SJTU ML Course Overview

Unsupervised Learning

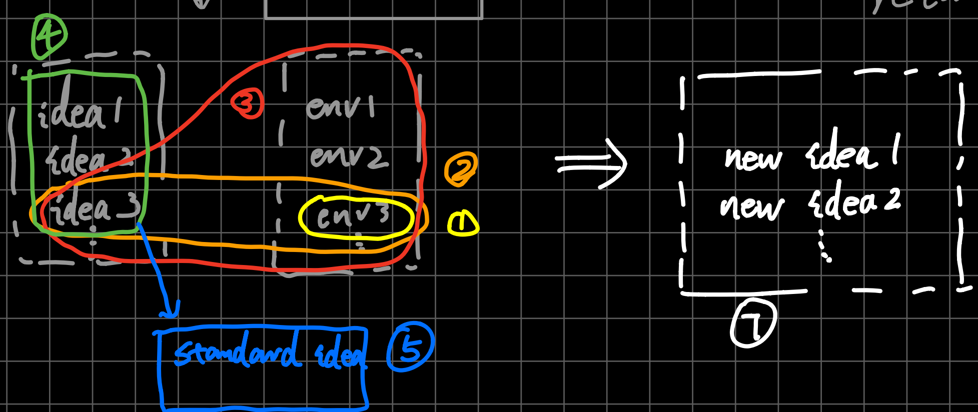
- 聚类分析 clustering
  - K均值聚类 (k-means)
  - 层次聚类 (Hierarchical)
  - 高斯混合模型聚类 (GMM)
- 最大似然估计(MLE)+EM算法
- 线性模型
  - PCA
  - ICA
  - FA
- 学习理论
  - Bias-Variance
  - Bayesian method
  - CL
    - CCL
    - FSCL
    - RPCL
- 模型选择
  - AIC
  - BIC
- 监督学习
  - Linear Regression
  - SVM
  - 神经网络
    - Perceptrons
    - BP
  - DL
    - AE.VAE
    - GAN
    - LMSER
    - Dense-Net.Resnet.CNN.RNN
- 因果发现

What to learn in this course?
数学理论和具体数学推导 😊

Bloom's Taxonomy (布伦分类法)

		action	object
High level recognition	create	create	new, original work
	evaluate	justify	standard
	analyse	build	connections among ideas
low level recognition	apply	use	original ideas in new environment
	understand	explain	concepts, facts
	remember	recall	concepts, facts

explaination ⟹



- Dimensionality Reduction
  - Feature Selection
    - Random Forest
    - Variance
  - Feature Projection
    - PCA
    - LDA
    - t-SNE
    - UMAP
  - Feature Learning
    - VAE (implement by myself)
    - VAE (existing package)

Gaussian distribution:

$$N(x|u,\Sigma) = \frac{1}{\sqrt{(2\pi)^D \cdot \Sigma}} \cdot e^{\left(-\frac{1}{2}\cdot (x-u)^T \Sigma^{-1} (x-u)\right)}$$

性质:

- $E(X) = u$
- $E(X \cdot X^T) = uu^T + \Sigma \xRightarrow{\text{Tips}} E(X^2) - E^2(X) = \Sigma$
- $\operatorname{cov}(X) = E\left((X-u)\cdot(X-u)^T\right) = \Sigma$

Bayes Theorem:

Condition Prob: $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)\cdot P(A)}{P(B)}$

$\Downarrow$

Product rule: $P(A,B) = P(A\cap B) = P(A|B)\cdot P(B)$

$\Downarrow$

Sum rule: $P(B) = \sum_A P(B,A)$

$$= \sum_A P(B|A)\cdot P(A)$$

Distance Metrics:

① Manhattan distance: $d_{fg} = \underbrace{\sum_c}_{\text{每分量}} |e_{fc} - e_{gc}|$

② Euclidean distance: $d_{fg} = \sqrt{\sum_c (e_{fc} - e_{gc})^2}$

③ Mahalanobis distance: $d_{fg} = \sqrt{(e_f - e_g)^T \cdot \underbrace{\Sigma^{-1}}_{\operatorname{cov}(X)} \cdot (e_f - e_g)}$

Where Gaussian Distribution comes from?

$\Rightarrow$ more powerful to convert distance to probability

distance — likely — Probability

$$(x-u)^2 \xrightarrow[\text{likely}]{\text{closer} \Leftrightarrow \text{more}} e^{-\lambda \cdot (x-u)^2} \xrightarrow{\text{intergrate}} \text{Gaussian distribution}$$

$$\frac{1}{\sqrt{(2\pi)^p \Sigma}} \cdot e^{\underbrace{-\frac{1}{2}}_{\lambda} \cdot \underbrace{D_M^2(x)}_{\text{Distance}}}$$

$D_M(x) = \sqrt{(x-u)^T \cdot \Sigma^{-1} \cdot (x-u)}$ $\Rightarrow$ Mahalanobis distance

GMM (Gaussian Mixture Model) $\Rightarrow$ 如何得到的

For point x, define $\vec{z} = (z_1, z_2 \cdots z_k)$, $\sum_{i=1}^{k} z_i = 1$

(1 行仅一个 z_i 值为 1，表明 x 属于 k 个 cluster 中哪一个)

$P(x \mid z_k = 1) = \mathcal{N}(x \mid u_k, \Sigma_k) \Rightarrow x \in$ cluster k,

each cluster is prepresented as a Gaussian distribution

$\Downarrow$

此前在 K-means 中被表示为 均值点 u_k.

Mixing weight for each cluster:

$P(z_k = 1) = \pi_k$, $\sum_{k=1}^{K} \pi_k = 1$

根据 Sum rule:

$$p(x) = \sum_{k=1}^{K} \underbrace{p(z_k=1)}_{\text{属于 cluster k 的概率，即权重}} \cdot p(x \mid z_k=1)$$

$$= \sum_{k=1}^{K} \pi_k \cdot N(x \mid u_k, \Sigma_k)$$

$$p(X) = p(x_1) \cdot p(x_2) \cdots p(x)$$

Goal of GMM:

maximize $p(X \mid u, \Sigma, \pi) \Rightarrow$

Given data set $X = (x_1, x_2 \cdots x_N)^T$，通过调整 Gaussian Distribution $\underbrace{u, \Sigma}$ 和 Mixing Weight π，

$(u_1, u_2 \cdots u_k)$ $(\Sigma_1, \Sigma_2 \cdots \Sigma_k)$

来 maximize the likelihood.

$\Downarrow$

等效于减少 distance，此为 clustering 的目的

How to maximize $\Rightarrow$ 求导

$$\frac{\partial \ln p(X \mid u, \Sigma)}{\partial u} = 0 \Rightarrow u_{ML} = \frac{\sum_{n=1}^{N} x_n}{N}$$

$$\frac{\partial \ln p(X \mid u, \Sigma)}{\partial \Sigma} = 0 \Rightarrow \Sigma_{ML} = \frac{1}{N} \cdot \sum_{n=1}^{N} (x_n - u_{ML})(x_n - u_{ML})^T$$

Matrix-cook-book

$$
\begin{aligned}
\partial \mathbf{A} &= 0 \qquad (\mathbf{A} \text{ is a constant}) \\
\partial(\alpha \mathbf{X}) &= \alpha \partial \mathbf{X} \\
\partial(\mathbf{X}+\mathbf{Y}) &= \partial \mathbf{X} + \partial \mathbf{Y} \\
\partial(\mathrm{Tr}(\mathbf{X})) &= \mathrm{Tr}(\partial \mathbf{X}) \\
\partial(\mathbf{X}\mathbf{Y}) &= (\partial \mathbf{X})\mathbf{Y} + \mathbf{X}(\partial \mathbf{Y}) \\
\partial(\mathbf{X} \circ \mathbf{Y}) &= (\partial \mathbf{X}) \circ \mathbf{Y} + \mathbf{X} \circ (\partial \mathbf{Y}) \\
\partial(\mathbf{X} \otimes \mathbf{Y}) &= (\partial \mathbf{X}) \otimes \mathbf{Y} + \mathbf{X} \otimes (\partial \mathbf{Y}) \\
\partial(\mathbf{X}^{-1}) &= -\mathbf{X}^{-1}(\partial \mathbf{X})\mathbf{X}^{-1} \\
\partial(\det(\mathbf{X})) &= \det(\mathbf{X})\mathrm{Tr}(\mathbf{X}^{-1}\partial \mathbf{X}) \\
\partial(\ln(\det(\mathbf{X}))) &= \mathrm{Tr}(\mathbf{X}^{-1}\partial \mathbf{X}) \\
\partial \mathbf{X}^T &= (\partial \mathbf{X})^T \\
\partial \mathbf{X}^H &= (\partial \mathbf{X})^H
\end{aligned}
$$

⇒利用负指数求导法则来记

⇒ 共轭转置

$A^H = (\overline{A})^T$，先共轭，再转置