Mathematical Introduction for Data Analysis **Mar. 13, 2017**

Lecture 6. Supervised PCA, Dual PCA (MDS), and RKHS

Instructor: Yuan Yao, HKUST *Scribe: Zhao Yuqi*

1 Review of PCA

In the previous lectures, we discussed Principal Component Analysis(PCA) and its generalization.

1. Classical PCA in two different view:
   (a) Geometry: Looking for best affine space to approximate a high dimensional Euclidean data.
   (b) Statistics: For a signal-noise model $Y = X + \epsilon$, recovering signal direction from principal component analysis on noisy measurements Y.
2. Robust PCA: Looking for the decomposition $X = L + S$, where
   - L is a low rank matrix;
   - S is a sparse matrix.
3. Sparse PCA: Locating sparse principal component.

Those algorithms are all efficient unsupervised dimension reduction methods. However, with label attaching data, some supervised feature extraction techniques can be used to increase the computational efficiency.

2 Linear discriminant analysis

[Fisher's] *Linear discriminant analysis*(LDA), like PCA, is look for linear combinations of features which best explain the data. However, LDA explicitly attempts to model the difference between the classes of data.

For given data (X_i, y_i), where $X_i \in \mathbb{R}^p$, and y_i is discrete in $\{1, 2, ..., K\}$ not ordered.

In LDA, we want obtain feature vectors such that captures most variance between class and meanwhile discard the variance within class.

Consider between class covariance matrix

$$\hat{\Sigma}_B^{p\times p} = \frac{1}{K}\sum_{k=1}^{K}(\hat{\mu_k} - \hat{\mu})(\hat{\mu_k} - \hat{\mu})^T;$$

and within class covariance matrix

$$\hat{\Sigma}_W^{p\times p} = \frac{1}{N-K}\sum_{k=1}^{K}\sum_{y_i=k}(X_i - \hat{\mu}_k)(X_i - \hat{\mu}_k)^T,$$

where $\hat{\mu}$ is sample mean and $\hat{\mu}_k$ is within class means, i.e.

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{y_i=k} X_i.$$

Therefore

$$R(w) = \frac{w^T \hat{\Sigma}_B w}{w^T \hat{\Sigma}_W w}$$

measures, in some sense, 'signal-to-noise ratio' in terms of direction w.

Intuitively, if $\hat{\Sigma}_W$ is invertible, the eigenvector corresponding the largest eigenvalues of $\hat{\Sigma}_W^{-1}\hat{\Sigma}_B$ will maximize S. Accordingly, the best feature vectors would be eigenvectors corresponding top k eigenvalues, i.e.

$$U_k = [u_1, u_2, ..., u_k], \quad u_k \in \mathbb{R}^n,$$

where $\hat{\Sigma}_B u_k = \hat{\Sigma}_W \lambda_k u_k$ and $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_k$.

Note. For *Generalized Eigen Decomposition*(G.E.D) problem $\hat{\Sigma}_B \phi = \lambda \hat{\Sigma}_W \phi$, it is more efficient to solve Eigen Decomposition problem $\hat{\Sigma}_W^{-\frac{1}{2}} \hat{\Sigma}_B \hat{\Sigma}_W^{-\frac{1}{2}} \varphi = \lambda \varphi$ first and scale φ by $\hat{\Sigma}_W$, i.e. $\phi = \hat{\Sigma}_W^{-\frac{1}{2}} \varphi$.

Algorithm 1 Linear Discriminate Analysis

Input:

Data with label $\{X_i, y_i\}_{i=1}^N$

Output:

Effective dimension reducing directions U_k

1: Compute sample mean and within class means

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} X_i, \qquad \hat{\mu}_k = \frac{1}{N_k} \sum_{y_i=k} X_i \,;$$

2: Compute Between class covariance matrix

$$\hat{\Sigma}_B^{p \times p} = \frac{1}{K} \sum_{k=1}^{K} (\hat{\mu_k} - \hat{\mu})(\hat{\mu_k} - \hat{\mu})^T;$$

3: Compute Within class covariance matrix

$$\hat{\Sigma}_W^{p \times p} = \frac{1}{N-K} \sum_{k=1}^{K} \sum_{y_i=k} (X_i - \hat{\mu}_k)(X_i - \hat{\mu}_k)^T;$$

4: Generalized Eigen decomposition $\hat{\Sigma}_B = \hat{\Sigma}_W U \Lambda U^T$ with $\Lambda = diag(\lambda_1, \lambda_2, ...\lambda_n)$ where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_k$;

5: Choose eigenvectors corresponding to top k nonzero eigenvalues, U_k i.e.

$$U_k = [u_1, \ldots, u_k], \quad u_k \in \mathbb{R}^n;$$

6: **return** U_k .

3 Sliced Inverse Regression

[Ker-Chau Li] *Sliced inverse regression*(SIR) is a dimension reduction method using inverse regression. The idea is to find a smooth regression function that operates on a variable set of projections.

3.1 Inverse Regression versus Regression

Regression: $f(X) = \mathbb{E}[y|X]$, a curve in $\mathbb{R}^{p+1}$
Inverse Regression: $g(y) = \mathbb{E}[X|y]$, a curve (1-d manifold) in $\mathbb{R}^{p+1}$ called the principal curve or inverse regression curve.

3.2 Model

Given a response variable Y and a random vector $X \in \mathbb{R}^p$ of explanatory variables, SIR is based on the model

$$Y = g(\Gamma X, \epsilon) \ ,$$

where $\Gamma^{k\times p}$ is a unknown projection and $k < p$, and g is infinite dimensional with some linearity condition, not arbitrary.

Algorithm 2 Sliced Inverse Regression

Input:
Data with label $\{X_i, y_i\}_{i=1}^N$, where $X_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$ is continuous(or ordered discrete)
Output:
Effective dimension reducing directions Γ_k

1: Divide the range of y_i into S nonoverlapping slices $H_s(s = 1, ..., S)$. N_s is the number of observations within each slice ;

2: Compute the sample mean and total covariance matrix

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i, \qquad \hat{\Sigma}^{p\times p} = \frac{1}{K}\sum_{i=1}^{N}(X_i - \hat{\mu})(X_i - \hat{\mu})^T \ ;$$

3: Compute the mean of X_i over all slices and Between slices covariance matrix

$$\hat{\mu}_k = \frac{1}{N_s}\sum_{y_i \in H_s} X_i, \qquad \hat{\Sigma}_B^{p\times p} = \frac{1}{N}\sum_{h}^{K}(\hat{\mu_k} - \hat{\mu})(\hat{\mu_k} - \hat{\mu})^T \ ;$$

4: Generalized Eigen decomposition $\hat{\Sigma}_B = \hat{\Sigma} U \Lambda U^T$ with $\Lambda = diag(\lambda_1, \lambda_2, ...\lambda_n)$ where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_k$;

5: Choose generalized eigenvectors corresponding to top k nonzero eigenvalues, Γ_k i.e.

$$\Gamma_k = [u_1, \ldots, u_k], \quad u_k \in \mathbb{R}^n;$$

6: **return** Γ_k .

3.3 Localized SIR

[Wu-Liang-Mukherjee, 2009]*Localized Sliced Inverse Regression*(LSIR) allows for supervised dimension reduction by projection onto a linear subspace that captures the nonlinear subspace relevant to predicting the response.

Algorithm 3 Localized Sliced Inverse Regression

Input:

Data with label $\{X_i, y_i\}_{i=1}^N$, where $X_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$

Output:

Effective dimension reducing directions Γ_k

1: Compute total covariance matrix $\hat{\Sigma}$ as in SIR;

2: Divide the range of y_i into S nonoverlapping slices $H_s (s = 1, ..., S)$;

3: For each sample (X_i, y_i) compute

$$\hat{\mu}_{i,loc} = \frac{1}{|s_i|} \sum_{j \in s_i} X_j,$$

where $s_i = \{j : x_j$ belongs to the k nearest neighboors of x_i in $H_s\}$ and s indexes the slice H_s to which i belongs;

4: Compute a localized version of $\hat{\Sigma}_B$

$$\hat{\Sigma}_{loc} = \frac{1}{N} \sum_i (\hat{\mu}_{i,loc} - \hat{\mu})(\hat{\mu}_{i,loc} - \hat{\mu})^T \ ;$$

5: Solve the generalized Eigen decomposition problem $\hat{\Sigma}_{loc} = \hat{\Sigma} U \Lambda U^T$ with $\Lambda = diag(\lambda_1, \lambda_2, ...\lambda_n)$ where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_k$;

6: Choose generalized eigenvectors corresponding to top k nonzero eigenvalues, Γ_k i.e.

$$\Gamma_k = [u_1, \ldots, u_k], \quad u_k \in \mathbb{R}^n \ ;$$

7: **return** Γ_k .

4 Classical Multidimensional Scaling

Classical Multidimensional Scaling(CMDS,MDS) aims to construct Euclidean coordinates for each object in N-dimensional space such that the given between-object distances are preserved as well as possible, in particular

Given pairwise distances between n data point

$$D = [d_{ij}^2] \quad where \quad d_{ii} = 0, d_{ij} \geq 0$$

construct a system of Euclidean coordinates $\{x_i\}_{i=1}^n \in \mathbb{R}^k \quad s.t. \quad \|x_i - x_j\| = d_{ij} \quad \forall\, i, j$

4.1 Forward Problem

Given a set of points $x_1, x_2, ..., x_n \in \mathbb{R}^p$, let

$$X = [x_1, x_2, ..., x_n]^{p \times n} \ .$$

The distance between point x_i and x_j satisfies

$$d_{ij}^2 = \|x_i - x_j\|^2 = (x_i - x_j)^T(x_i - x_j) = {x_i}^T x_i + {x_j}^T x_j - 2{x_i}^T x_j \ .$$

4.2 Inverse Problem

Given d_{ij}, find a $\{x_i\}$ satisfying the relations above. Clearly the solutions are not unique as any Euclidean transform on $\{x_i\}$ gives another solution. General ideas of classic (metric) MDS is:

1. transform squared distance matrix $D = [d_{ij}^2]$ to an inner product form;
2. compute the eigen-decomposition for this inner product form.

4.3 Relationship between $\tilde{X}$ and D

Define inner product matrix

$$K_{ij} = \langle x_i, x_j \rangle,$$

and let $k = \text{diag}\{K_{ii}\} \in \mathbb{R}^n$, then

$$\begin{aligned} D_{ij} &= k_{ii} + k_{jj} - 2k_{ij} \quad or \\ D &= k \cdot \mathbf{1}^T + \mathbf{1} \cdot k^T - 2K, \end{aligned}$$

where $\mathbf{1} = (1, 1, ..., 1)^T \in \mathbb{R}^n$.

For the centered data

$$\tilde{X}^{p\times n} = XH,$$

the inner produce matrix is

$$\begin{aligned} \tilde{K}^{n\times n} &= \tilde{X}^T\tilde{X}^T \\ &= H^T(X^TX)H \\ &= HKH. \end{aligned}$$

Construct

$$\begin{aligned} B &= -\frac{1}{2}H^TDH \\ &= -\frac{1}{2}H^T(k \cdot \mathbf{1}^T + \mathbf{1} \cdot k^T - 2K)H \\ &= H^TKH \\ &= \tilde{K}, \end{aligned}$$

since

$$H^T\mathbf{1} = (I - \frac{1}{n}\mathbf{1}\mathbf{1}^T) \cdot \mathbf{1} = 0.$$

Therefore, Eign-decomposition applied to $B = -\frac{1}{2}HDH^T$ will give rise the Euclidean coordinates $\tilde{X}$.

In practice, one often chooses top k nonzero eigenvectors of B for a k-dimensional Euclidean embedding of data.

Algorithm 4 Classical MDS Algorithm

Input:

A squared distance matrix $D^{n\times n}$ with $D_{ij} = d_{ij}^2$;

Output:

Euclidean k-dimensional coordinates $\widetilde{X}_k \in \mathbb{R}^{k\times n}$ of data.

1: Compute $B = -\frac{1}{2} H \cdot D \cdot H^T$, where H is a centering matrix;

2: Eigenvalue decomposition $B = U\Lambda U^T$ with $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$;

3: Choose top k nonzero eigenvalues and corresponding eigenvectors, $\widetilde{X}_k = U_k \Lambda_k^{\frac{1}{2}}$ where

$$U_k = [u_1, \ldots, u_k], \quad u_k \in \mathbb{R}^n,$$

$$\Lambda_k = \text{diag}(\lambda_1, \ldots, \lambda_k)$$

with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k > 0$;

4: **return** $\widetilde{X}_k$;

5 Generalized MDS

Definition (Positive Semi-definite)**.** Suppose $A^{n\times n}$ is a real symmetric matrix,then:
A is *positive semi-definite*(p.s.d. *or* $A \succeq 0$) $\iff \forall v \in \mathbb{R}^n, v^T A v \geq 0 \iff A = Y^T Y$

Property. Suppose $A^{n\times n}$, $B^{n\times n}$ are real symmetric matrix, $A \succeq 0$, $B \succeq 0$. Then we have:

1. $A + B \succeq 0$;
2. $A \circ B \succeq 0$;

where $A \circ B$ is called Hadamard product and $(A \circ B)_{i,j} := A_{i,j} \times B_{i,j}$.

Definition (Conditionally Negative Definite)**.** Let $A^{n\times n}$ be a real symmetric matrix. A is *conditionally negative definite*(c.n.d.) if

$$\forall v \in \mathbb{R}^n, \quad s.t. \quad \mathbf{1}^T v = \sum_{i=1}^n v_i = 0, \text{ there holds } \quad v^T A v \leq 0$$

Lemma 5.1 (Young/Househölder-Schoenberg '1938)**.** For any signed probability measure α ($\alpha \in \mathbb{R}^n, \sum_{i=1}^n \alpha_i = 1$),

$$B_\alpha = -\frac{1}{2} H_\alpha C H_\alpha^T \succeq 0 \iff C \text{ is c.n.d.}$$

where H_α is Householder centering matrix: $H_\alpha = \mathbf{I} - \mathbf{1} \cdot \alpha^T$.

Note: $u = \frac{1}{n} \cdot \mathbf{1}$ gives the previous result: squared distance matrix is c.n.d.

Sometimes, we may want to transform a square distance matrix to another square distance matrix. The following theorem tells us the form of all the transformations between squared distance matrices.

Theorem 5.2 (Schoenberg Transform)**.** Given D a squared distance matrix, $C_{i,j} = \Phi(D_{i,j})$. Then

$$C \text{ is a squared distance matrix } \iff \Phi \text{ is a Schoenberg Transform.}$$

A *Schoenberg Transform* Φ is a transform from $\mathbb{R}^+$ to $\mathbb{R}^+$, which takes d to

$$\Phi(d) = \int_0^\infty \frac{1 - \exp(-\lambda d)}{\lambda} g(\lambda) d\lambda,$$

where $g(\lambda)$ is some nonnegative measure on $[0, \infty)$ s.t

$$\int_0^\infty \frac{g(\lambda)}{\lambda} d\lambda < \infty.$$

Examples of Schoeberg transforms include

- $\phi_0(d) = d$ with $g_0(\lambda) = \delta(\lambda)$;
- $\phi_1(d) = \dfrac{1 - \exp(-ad)}{a}$ with $g_1(\lambda) = \delta(\lambda - a)$ $(a > 0)$;
- $\phi_2(d) = \ln(1 + d/a)$ with $g_2(\lambda) = \exp(-a\lambda)$;
- $\phi_3(d) = \dfrac{d}{a(a+d)}$ with $g_3(\lambda) = \lambda \exp(-a\lambda)$;
- $\phi_4(d) = d^p$ $(p \in (0,1))$ with $g_4(\lambda) = \dfrac{p}{\Gamma(1-p)} \lambda^{-p}$.

6 Hilbert Space Embedding and Reproducing Kernels

6.1 Hilbert Space Embedding

Definition (Positive Definite Function)**.** A function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a *positive definite function*(p.d.f) if it is symmetric and positive definite, i.e.

$$\sum_{i,j} c_i c_j k(x_i, x_j) \geq 0,$$

with equality '=' holds iff $c_n = 0,\ \forall n \in \mathbb{N}$.

Equivalently, the matrix $\{k(x_i, x_j)\}^{n \times n}$ is positive definite $\forall x_i \in \mathcal{X},\ \forall n \in \mathbb{N}$

Theorem 6.1 (Schoenberg 38)**.** A separable space M with a metric function $d(x, y)$ can be isometrically embedded in a Hilbert space H, if and only if the family of functions $e^{-\lambda d^2}$ are positive definite for all $\lambda > 0$ (in fact we just need it for a sequence of λ_i whose accumulate point is 0).

6.2 Reproducing Kernels Hilbert Space

Definition. A *Reproducing Kernel Hilbert Space* (RKHS) is a Hilbert space associated with a p.d.f $k(\cdot, \cdot)$ such that:

1. $k_x \in H$, where $k_x = k(x, \cdot)$;
2. linear span: $L := \sum_x c_x k_x$;

3. inner product: $\langle k_x, k_y\rangle = k(x,y)$ extends to L: $\langle \sum_x c_x k_x, \sum_y c_y k_y\rangle = \sum_{x,y} c_x c_y k(x,y)$;

4. closure under induced norm $\|k_x\|_H = \sqrt{k(x,x)}$.

Theorem 6.2 (Moore Aronszajn)**.** Suppose k is a symmetric, positive definite kernel on a set $\mathcal{X}$. Then there is a unique Hilbert space of functions on $\mathcal{X}$ for which k is a reproducing kernel, i.e.,

$$f(x) = \langle f, k(\cdot, x)\rangle_H.$$

The Hilbert space is often denoted by H_k.

From here one can see

$$\left\|\sum_i c_i k_{x_i}\right\|_{\mathcal{H}}^2 = \sum_{i,j} c_i c_j k(x_i, x_j) = c^T K c, \quad c = (c_i) \in R^n,\ K = (k(x_i, x_j)) \in R^{n\times n}$$

Example. Some kernels in RKHS:

1. linear: $k(x,y) = \langle x, y\rangle$, positive semi-definite;
2. polynomial: $k(x,y) = (\langle x, y\rangle + a_0)^d$, positive semi-definite;
3. Gaussian kernel (radial basis function): $k(x,y) = \exp(-\gamma\|x-y\|^2)$ $(\gamma > 0)$, positive definite;
4. sigmoid: $k(x,y) = \tanh(\langle x, y\rangle + a_0)$, *NOT positive semi-definite*;
5. piecewise polynomial, splines, Sobolev spaces (Wahba'1990);
6. wavelet (Daubechies, Ten lectures on wavelets)
7. Mercer kernel: $\mathcal{X}$ is compact, $k(x,y) \in C(\mathcal{X}, \mathcal{X})$, e.g. linear, polynomial and Gaussian kernels are Mercer's kernel.

By Riesz representation, for every $x \in \mathcal{X}$ there exists $E_x \in \mathcal{H}$ such that $f(x) = \langle f, E_x\rangle$.

The evaluation functional over the Hilbert space H is a linear functional that evaluates each function at a point x,

$$L_x : f \mapsto f(x) \quad \forall f \in H.$$

If $0 < k(x,x) \le \kappa$ for all $x \in \mathcal{X}$, then the evaluation functional L_x is bounded, $|L_x(f)| = |f(x)| = |\langle f, E_x\rangle| \le \|f\|_H \|E_x\|_H \le \sqrt{\kappa}\|f\|_H$ where $\|E_x\|_H \le \sqrt{\langle E_x, E_x\rangle} \le \sup_{x\in\mathcal{X}} \sqrt{k(x,x)} \le \sqrt{\kappa}$. Bounded evaluation functional imposed certain stability: *a small perturbation on model f will lead to a small error in its predicted value $f(x)$*. This is a desired property in statistical modeling, which leads to the following fact.

[Wahba 1990] All function models in Hilbert space in statistics are RKHS.

Theorem 6.3 (Mercer)**.** A continuous p.d.f k defined on compact domain $\mathcal{X} \times \mathcal{X}$ has the expansion:

$$k(x,y) = \sum_{i=1}^{\infty} \gamma_i \phi_i(x)\phi_i(y),$$

where

$$\int_{\mathcal{X}} k(x,y)\phi_i(y)\, d\mu(y) = \gamma_i \phi_i(x),$$

and

- eigen values $\gamma_i \to 0$;
- eigen functions $\{\phi_i\}$ form an orthonormal basis of $L_2(\mathcal{X})$.

The convergence is absolute and uniform over $\mathcal{X} \times \mathcal{X}$.

Any functions $f, g \in H_K$ admit a series expansion

$$f = \sum_i a_i \phi_i, \quad g = \sum_i b_i \phi_i$$

whose H_K-inner product is equivalent to

$$\langle f, g \rangle_H = \sum_i \frac{a_i b_i}{\gamma_i}.$$

6.3 Regularization and Reproducing Kernel Hilbert Spaces

A general class of regularization problems has the form

$$\min_{f \in \mathcal{H}} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{H}}^2,$$

which is a infinite-dimensional problem.

It can be shown that the solution of this problem admits a finite-dimensional representation in the form

$$f(x) = \sum_{i=1}^{n} \alpha_i k_{x_i}(x). \tag{1}$$

In fact, assume that in general the solution has the following form

$$f = \sum_i \alpha_i k_{x_i} + g$$

where $g \perp \text{span}\{k_{x_i}\}$. Hence $\langle g, k_{x_i} \rangle = g(x_i) = 0$ for all i, which implies that

$$\ell(y_i, f(x_i)) = \ell(y_i, \left\langle \sum_i \alpha_i k_{x_i} + g, k_{x_i} \right\rangle) = \ell(y_i, \left\langle \sum_i \alpha_i k_{x_i}, k_{x_i} \right\rangle).$$

Other the other hand, by Pythagorean Theorem

$$\|f\|_{\mathcal{H}}^2 = \|\sum_i \alpha_i k_{x_i}\|_{\mathcal{H}}^2 + \|g\|_{\mathcal{H}}^2$$

so a minimization of such a penalty pushes $g = 0$ if the kernel k is positive definite (Note that a positive semi-definite kernel: $\|g\| = 0 \not\Rightarrow g = 0$). Therefore a minimizer must have the form Eq. (1). This is often called as *Representer's Theorem* in machine learning due to its origin from Riesz Representation Theorem.

Example. Some loss functions in Support Vector Machines
Hinge loss in SVM Classification: $\ell(y, f(x)) = \max(0, 1 - yf(x))$
ϵ-insensitive loss in SVM Regression: $\ell(y, f(x)) = \max(0, |y - f(x)| - \epsilon)$

6.4 Support Vector Machines: Classification

[Vapnik 1992] Consider a linear classifier in $\mathbb{R}^d$,

$$f(x) = a^T x + a_0, \quad a \in \mathbb{R}^d,\ a_0 \in \mathbb{R},$$

SVM Classification (SVMC) aims to solve the following problem

$$\min_{(a,a_0)\in\mathbb{R}^{d+1}} \frac{1}{n}\sum_{i=1}^{n} \ell_{hinge}(y_i, f(x_i)) + \lambda\|a\|^2$$

where $\ell_{hinge}(y, f(x)) = \max(0, 1 - yf(x))$ is the hinge loss function. This is equivalent to

$$\begin{aligned} (Primal) \quad \min_{(a,a_0,\zeta)} \quad & \frac{1}{n}\sum_i \zeta_i + \lambda\|a\|^2 \\ s.t. \quad & 1 - y_i(a^T x_i + a_0) \le \zeta_i \\ & \zeta_i \ge 0,\ i = 1, \dots, n \end{aligned}$$

Define the Lagrangian

$$\max_{\alpha\ge 0,\beta\ge 0} \min_{(a,a_0,\zeta)} L(\zeta, a, a_0; \alpha, \beta) := \frac{1}{n}\sum_i \zeta_i + \lambda\|a\|^2 - \sum_i \alpha_i(\zeta_i - 1 + y_i(a^T x_i + a_0)) - \sum_i \beta_i\zeta_i$$

The saddle point condition

$$0 = \frac{\partial L}{\partial \zeta} = \frac{1}{n} - \alpha - \beta \Rightarrow 0 \le \alpha \le \frac{1}{n} \text{ for } \alpha \ge 0, \beta \ge 0$$

$$0 = \frac{\partial L}{\partial a_0} = \sum_i \alpha_i y_i$$

$$0 = \frac{\partial L}{\partial a} = 2\lambda a - \sum_i \alpha_i y_i x_i \Rightarrow a = \frac{1}{2\lambda}\sum_i \alpha_i y_i x_i$$

Now replacing $\hat{\alpha} = \frac{1}{2\lambda}\alpha$, we see

$$f(x) = a^T x + a_0 = a_0 + \sum_{i,j} \hat{\alpha}_i y_i \langle x_i, x\rangle = a_0 + \sum_{i,j} \hat{\alpha}_i y_i k(x_i, x)$$

$$\|a\|^2 = \sum_{i,j} \hat{\alpha}_i y_i \hat{\alpha}_j y_j \langle x_i, x_j\rangle = \sum_{i,j} \hat{\alpha}_i y_i \hat{\alpha}_j y_j k(x_i, x_j)$$

Hence we reach the dual SVMC problem

$$\begin{aligned} (Dual) \quad \max_{\hat{\alpha}} \quad & \sum_i \hat{\alpha}_i - \frac{1}{2}\sum_{i,j} \hat{\alpha}_i y_i \hat{\alpha}_j y_j k(x_i, x_j) \\ s.t. \quad & \sum_i \hat{\alpha}_i y_i = 0 \\ & 0 \le \hat{\alpha}_i \le \frac{1}{2n\lambda} =: C,\ i = 1, \dots, n \end{aligned}$$

So parameter C controls the regularity: the smaller is C, the smoother is the function (less variation).

From the complementary condition

$$\alpha_i(\zeta_i - 1 + y_i(a^T x_i + a_0)) = 0$$

and

$$\beta_i \zeta_i = 0,$$

we can see

1. for i's correctly identified with margin at least 1, $y_i f(x_i) > 1 \Rightarrow \alpha_i = 0$;
2. for i's within the margin $y_i f(x_i) < 1 \Rightarrow \zeta_i > 0 \Rightarrow \beta_i = 0 \Rightarrow \alpha_i = 1/n > 0$ or $\hat{\alpha}_i = C > 0$, maximized;
3. for i's on the margin boundary $y_i f(x_i) = 1 \Rightarrow \zeta_i = 0 \Rightarrow \beta_i > 0 \Rightarrow 0 < \alpha_i < 1/n$ or $0 < \hat{\alpha}_i < C$, one can decide $a_0 = y_i - a^T x_i$.

The samples corresponding to $\hat{\alpha}_i \neq 0$, lying in the latter two cases above, are called *support vectors* since they have classification margin less than 1 and influence the location of decision hyperplane.

The procedure above can be generalized to infinite dimensional space with $f = a_0 + \sum_k a_k \phi_k$ and $\|f\|^2 = \sum_k a_k^2/\gamma_k$ for $k(x, x') = \sum_k \gamma_k \phi_k(x)\phi_k(x')$ where $k \in \mathbb{N}$. The primal problem is of infinite dimension, however due to the representer theorem, the dual problem is still the same.