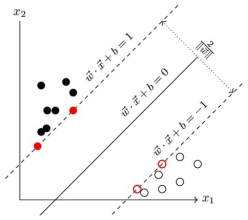
Hard SVM

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

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



$$L(w,b,\lambda)$$

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

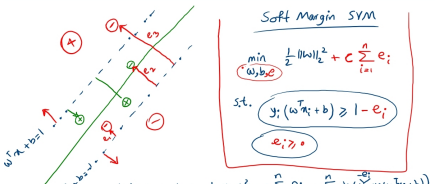
$$\nabla_b L = 0 \Rightarrow \sum_{i=1}^{n} \lambda_i y_i = 0$$

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

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

Quadratic Programming

Soft Margin SVM



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

$$s.t. \quad y_i(w^T x_i + b) \geq 1 - e_i$$

$$e_i \geq 0$$

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

dual var

$$\nabla_{e_j} L = C - \lambda_j - \mu_j = 0 \Rightarrow \mu_j = C - \lambda_j \geq 0 \Rightarrow 0 \leq \lambda_j \leq C$$

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

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

Soft SVM

Contrained Optimization

$$\min_{x} f(x)$$

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

$$h_i(x) = 0 \quad i=1,\dots,p$$

KKT conditions

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

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

KKT:

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

x^*, λ^*, μ^*

$$g(\mu,\lambda) = \min_{x} L(x,\mu,\lambda)$$

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

Dual solution ≤ primal solution

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

Strong Duality: $g(\lambda^*,\mu^*) = f(x^*)$

Slater's Condition: ① convex ② $\exists x \; g_i(x) < 0$

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:

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

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

$h_1 \dots h_T$

Base classifier / weak Learner

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

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

$\omega_i = ?$

$w^{(0)}$ h_1 (α_1) → $w^{(1)}$ h_2 (α_2) → $w^{(2)}$ h_3 (α_3) → … → $w^{(T-1)}$ h_T (α_T)

$\alpha_1 h_1 + \alpha_2 h_2$

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

h_T α_T

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

$w_i^{(T-1)}$

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

M C

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

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

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

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

$\epsilon_T \Rightarrow \frac{1}{2}$

$\epsilon_T = 0$

$\epsilon_T = 1$

Bagging

$1 - \frac{1}{e} = 0.66$

majority vote → $\hat{y}$

Bootstrap Aggregation

PCA

$$\max_{u} \; var(z)$$

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

$E[x] = 0$ ✓

$$z = u^T x \Rightarrow E[z] = E[u^T x] = u^T E[x] = 0$$

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

$$\max_{u} \; \frac{1}{n}\sum_{i=1}^{n} z_i^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i^T u)^2 = \frac{1}{n}\sum_{i=1}^{n} u^T x_i x_i^T u$$

$(u^T x_i)(x_i^T u)$

$$= u^T \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T\right) u$$

S: Sample cov. Matrix

$$= u^T S u$$

$$\max_{u} \; u^T S u \quad s.t. \quad \|u\|_2^2 = 1$$

$$L(u) = u^T S u - \lambda(\|u\|_2^2 - 1)$$

$u^T u$

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

$$\Rightarrow \boxed{Su = \lambda u}$$

eigenvector, eigenvalue

$x \in \mathbb{R}^d$, $S_{d \times d}$

$$\max \; u^T S u = \max \; u^T \lambda u = \max \; \lambda \underbrace{u^T u}_{1} \neq \max(1)$$

LDA

$$\max_{\omega} \; \frac{(\tilde{m}_1 - \tilde{m}_2)^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

$g(x) = \omega^T x$

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

m_1

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

$$\tilde{s}_1^2 = \sum_{i=1}^{n} z_i z_i^T = \sum_{i=1}^{n} (\omega^T x_i)(\omega^T x_i)^T = \sum_{i} \omega^T x_i x_i^T \omega$$

$$= \omega^T \left(\sum_{i=1}^{n} x_i x_i^T\right) \omega = \omega^T s_1^2 \omega$$

s_1^2

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

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

S_w: within cluster scatter

S_B: Between cluster Scatter

$$\max_{\omega} \; J(\omega) = \frac{\omega^T S_B \omega}{\omega^T S_w \omega}$$

$S_B = (m_1 - m_2)(m_1 - m_2)^T$ $d \times 1$ $d \times d$

Rayleigh Quotient

$$\max_{\omega} \; \omega^T S_B \omega \quad s.t. \quad \omega^T S_w \omega = K$$

$$\max_{\omega} L(\omega) = \omega^T S_B \omega - \lambda(\omega^T S_w \omega - k)$$

$\lambda \geq 0$

$$2S_B \omega - 2\lambda S_w \omega = 0 \Rightarrow S_B \omega = \lambda S_w \omega$$

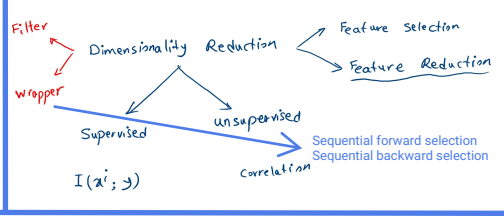
$$\underbrace{S_w^{-1} S_B}_{A} \omega = \lambda \omega$$

$$\lambda \omega = S_w^{-1} S_B \omega \Rightarrow$$

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

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

$$\Rightarrow \omega^* = S_w^{-1} (m_1 - m_2)$$



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
  - $\mu^{(0)} = \mu_1^{(0)}, \mu_2^{(0)}, \mu_3^{(0)}, \ldots, \mu_k^{(0)}$
- Interate t=0, 1, 2, . . .
  - **Classify**: Assign each point j∈{1,2,…,m} to nearest center
    $$C^{(t)}(j) \leftarrow \arg\min_{i=1,\ldots,k} \|\mu_i^{(t)} - x_j\|^2$$
  - **Recenter**: μ_i becomes centroid if its points:
    $$\mu_i^{(t+1)} \leftarrow \arg\min_{\mu} \sum_{j:C^{(t)}(j)=i} \|\mu - x_j\|^2 \qquad i \in \{1,\ldots,k\}$$
  - 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 $L(\theta)$

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

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

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

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

→ $H = \{z_1, z_2, z_3, \ldots, z_n\}$ $z_i = [\ldots 1 \ldots]$ (1 at position k, $K\times 1$)

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

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

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

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

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

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

$= \frac{p(x_i|z_{ik}=1, \theta^t) P(z_{ik}=1|\theta^t)}{p(x_i|\theta^t)}$

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

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

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

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

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

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

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

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