Introduction to Machine Learning

Wenting Tu

SHUFE, SIME

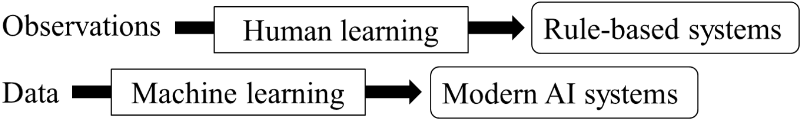
Outline

What is Machine Learning

Practical Applications of Machine Learning

What is Machine Learning

- Rule-based systems VS Learning-based systems

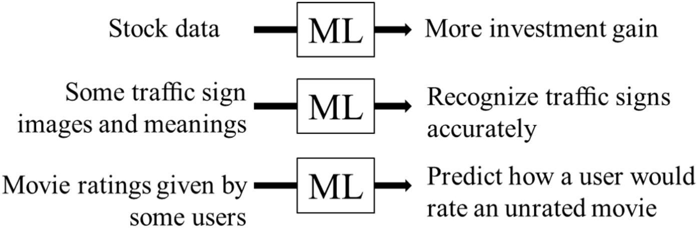


- Issues with rule-based systems
  - · Very labor intensive to build.
  - · Only work very well for areas they cover.
  - · Don’t naturally handle uncertainty.

  Disappointment in expert systems (late 80s / early 90s) led to an “AI Winter”.

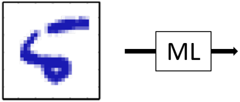
Formal Definition

A computer program is said to learn from ***experience E*** with respect to ***some class of tasks T*** and ***performance measures P***, if its performance at tasks in ***T***, as measured by ***P***, improved with ***E***.



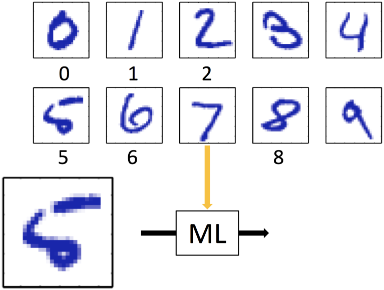
Terminologies

- **About *Data***
  - Instance / Example
  - Attribute / Feature
  - Feature vector
  - Feature space / Input space
  - Label
  - Label space / Output space



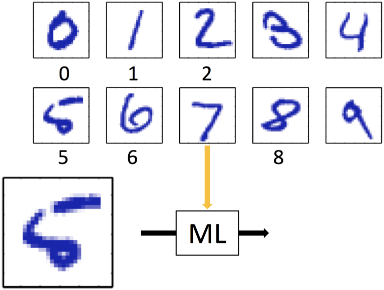
Terminologies

- **About *session***
  - Training/Learning
  - Testing



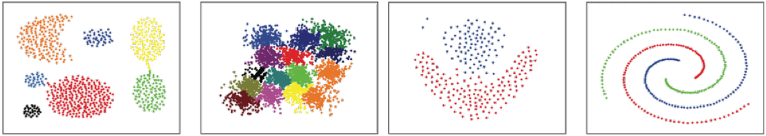
Terminologies

- **About *task***
  - Classification
  - Regression



Terminologies

- **About *task***
  - Clustering



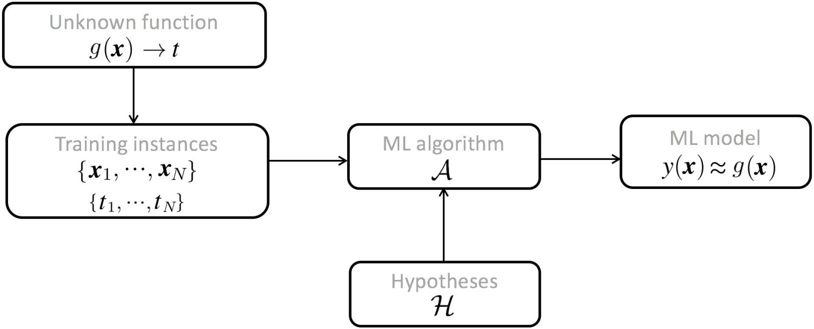
Terminologies

- **About *learning***
  - Supervised learning
  - Unsupervised learning
  - Semi-supervised learning
  - Transfer learning
  - Life-long learning

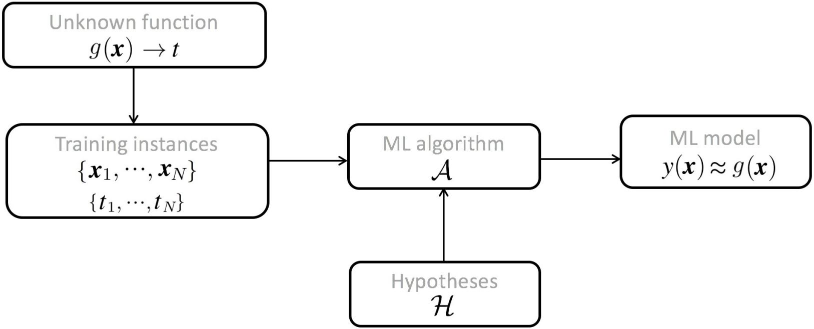
Terminologies

- **About *learning***
  - Reinforcement learning

How (supervised) ML works



Empirical risk minimization



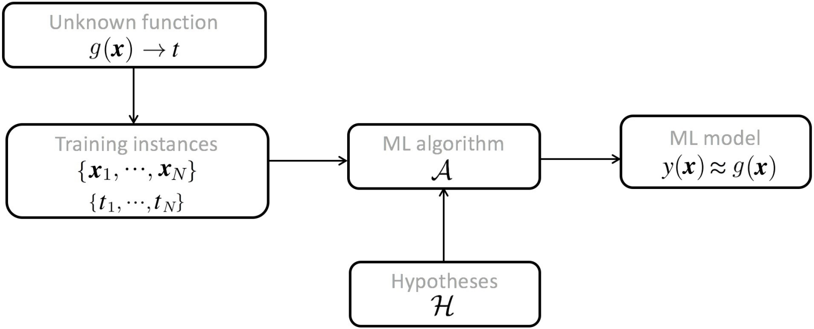
- Loss function

  $L(t, y(\boldsymbol{x}))$

- Expected risk

  $\mathbb{E}[L] = \iint L(t, y(\boldsymbol{x}))p(\boldsymbol{x}, t)\mathrm{d}\boldsymbol{x}\mathrm{d}t$

Empirical risk minimization



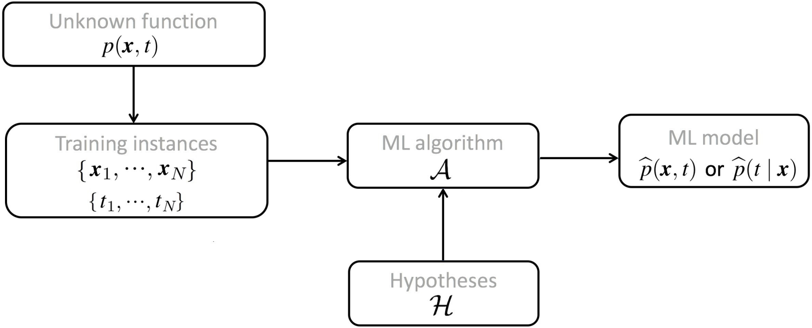
- Empirical risk

$$E = \sum_{n=1}^{N} L(t_n, y(\boldsymbol{x_n}))$$

- Empirical risk minimization

$$y^* = \arg\min_y E$$

Decision Theory (for Regression)



$$L(t, y(\boldsymbol{x})) = \{y(\boldsymbol{x}) - t\}^2$$

$$\mathbb{E}[L] = \iint \{y(\boldsymbol{x}) - t\}^2 p(\boldsymbol{x}, t) \mathrm{d}\boldsymbol{x} \mathrm{d}t$$

$$\frac{\partial \mathbb{E}[L]}{\partial y(\boldsymbol{x})} = 2 \int \{y(\boldsymbol{x}) - t\} p(\boldsymbol{x}, t) \mathrm{d}t = 0$$

$$y(\boldsymbol{x}) = \frac{\int t p(\boldsymbol{x}, t) \mathrm{d}t}{p(\boldsymbol{x})} = \int t p(t \mid \boldsymbol{x}) \, \mathrm{d}t = \mathbb{E}_t[t \mid \boldsymbol{x}]$$

Decision Theory (for Classification)

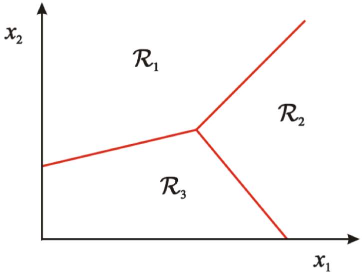
- Binary and multiclass classification

  $\{\mathcal{C}_k\} = \{-1, +1\}$

  $\{\mathcal{C}_k\} = \{1, \ldots, K\}$

- Decision Regions and Boundaries

  $\mathcal{R}_k = \{\boldsymbol{x} | y(x) \rightarrow \mathcal{C}_k\}$



Decision Theory (for Classification)

- Optimal decision for binary classification

$$\begin{aligned} p(\text{mistake}) &= p\left(\boldsymbol{x} \in \mathcal{R}_1, \mathcal{C}_2\right) + p\left(\boldsymbol{x} \in \mathcal{R}_2, \mathcal{C}_1\right) \\ &= \int_{\mathcal{R}_1} p\left(\boldsymbol{x}, \mathcal{C}_2\right) \mathrm{d}\boldsymbol{x} + \int_{\mathcal{R}_2} p\left(\boldsymbol{x}, \mathcal{C}_1\right) \mathrm{d}\boldsymbol{x} \end{aligned}$$

$$\mathcal{R}_1 = \left\{\boldsymbol{x} | p\left(\boldsymbol{x}, \mathcal{C}_1\right) > p\left(\boldsymbol{x}, \mathcal{C}_2\right)\right\}$$

$$\mathcal{R}_2 = \left\{\boldsymbol{x} | p\left(\boldsymbol{x}, \mathcal{C}_1\right) \leqslant p\left(\boldsymbol{x}, \mathcal{C}_2\right)\right\}$$

$$p\left(\boldsymbol{x}, \mathcal{C}_k\right) = p\left(\mathcal{C}_k | \boldsymbol{x}\right) p(\boldsymbol{x})$$

$$\mathcal{R}_k = \left\{\boldsymbol{x} \,| p\left(\mathcal{C}_k | \boldsymbol{x}\right) \text{ is largest } \right\}$$

Decision Theory (for Classification)

- Optimal decision for multiclass classification

$$p(\text{mistake}) = \sum_{k=1}^{K} \sum_{j \neq k} p\left(\boldsymbol{x} \in \mathcal{R}_j, \mathcal{C}_k\right) = \sum_{k=1}^{K} \sum_{j \neq k} \int_{\mathcal{R}_j} p\left(\boldsymbol{x}, \mathcal{C}_k\right) \mathrm{d}\boldsymbol{x}$$

$$p(\text{correct}) = \sum_{k=1}^{K} p\left(\boldsymbol{x} \in \mathcal{R}_k, \mathcal{C}_k\right) = \sum_{k=1}^{K} \int_{\mathcal{R}_k} p\left(\boldsymbol{x}, \mathcal{C}_k\right) \mathrm{d}\boldsymbol{x}$$

$$\mathcal{R}_k = \left\{\boldsymbol{x} \left| p\left(\mathcal{C}_k | \boldsymbol{x}\right) \text{ is largest } \right.\right\}$$

Decision Theory (for Classification)

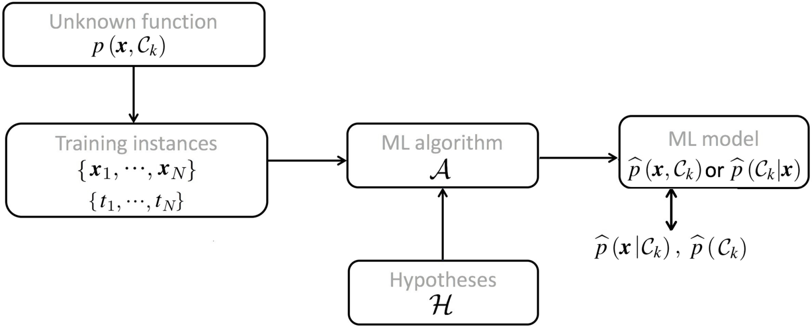
- The role of class posterior probability

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p\left(\boldsymbol{x}, \mathcal{C}_k\right) \mathrm{d}\boldsymbol{x}$$

$$\mathcal{R}_j = \left\{ \boldsymbol{x} \,\middle|\, \sum L_{kj} p\left(\mathcal{C}_k | \boldsymbol{x}\right) \text{ is smallest } \right\}$$

Decision Theory (for Classification)

- The role of class posterior probability



Outline

Practical Applications of Machine Learning

- Machine Learning for Finance
- Machine Learning for Medical Diagnosis
- Machine Learning for Education
- Machine Learning for Transportation
- Machine Learning for Internet
- $\cdots$

Thanks

Some images and slides are from the internet.
If related to copyright, please contact me.

tu.wenting@mail.shufe.edu.cn