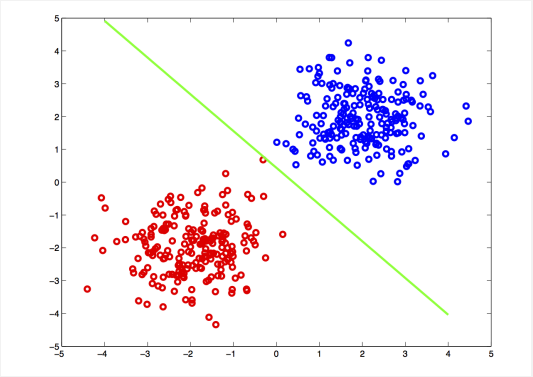


Training/Learning Process

- Basic categories:
  - **Supervised Learning**:
    - Data comes with additional attributes that we want to predict.
    - With ground truth value.
    - Training based on data + ground truth
  - **Unsupervised Learning**:
    - Training data without corresponding target values.
    - Training based on data only.
  - **Reinforcement Learning**:
    - Interacts with dynamic environment.

Supervised Learning

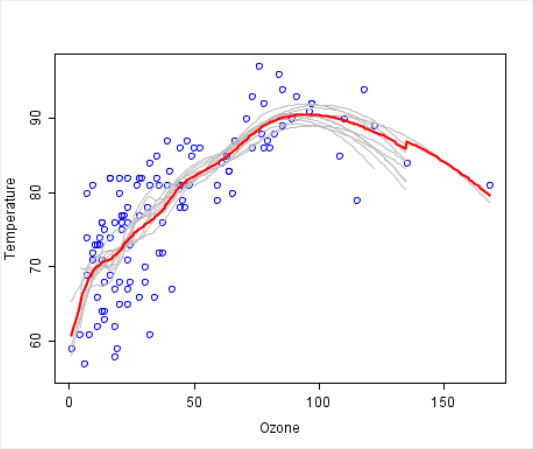
Classification



- For discrete data.
- Trained with samples of two or more classes.
- Generate the **classifier**.
- Classify the future data using this classifier.
- e.g., handwritten digits recognition.

Supervised Learning

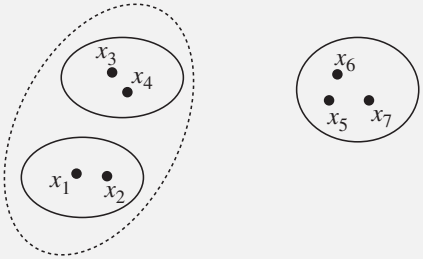
Regression



- For continuous data.
- Trained with samples.
- Generate the **trend**.
- Predict the future data using this trend function.
- e.g., temperature vs ozone.

Unsupervised Learning

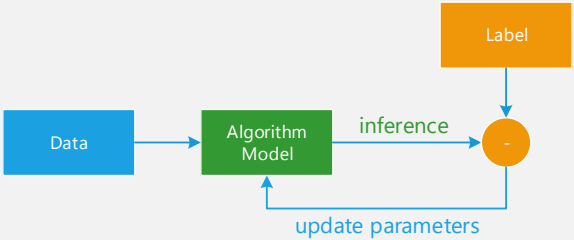
Clustering



- For data without label.
- Discover groups of similar examples.
- Generate the **clusters**.
- Used to determine the distribution of data.

Training Data

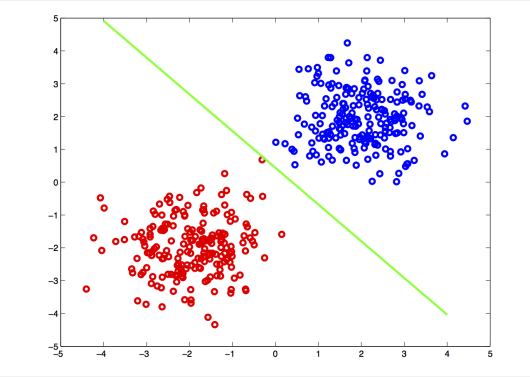
- We need large enough dataset to train a computer.
- For supervised learning, we need both data and label.
- Public datasets:
  - MNIST for handwritten digits.
  - ImageNet for large scale natural images.
- Create your own dataset with data samples and ground truth.



Learning Algorithm and Model

- Classifier based on Bayes decision theory.
  - Minimize the probability of error.
- Classifier based on Cost function optimization.
  - More complicated criteria.

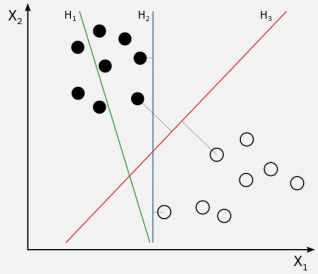
Euclidean Distance Classifier

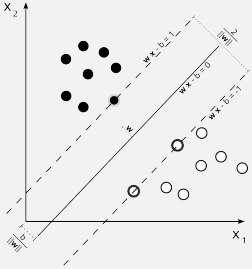


- Based on euclidean distance.
- Assign x to class i, if

$$||x - m_i|| < ||x - m_j|| \tag{1}$$

Support Vector Machine (SVM) Classifier

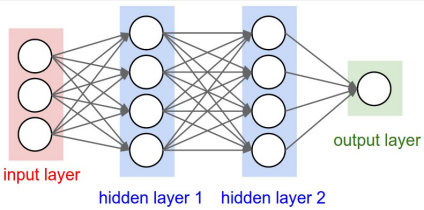




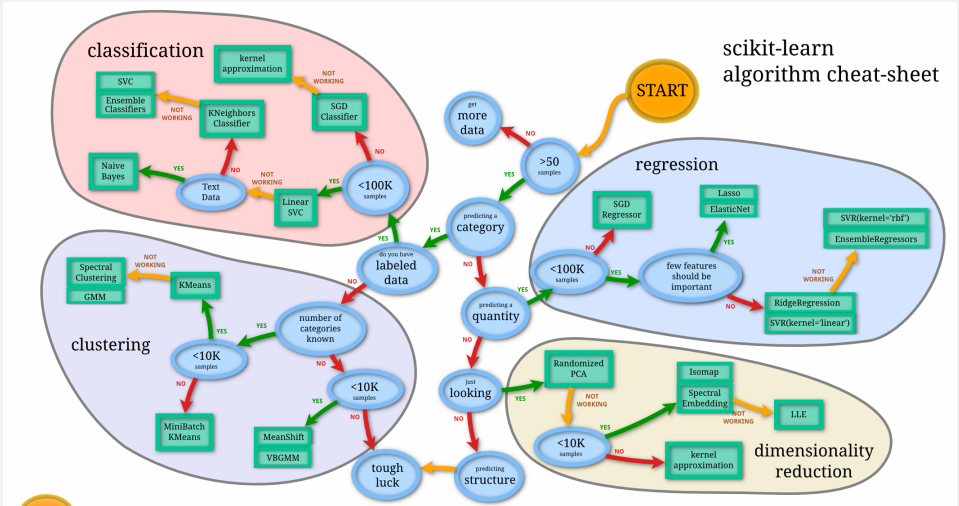
- Not only minimize the error probability.
- But also seek to maximize the margin.
- Increase robustness.

Neural Network

- **Artificial Neural Network**:
  - Small amount of layers.
  - No connection for neurons in the same layer.
  - Fully connected between neurons in adjacent layers.
  - Used to solve simple classification tasks.



Machine Learning in Python



Scikit-Learn

- Installation:

```
>>> pip install -U scikit-learn
or
>>> conda install scikit-learn
```

- Official Site:
  `http://scikit-learn.org/stable/`

Scikit-Learn

Load Dataset

- `Scikit-Learn` comes with few standard datasets:

```
>>> from sklearn import datasets
>>> digits = datasets.load_digits()
# data is in 2D format (num_sample, num_feature)
>>> print(digits.data)
[[  0.   0.   5. ...,   0.   0.   0.]
 [  0.   0.   0. ...,  10.   0.   0.]
 [  0.   0.   0. ...,  16.   9.   0.]
 ...,
 [  0.   0.   1. ...,   6.   0.   0.]
 [  0.   0.   2. ...,  12.   0.   0.]
 [  0.   0.  10. ...,  12.   1.   0.]]
# the ground truth
>>> digits.target
array([0, 1, 2, ..., 8, 9, 8])
```

Scikit-Learn

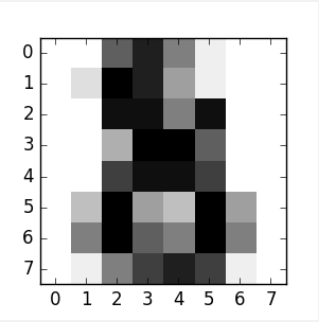
Learning Process

```
>>> from sklearn import svm
# create classifier
>>> clf = svm.SVC(gamma=0.001, C=100.)
# train the classifier
>>> clf.fit(digits.data[:-1], digits.target[:-1])
SVC(C=100.0, cache_size=200, class_weight=None, coef0=0.0,
decision_function_shape=None, degree=3, gamma=0.001, kernel='rbf',
max_iter=-1, probability=False, random_state=None, shrinking=True,
tol=0.001, verbose=False)
```

Scikit-Learn

Testing

```
>>> clf.predict(digits.data[-1:])
array([8])
```



Summary

- For machine learning:
  - Training dataset/Testing dataset
  - Learning algorithm/Model
  - Training the model
  - Testing the model
- Use `Scikit-Learn` to implement machine learning algorithm.