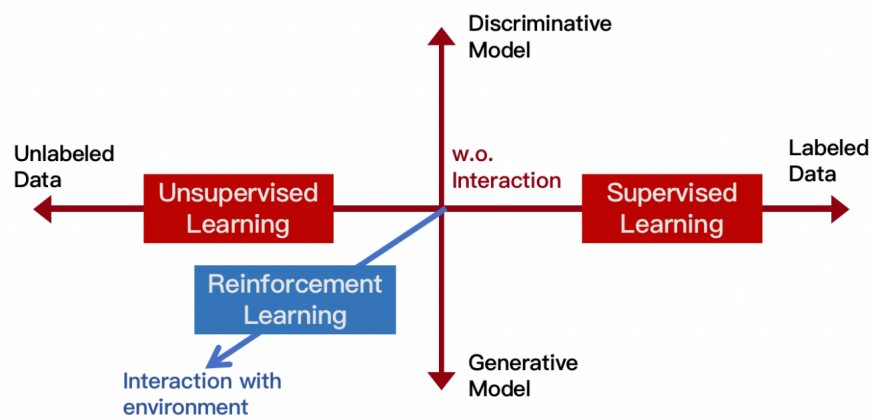
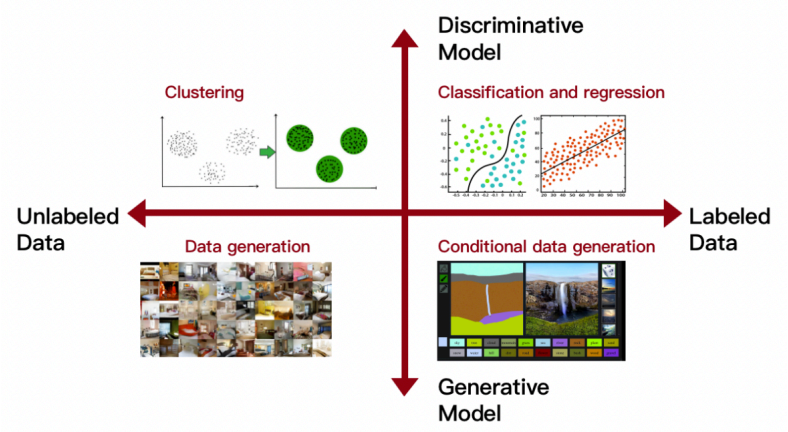
Lecture 1: Introduction to Machine Learning

Shukai Gong

1 Typical Paradigms of ML





- **Labeled Data + Discriminative Model:** Face recognition
- **Unlabeled Data + Discriminative Model:** Face clustering
- **Labeled Data + Generative Model:** Face interpolation / Conditional Manipulation
- **Unlabeled Data + Generative Model:** Face random generation

2. Key Factors in ML:

1. Computable **objective functions** (quantitative evaluation metric)
2. Computable **data representation** (vectors, matrices, etc.)
3. **Inference models** taking data representation as input.
4. Effective and efficient **learning algorithms**.

3. Aims in ML:

1. **Optimize** a performance criterion using data or past experience
2. **Generalize** to unseen data

2 Data Representation

Three common practices of data representation:

1. **Element-wise Representation:** $\boldsymbol{x} = [x_i]$ where $\boldsymbol{a}_i \in \mathbb{R}^N$,$\boldsymbol{A} = [a_{ij}]$
2. **Column-wise Representation:** $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_N]$ where $\boldsymbol{a}_i \in \mathbb{R}^N$
3. **One-hot:** Discrete data can be represented as $\boldsymbol{x} = [0, \cdots, 1, \cdots, 0]$

2.1 Space

1. **Sample Space:** $\mathcal{X}$ where $\boldsymbol{x} \in \mathcal{X}$
2. **Metric Space:** $(\mathcal{X}, d_{\mathcal{X}})$ where $d_{\mathcal{X}}$ is a distance of metric of samples.
3. **Probability Measure Space:** $\mathbb{P}_{\mathcal{X}}$ satisfying

$$\mathbb{P}_{\mathcal{X}} = \{\mu : \int_{x \in \mathcal{X}} \mu(x) \mathrm{d}x = 1, \mu(x) \geq 0, \forall x \in \mathcal{X}\}$$

   - $\mu \in \mathbb{P}_{\mathcal{X}}$ is a probability measure on $\mathcal{X}$

4. **Metric-Measure Space:** $\mathcal{X}_{d_{\mathcal{X}}, \mu_{\mathcal{X}}} := (\mathcal{X}, d_{\mathcal{X}}, \mu_{\mathcal{X}})$
   One example of MM-space can be a 2-dimensional space with a Gaussian distribution, where $\mathcal{X} = \mathbb{R}^2$, $d_{\mathcal{X}}$ is the Euclidean distance, and $\mu_{\mathcal{X}}$ is the Gaussian distribution.

   - Most ML tasks correspond to **reconstruct the MM-space** from observed data: Given $\{x\} \subset \mathcal{X}$
     - **Data representation:** Find a map $f : \mathcal{X} \to \mathcal{Z}_{d_{\mathcal{Z}}, \mu_{\mathcal{Z}}}$
     - **Metric learning:** Learn a metric $d_{\mathcal{X}}$
     - **Estimate $\mu_{\mathcal{X}}$:** Learn a density estimator $\hat{\mu}_{\mathcal{X}}$
   - **Vectorized data representation** often leads to a good metric space - Euclidean space.

3 A Basic Paradigm of ML

As preparation, we have

- **Data:** $X = \{\boldsymbol{x}_i\}_{i=1}^N \subset \mathcal{X}$, optionally with labels $Y = \{y_i\}_{i=1}^N \subset \mathcal{Y}$
- **A loss function:** $L : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$
- **A model with parameter** M_θ

Our **training task** is to

$$\min_{\theta \in \Omega} \sum_{i=1}^{N} L(M_\theta(\boldsymbol{x}_i), y_i)$$

where $f(\theta) = \sum_{i=1}^{N} L(M_\theta(\boldsymbol{x}_i), y_i)$ is the **objective function** of the variable θ and Ω is the **feasible domain**.

(Ex.)As preparation, we have

- **Data:** $x_n \in \mathbb{R}^2$: the sale and the revenue of the n-th company, $y_n \in \mathbb{R}$: the stock value of the n-th company
- **Model:** A **linear regression model** with Gaussian noise

$$\boldsymbol{y} = \boldsymbol{x}^\top \theta + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$$

- **Loss function:** MSE: $L(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2$

Our **training task** is to

$$\min_{\theta \in \Omega} \sum_{i=1}^{N} \|\boldsymbol{y}_i - \boldsymbol{x}_i^\top \theta\|_2^2$$

3.1 Keypoints of the Learning Problem

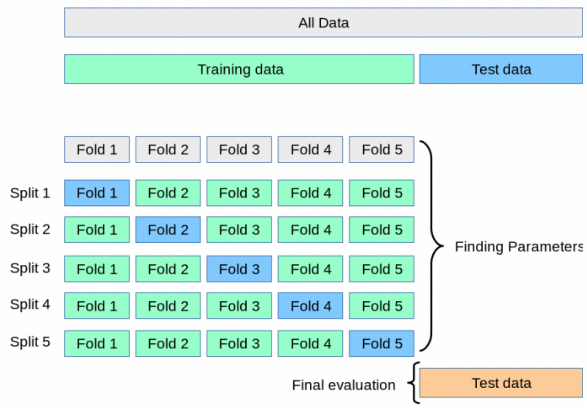
1. **Data Processing:** data processing is conducted to suppress the unfairness of features.
   - **Shifting and Scaling:** Standardize x into $\dfrac{x - \mu}{\sigma}$.
     The basic idea is to make each feature has zero mean and unit variance.
   - **Whitening:** Given $\boldsymbol{X} = [\boldsymbol{x_1}, \cdots, \boldsymbol{x_D}]$ with D features, and the estimate covariance matrix measuring the covariance between feature i and j is:

     $$\hat{\boldsymbol{\Sigma}} = \frac{1}{N-1}(\boldsymbol{X} - \boldsymbol{1_N}\hat{\boldsymbol{\mu}}^\top)^\top(\boldsymbol{X} - \boldsymbol{1_N}\hat{\boldsymbol{\mu}}^\top)$$

     where $\boldsymbol{1}_N$ is a N-dimensional vector with all elements equal to 1. Then we can whiten the data by

     $$\tilde{\boldsymbol{X}} = (\boldsymbol{X} - \boldsymbol{1_N}\hat{\boldsymbol{\mu}}^\top)\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}$$

     The basic idea is to make each feature zero-mean, unit-variance and uncorrelated to each other.
2. **Evaluation:**
   - **Loss Function:** $MSE = |y - \hat{y}|^2, MAE = |y - \hat{y}|$
   - **Cross Validation:**



3. **Training:** Model Selection

Akaike Information Criterion(AIC)

To estimate the relative amount of information **lost** by a given model, and achieve a trade-off between **good-of-fitness** and **model simplicity**. Suppose that we have a statistical model of some data $\boldsymbol{X}$. Let K be the number of model parameters and $\hat{L} = \max P(\boldsymbol{X}|\hat{\boldsymbol{\theta}})$ be the maximum likelihood for the model, then

$$AIC = 2K - 2\log \hat{L}$$

Given M models and their AIC values $\{AIC_m\}_{m=1}^{M}$, the relative likelihood of model m is $\exp\left(\frac{AIC_{\min} - AIC_m}{2}\right)$. The smaller the distance between $AIC_{\min}$ and AIC_m, the less information is lost by the model m compared to the best model.

Bayesian Information Criterion(BIC)

Suppose that we have a statistical model of some data $\boldsymbol{X}$. Let K be the number of model parameters, N be the number of data points, and $\hat{L} = \max P(\boldsymbol{X}|\hat{\boldsymbol{\theta}})$ be the maximum likelihood for the model, then

$$BIC = K\log N - 2\log \hat{L}$$

For example, in Polynomial Regression, K = polynomial order, and the N = data points size, $L = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}\exp\left(-\frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{2\sigma^2}\right)$ (Normally we omit the coefficients to make it clearer)