Types of Clustering

- A clustering is a process to find a set of clusters
- Important distinction between **hierarchical** and **partitional** sets of clusters
- **Partitional** Clustering ← K-means
  - A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
- **Hierarchical** clustering ←
  - A set of nested clusters organized as a hierarchical tree
- **Density based** clustering → EM
  - Discover clusters of **arbitrary** shape.
  - Clusters **dense regions** of objects **separated by regions of low density**

K-means Recap ...

- Randomly initialize k centers
  - $\mu^{(0)} = \mu_1^{(0)}, \mu_2^{(0)}, \mu_3^{(0)}, \dots, \mu_k^{(0)}$
- Interate t=0, 1, 2, . . .
  - **Classify**: Assign each point j∈{1,2,…,m} to nearest center $C^{(t)}(j) \leftarrow \arg\min_{i=1,\dots,k} \|\mu_i^{(t)} - x_j\|^2$
- **Recenter**: μ_i becomes centroid if its points: $\mu_i^{(t+1)} \leftarrow \arg\min_{\mu} \sum_{j:C^{(t)}(j)=i} \|\mu - x_j\|^2 \quad i \in \{1,\dots,k\}$
- Equivalent to μ_i ← average of its points!

What is K-means optimizing? (Objective Function)

- **Potential function** (objective function) F(μ,C) of centers μ and point allocation C:

$$F(\mu, C) = \sum_{j=1}^{m} \|\mu_{C(j)} - x_j\|^2 = \sum_{i=1}^{k} \sum_{j:C(j)=i} \|\mu_i - x_j\|^2$$

- Optimal K-means:
  - $\min_\mu \min_C F(\mu, C)$

EM / GMM

$$\hat{\theta} = \arg\max_\theta E_z\left[\sum_{i=1}^{n} \log P(x_i, z_i|\theta)\right]$$

complete Log Likelihood $L(\theta)$

$\boxed{Q(\theta) = E_z[L(\theta)]}$ E-step

$\frac{dQ(\theta)}{d\theta} = 0 \quad \max_\theta Q(\theta)$ M-step

} EM

$$P(x|\theta) = \sum_{k=1}^{K} \alpha_k N(x|\mu_k, \Sigma_k) \qquad \sum_{k=1}^{K} \alpha_k = 1 \quad \alpha_k \geq 0$$

$D = \{x_1, x_2, x_3, \dots, x_n\}$

$H = \{z_1, z_2, z_3, \dots, z_n\}$ $\quad z_i = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \end{bmatrix} \to k$, $K \times 1$

$$P(x_i, z_i|\theta) = \prod_{k=1}^{K} \left(\alpha_k N(x_i|\mu_k, \Sigma_k)\right)^{z_{ik}}$$

$$= \cancel{\sum_{k=1}^{K} z_k \alpha_k N(x|\mu_k, \Sigma_k)}$$

$$L(\theta) = \log P(D, H|\theta) = \log P(x_1, z_1, x_2, z_2, \dots, x_n, z_n|\theta) = \log \prod_{i=1}^{n} P(x_i, z_i|\theta) = \sum_{i=1}^{n} \log P(x_i, z_i|\theta)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik}\left(\log \alpha_k + \log N(x_i|\mu_k, \Sigma_k)\right)$$

$$Q(\theta) = E_{z}[L(\theta)] = \sum_{i=1}^{n} \sum_{k=1}^{K} \underbrace{E[z_{ik}]}_{\gamma_{ik}^t} \left(\log \alpha_k + \log N(x_i|\mu_k, \Sigma_k)\right)$$

$$\underset{z_i|x_i}{E}[z_{ik}] = Pr(z_{ik} = 1 | x_i, \theta^t) = \frac{P(x_i|z_{ik}=1, \theta^t) P(z_{ik}=1|\theta^t)}{P(x_i|\theta^t)}$$

$$= \frac{N(x_i|\mu_k^t, \Sigma_k^t)\ \alpha_k^t}{\sum_{k=1}^{K} \alpha_k N(x_i|\mu_k^t, \Sigma_k^t)} = \gamma_{ik}^t$$

$$\frac{\partial Q(\theta)}{\partial \alpha_k} = 0$$

$$\boxed{\hat{\alpha}_k = \frac{\sum_{i=1}^{n} \gamma_{ik}^t}{n}} \qquad \boxed{\hat{\mu}_k = \frac{\sum_{i=1}^{n} \gamma_{ik}^t x_i}{\sum_{i=1}^{n} \gamma_{ik}^t}}$$

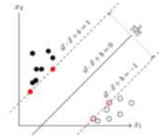
$$Q(\theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik}^t \left(\log \alpha_k + \log N(x_i|\mu_k, \Sigma_k)\right) - \lambda\left(\sum_{k=1}^{K} \alpha_k - 1\right)$$

$$\frac{\partial Q(\theta)}{\partial \alpha_j} = \sum_{i=1}^{n} \gamma_{ij}^t \frac{1}{\alpha_j} - \lambda = 0 \to \boxed{\alpha_j = \frac{\sum_{i=1}^{n} \gamma_{ij}^t}{\lambda}} \quad j = 1, \dots, K$$

$$1 = \sum_{j=1}^{K} \alpha_j = \frac{\sum_{j=1}^{K} \sum_{i=1}^{n} \gamma_{ij}^t}{\lambda}$$

$$\alpha_j = \frac{\sum_{i=1}^{n} \gamma_{ij}^t}{\sum_{j=1}^{K} \sum_{i=1}^{n} \gamma_{ij}^t} = \frac{\sum_{i=1}^{n} \gamma_{ij}^t}{n}$$

Hard SVM



$$\min_{w,b} \frac{1}{2}\|w\|_2^2$$

s.t. $y_i(w^T x_i + b) \geq 1 \quad i = 1, \dots, n$

$L(w, b, \lambda)$

$\nabla_w L = 0 \Rightarrow \boxed{w = \sum_{i=1}^{n} \lambda_i y_i x_i}$

$\nabla_b L = 0 \Rightarrow \sum_{i=1}^{n} \lambda_i y_i = 0$

$$\max_\lambda -\frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j x_i^T x_j + \sum_{i=1}^{n} \lambda_i$$

$\lambda_i \geq 0, \quad \sum \lambda_i y_i = 0$

Quadratic Programing

Soft SVM

Soft Margin SVM

$$\min_{w,b,\xi} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \xi_i$$

s.t. $y_i(w^T x_i + b) \geq 1 - \xi_i$, $\xi_i \geq 0$

$$L(w, b, \xi, \lambda, \mu) = \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \lambda_i \left(1 - \xi_i - y_i(w^T x_i + b)\right) + \sum_{i=1}^{n} \mu_i(-\xi_i)$$

$\nabla_{\xi} L = C - \lambda_i - \mu_i = 0 \Rightarrow \boxed{\mu_i = C - \lambda_i \geq 0} \Rightarrow \lambda_i \leq C$

$$\max_\lambda -\frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j x_i^T x_j + \sum_i \lambda_i$$

$\sum_{i=1}^{n} \lambda_i y_i = 0, \quad 0 \leq \lambda_i \leq C$

Contrained Optimization

$$\min_x f(x)$$

s.t. $g_i(x) \leq 0 \quad i = 1, \dots, m$

$h_j(x) = 0 \quad j = 1, \dots, p$

KKT conditions

Lagrangian: $L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \mu_j h_j(x)$

$\lambda_i \geq 0 \quad i = 1, \dots, m$

KKT:

1. $\nabla_x L(x, \lambda, \mu) = 0$ $\quad x^*, \lambda^*, \mu^*$
2. $h_j(x) = 0$, $g_i(x) \leq 0$
3. $\lambda_i \geq 0$
4. $\lambda_i g_i(x) = 0$ → Complementary Slackness

$$g(\mu, \lambda) = \min_x L(x, \mu, \lambda)$$

$$g(\mu, \lambda) \leq f(x^*)$$

Dual solution ≤ Primal solution

$\boxed{g(\lambda^*, \mu^*) \leq f(x^*)}$ weak duality

Strong Duality:

$\boxed{g(\lambda^*, \mu^*) = f(x^*)}$

1. Convex
2. $\boxed{\exists x \quad g_i(x) < 0}$ ← Slater's Condition

$$\min_{w,b} \frac{1}{2}\|w\|_2^2$$

s.t. $y_i(w^T x_i + b) \geq 1 \quad i = 1, \dots, n$ → $1 - y_i(w^T x_i + b) \leq 0$

$$L(w, b, \lambda) = \frac{1}{2}\|w\|_2^2 + \sum_{i=1}^{n} \lambda_i \left(1 - y_i(w^T x_i + b)\right)$$

$\frac{\partial L}{\partial w} = w + \sum_{i=1}^{n} -\lambda_i y_i x_i = 0 \Rightarrow \boxed{w = \sum_{i=1}^{n} \lambda_i y_i x_i}$

$\frac{\partial L}{\partial b} = -\sum_{i=1}^{n} \lambda_i y_i = 0 \Rightarrow \boxed{\sum_{i=1}^{n} \lambda_i y_i = 0}$

$$g(\lambda) = \frac{1}{2}\left(\sum_i \lambda_i y_i x_i\right)^T\left(\sum_i \lambda_i y_i x_i\right) + \sum_{i=1}^{n} \lambda_i - \sum_i \lambda_i y_i \left(\sum_j \lambda_j y_j x_j\right)^T x_i - \underbrace{\sum_i \lambda_i y_i b}_{0}$$

$$= \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j x_i^T x_j - \sum_i \sum_j \lambda_i \lambda_j y_i y_j x_i^T x_j + \sum_i \lambda_i$$

$$\Rightarrow \boxed{g(\lambda) = -\frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j x_i^T x_j + \sum_i \lambda_i} \quad \lambda_i \geq 0$$

$\boxed{\max_\lambda g(\lambda) \quad \text{s.t. } \lambda_i \geq 0}$ → $\lambda^* = \begin{bmatrix} \lambda_1^* \\ \vdots \\ \lambda_n^* \end{bmatrix}$

$\boxed{w^* = \sum_{i=1}^{n} \lambda_i^* y_i x_i}$

Complementary Slackness:

$\lambda_i^* g_i(x^*) = 0$

$$\boxed{\underbrace{\lambda_k^*}_{\neq 0}\left(y_k(w^{*T} x_k + b)\right) = 0}$$

Techniques for Constructing New Kernels.

Given valid kernels $k_1(\mathbf{x}, \mathbf{x}')$ and $k_2(\mathbf{x}, \mathbf{x}')$, the following new kernels will also be valid:

$k(\mathbf{x}, \mathbf{x}')$	=	$ck_1(\mathbf{x}, \mathbf{x}')$	(6.13)
$k(\mathbf{x}, \mathbf{x}')$	=	$f(\mathbf{x})k_1(\mathbf{x}, \mathbf{x}')f(\mathbf{x}')$	(6.14)
$k(\mathbf{x}, \mathbf{x}')$	=	$q(k_1(\mathbf{x}, \mathbf{x}'))$	(6.15)
$k(\mathbf{x}, \mathbf{x}')$	=	$\exp(k_1(\mathbf{x}, \mathbf{x}'))$	(6.16)
$k(\mathbf{x}, \mathbf{x}')$	=	$k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}')$	(6.17)
$k(\mathbf{x}, \mathbf{x}')$	=	$k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}')$	(6.18)
$k(\mathbf{x}, \mathbf{x}')$	=	$k_3(\phi(\mathbf{x}), \phi(\mathbf{x}'))$	(6.19)
$k(\mathbf{x}, \mathbf{x}')$	=	$\mathbf{x}^T\mathbf{A}\mathbf{x}'$	(6.20)
$k(\mathbf{x}, \mathbf{x}')$	=	$k_a(\mathbf{x}_a, \mathbf{x}_a') + k_b(\mathbf{x}_b, \mathbf{x}_b')$	(6.21)
$k(\mathbf{x}, \mathbf{x}')$	=	$k_a(\mathbf{x}_a, \mathbf{x}_a')k_b(\mathbf{x}_b, \mathbf{x}_b')$	(6.22)

where $c > 0$ is a constant, $f(\cdot)$ is any function, $q(\cdot)$ is a polynomial with nonnegative coefficients, $\phi(\mathbf{x})$ is a function from $\mathbf{x}$ to $\mathbb{R}^M$, $k_3(\cdot,\cdot)$ is a valid kernel in $\mathbb{R}^M$, $\mathbf{A}$ is a symmetric positive semidefinite matrix, $\mathbf{x}_a$ and $\mathbf{x}_b$ are variables (not necessarily disjoint) with $\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_b)$, and k_a and k_b are valid kernel functions over their respective spaces.

روش‌های اثبات اعتبار کرنل:
انتگرال مرسر
اثبات PSD بودن ماتریس K
نوشتن کرنل به فرم ضرب داخلی

Mercer Theorem

$$\iint g(x)K(x, y)g(y)\, dx\, dy \geq 0$$

AdaBoost

$h_1, \dots, h_T$

$$\boxed{F_T(x) = \sum_{t=1}^{T} \alpha_t h_t(x)}$$

Base Classifier / weak Learner

$\alpha_t = ? \quad t = 1, \dots, T$

$w_i = ?$

$w^{(0)}\ h_1\ \alpha_1 \to w^{(1)}\ h_2\ \alpha_2 \to w^{(2)}\ h_3\ \alpha_3 \to \dots \to w^{(T-1)}\ h_T\ \alpha_T$

$\alpha_1 h_1 + \alpha_2 h_2$

$$F_T(x) = \sum_{t=1}^{T} \alpha_t h_t(x) = F_{T-1}(x) + \alpha_T h_T(x) \qquad h_T \quad \alpha_T$$

$$E = \sum_{i=1}^{n} e^{-y_i F_T(x_i)} = \sum_{i=1}^{n} \underbrace{e^{-y_i F_{T-1}(x_i)}}_{w_i^{(T-1)}} e^{-y_i \alpha_T h_T(x_i)}$$

$$= \sum_{i=1}^{n} w_i^{(T-1)} e^{-y_i \alpha_T h_T(x_i)} = \sum_{\underbrace{i: y_i \neq h_T(x_i)}_{M}} w_i^{(T-1)} e^{\alpha_T} + \sum_{\underbrace{i: y_i = h_T(x_i)}_{C}} w_i^{(T-1)} e^{-\alpha_T}$$

$$\frac{dE}{d\alpha_T} = e^{\alpha_T} \sum_{i \in M} w_i^{(T-1)} - e^{-\alpha_T} \sum_{i \in C} w_i^{(T-1)} = 0$$

$$\to \boxed{\alpha_T = \frac{1}{2} \ln \frac{\sum_{i \in C} w_i^{(T-1)}}{\sum_{i \in M} w_i^{(T-1)}}}$$

$$\epsilon_T = \frac{\sum_{i \in M} w_i^{(T-1)}}{\sum_{i=1}^{n} w_i^{(T-1)}} \qquad h_T \qquad 0 \leq \epsilon_T \leq 1$$

$$\boxed{\alpha_T = \frac{1}{2} \ln \frac{1 - \epsilon_T}{\epsilon_T}}$$

$\boxed{\epsilon_T \neq \frac{1}{2}}$

$\epsilon_T = 0$

$\epsilon_T = 1$

Bagging

$1 - \frac{1}{e}$ → odd

① → h_1

⋮

Ⓣ → h_T

majority vote → $\hat{y}$

Bootstrap

Aggregation

PCA

$$\max_u \text{var}(z) \quad \text{s.t. } \|u\|_2^2 = 1$$

$\boxed{E[x] = 0}$ ✓

$z = u^T x \Rightarrow E[z] = E[u^T x] = u^T \underbrace{E[x]}_{0} = 0$

$$\text{var}(z) = E[z^2] - \underbrace{E^2(z)}_{0} \simeq \frac{1}{n} \sum_{i=1}^{n} z_i^2$$

$$\max_u \frac{1}{n} \sum_{i=1}^{n} z_i^2 = \frac{1}{n} \sum_{i=1}^{n} \underbrace{(x_i^T u)^2}_{(u^T x_i)(x_i^T u)} = \frac{1}{n} \sum_{i=1}^{n} u^T x_i x_i^T u$$

s.t. $\|u\|_2^2 = 1$

$$= u^T \underbrace{\left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i^T\right)}_{S:\ \text{Sample cov. Matrix}} u = u^T S u$$

$$\boxed{\max_u u^T S u \quad \text{s.t. } \|u\|_2^2 = 1} \to L(u) = u^T S u - \lambda\left(\underbrace{\|u\|_2^2}_{u^T u} - 1\right)$$

$\nabla_u L = 0 \Rightarrow 2Su - \lambda 2u = 0$

$\Rightarrow \boxed{Su = \lambda u}$ — u: eigenvector, λ: eigenvalue

$x \in \mathbb{R}^d$

$S_{d \times d}$

$$\max u^T S u = \max u^T \lambda u = \max \lambda \underbrace{u^T u}_{1} \neq \boxed{\max \lambda}$$

Dimensionality Reduction

Filter / wrapper

Dimensionality Reduction → Feature selection / Feature Reduction

Supervised / unsupervised

Sequential forward selection
Sequential backward selection

$I(x^i; y)$ Correlation

LDA

$$\max_w \frac{(\tilde{m}_1 - \tilde{m}_2)^2}{\tilde{s}_1^2 + \tilde{s}_2^2} \qquad \boxed{g(x) = \underset{?}{w^T} x}$$

$$\tilde{m}_1 = \frac{1}{n} \sum_{i=1}^{n} z_i = \frac{1}{n} \sum_{i=1}^{n} w^T x_i = w^T \underbrace{\left(\frac{1}{n} \sum_{i=1}^{n} x_i\right)}_{m_1} = w^T m_1$$

$\tilde{m}_2 = w^T m_2$

$$\tilde{s}_1^2 = \sum_{i=1}^{n} z_i z_i^T = \sum_{i=1}^{n} (w^T x_i)(w^T x_i)^T = \sum_i w^T x_i x_i^T w = w^T \underbrace{\left(\sum_{i=1}^{n} x_i x_i^T\right)}_{s_1^2} w = w^T s_1^2 w$$

$\tilde{s}_2^2 = w^T s_2^2 w$

$$J(w) = \frac{(w^T m_1 - w^T m_2)^2}{w^T s_1^2 w + w^T s_2^2 w} = \frac{w^T \overbrace{(m_1 - m_2)(m_1 - m_2)^T}^{S_B} w}{w^T \underbrace{(s_1^2 + s_2^2)}_{S_W} w}$$

S_W: within cluster scatter

S_B: Between cluster scatter

$$\boxed{\max_w J(w) = \frac{w^T S_B w}{w^T S_W w}} \qquad S_B = \underset{d \times 1}{(m_1 - m_2)} \underset{1 \times d}{(m_1 - m_2)^T}$$

Rayleigh Quotient

$$\to \boxed{\max_w w^T S_B w \quad \text{s.t. } w^T S_W w = k}$$

$$\max_w L(w) = w^T S_B w - \lambda(w^T S_W w - k) \qquad \lambda \geq 0$$

$2S_B w - 2\lambda S_W w = 0 \Rightarrow \boxed{S_B w = \lambda S_W w}$

$\boxed{S_W^{-1} S_B w = \lambda w}$

$\boxed{\lambda w = S_W^{-1} S_B w}$ →

$S_B = (m_1 - m_2)(m_1 - m_2)^T$

$$\lambda w = S_W^{-1} (m_1 - m_2)\underbrace{(m_1 - m_2)^T w}_{\alpha} \Rightarrow w = \frac{\alpha}{\lambda} S_W^{-1} (m_1 - m_2)$$

$$\Rightarrow \boxed{w^* = S_W^{-1}(m_1 - m_2)}$$