Pattern Classification

Wing-Kin (Ken) Ma

Department of Electronic Engineering,
The Chinese University Hong Kong, Hong Kong

ELEG5481, Lecture 14

Acknowledgment: Jiaxian Pan for helping prepare the slides.

Introduction

- Suppose that we are given 50 pictures[1] of tigers, 50 pictures of dolphins, and 50 picture of monkeys.

- From the given pictures, we learn how tigers, dolphins and monkeys look like.

- Now, given a new picture, we want to know whether it is tiger, dolphin, or a monkey.

A tiger,
a dolphin,
or a monkey?

[1] All photographs are taken from the Internet.

Pattern classification problem setup

- Let $\mathcal{X}$ be set of all possible inputs and $\mathcal{Y}$ be the set of all classes.
- We are given a set of training examples (or training data) $x_1, x_2, \ldots, x_m \in \mathcal{X}$.
- Each data point x_i has been labeled to belong a certain class. Let $y_1, y_2, \ldots, y_m \in \mathcal{Y}$ be the class labels corresponding to $x_1, x_2, \ldots, x_m$.
  - For example, consider $\mathcal{Y} = \{-1, +1\}$. The data point x_i belongs to class "$+1$" if $y_i = 1$, and class "-1" if $y_i = -1$.
- Let $f : \mathcal{X} \to \mathcal{Y}$ be a classifier or decision function, which should do the following:
  - $f(x_i) = y_i$ for $i = 1, \ldots, m$, or, if not possible, maximizes the number of training examples satisfying $f(x_i) = y_i$.
  - for a new data $x \in \mathcal{X}$, predict its class by $f(x)$.
- Goal: learn a good classifier from the training data $\{(x_i, y_i)\}_{i=1}^m$.

Binary Classification by the Support Vector Machine (SVM)

- Consider the binary classification case. Let $\mathcal{Y} = \{-1, +1\}$.

- Consider a simple decision function

$$f(x) = \mathrm{sign}(w^T x + b)$$

  where $w \in \mathbf{R}^n$ and $b \in \mathbf{R}$. This classifier is known as the SVM.

- **Problem 1:** given $\{(x_i, y_i)\}_{i=1}^m$, find (w, b) such that

$$y_i = \mathrm{sign}(w^T x_i + b), \ \ i = 1, \ldots, m. \tag{$*$}$$

- Eq. $(*)$ is equivalent to

$$w^T x_i + b > 0, \ \text{if } y_i = 1, \qquad w^T x_i + b < 0, \ \text{if } y_i = -1,$$

  for $i = 1, \ldots, m$. Or, we can write

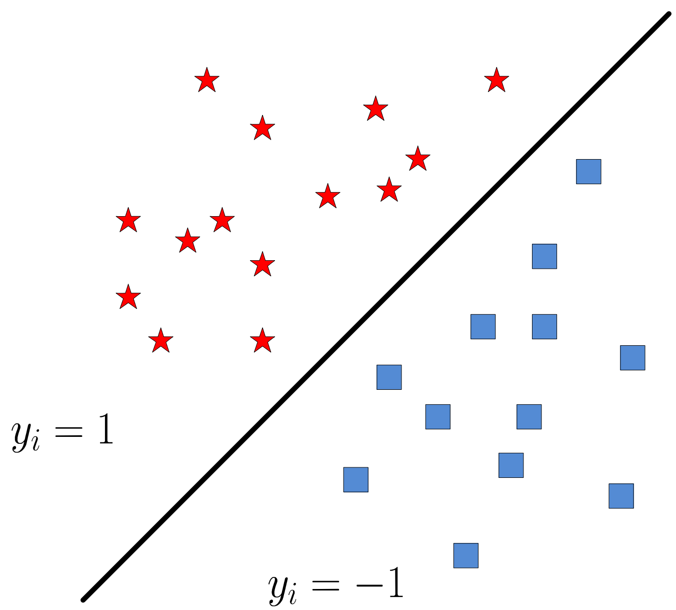
$$y_i(w^T x_i + b) > 0, \quad i = 1, \ldots, m.$$

- Problem 1 can be written as

$$\begin{aligned} &\text{find } w, b \\ &\text{s.t. } y_i(w^T x_i + b) > 0, \quad i = 1, \dots, m, \end{aligned}$$

  which is an LP feasibility problem.

- Geometrically, the problem is to find a hyperplane $\mathcal{H} = \{x \mid w^T x + b = 0\}$ that separates the data $\{x_i \mid y_i = 1\}$ from $\{x_i \mid y_i = -1\}$.



A Robust SVM Formulation

- Suppose that there are uncertainties in $\{x_i\}_{i=1}^m$, say, due to noise and modeling errors.

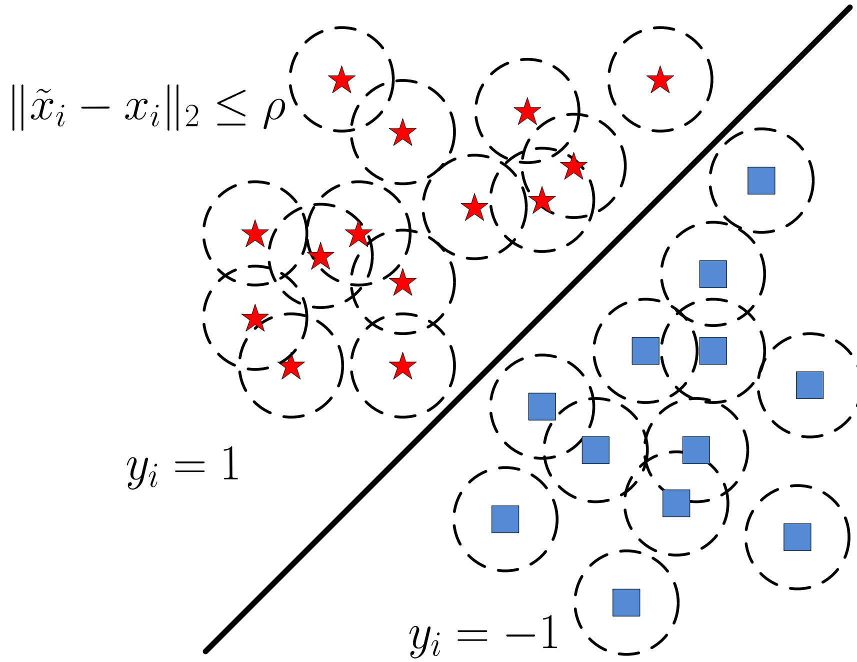
- Under such cases, the classifier design in Problem 1 is not robust.

- Consider the spherical uncertainty model:

$$\tilde{x}_i = x_i + e_i, \quad \|e_i\|_2 \leq \rho,$$

  for $i = 1, \ldots, m$, where x_i now denotes the "nominal" data point; $\tilde{x}_i$ the "true" data point; e_i the corresponding uncertainty vector; ρ the uncertainty level.

- We wish to maximize the uncertainty level while still separating the data.



- **Problem 2:**

$$\max_{w,b,\rho} \; \rho$$

$$\text{s.t. } y_i(w^T(x_i + e_i) + b) \geq 0, \quad \text{for all } \|e_i\|_2 \leq \rho, \; i = 1, \ldots, m.$$

- A recap of problem 2:

$$\max_{w,b,\rho} \rho$$
$$\text{s.t. } y_i(w^T(x_i + e_i) + b) \geq 0, \quad \text{for all } \|e_i\|_2 \leq \rho, \; i = 1, \ldots, m.$$

- By the Cauchy-Schwarz inequality, we have

$$\inf_{\|e_i\|_2 \leq \rho} y_i(w^T(x_i + e_i) + b) \geq 0 \Longleftrightarrow y_i(w^T x_i + b) - \rho\|w\|_2 \geq 0.$$

- Problem 2 is homogeneous—if $(w^\star, b^\star)$ is a solution, then $(\alpha \cdot w^\star, \alpha \cdot b^\star)$, for any $\alpha > 0$, is also a solution.

- Assume w.l.o.g. that $\rho\|w\|_2 = 1$. Problem 2 can be reformulated as

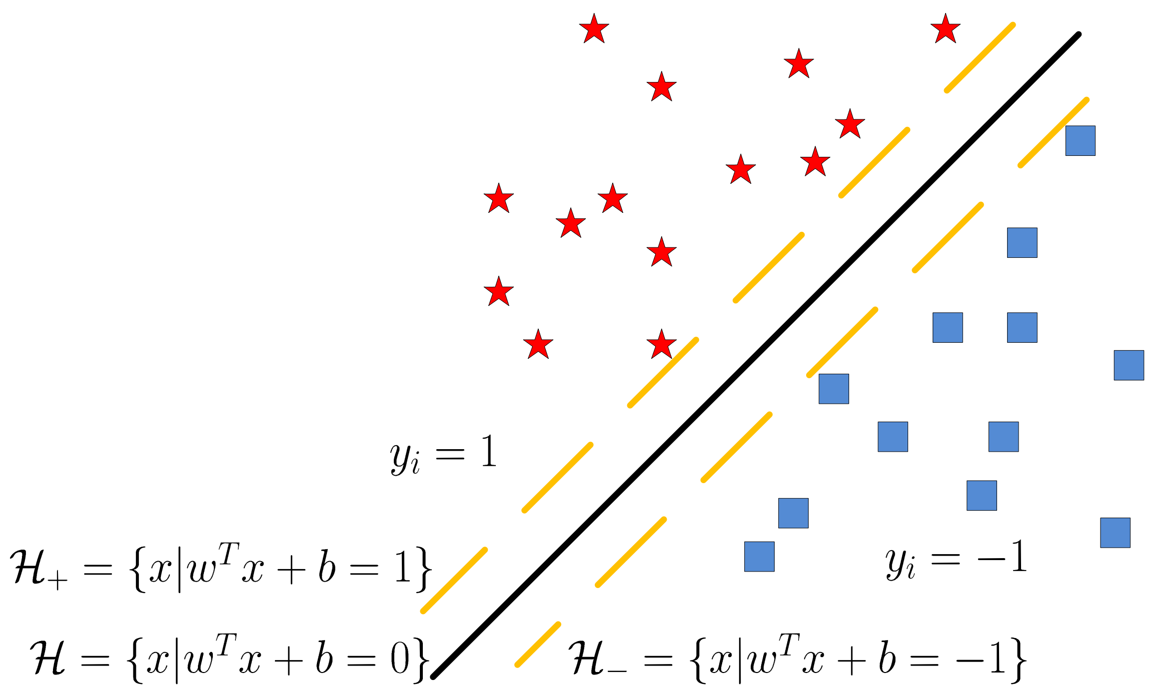
$$\min_{w,b} \|w\|_2^2$$
$$\text{s.t. } y_i(w^T x_i + b) \geq 1, \; i = 1, \ldots, m.$$

Alternative (and classical) Interpretation

- Define hyperplanes $\mathcal{H}_+ = \{x | w^T x + b = 1\}$ and $\mathcal{H}_- = \{x | w^T x + b = -1\}$.

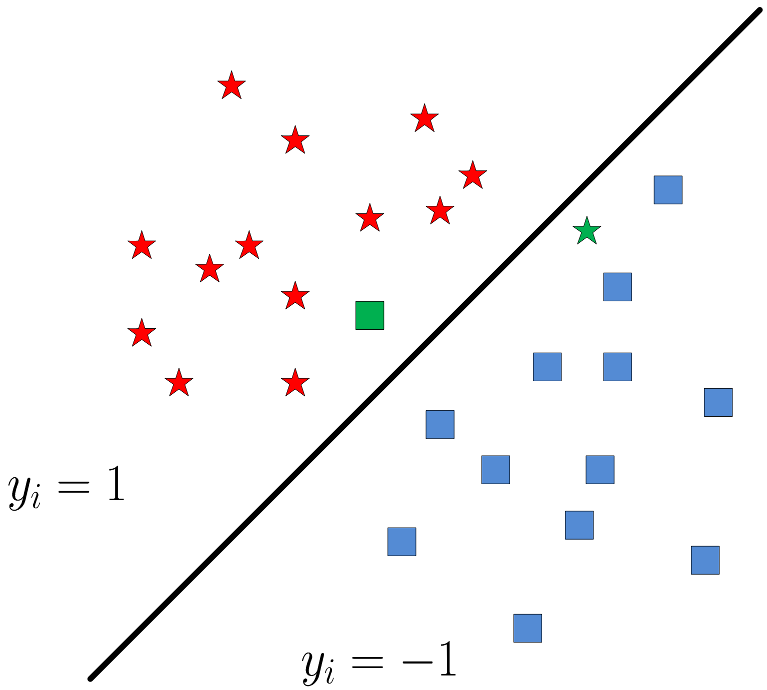
- The distance between $\mathcal{H}_+$ and $\mathcal{H}_-$ is $2/\|w\|_2$.

- Problem 2 is identical to that of maximizing the distance between the parallel hyperplanes $\mathcal{H}_+$ and $\mathcal{H}_-$.



The Non-Separable Data Case

- A given training data set $\{(x_i, y_i)\}_{i=1}^{m}$ is not always separable; i.e., there does not exist a hyperplane that separates $\{x_i \mid y_i = -1\}$ and $\{x_i \mid y_i = 1\}$.



- As a compromise, a minimum "loss" should be sought.

A Soft Margin SVM Formulation

- Let $\psi : \mathbf{R} \to \{0, 1\}$ be a step loss function:

$$\psi(x) = \begin{cases} 0, & x \leq 0 \\ 1, & x > 0. \end{cases}$$

- **Problem 3 (an ℓ_0-norm-like soft margin SVM)**:

$$\min_{w,b} \|w\|_2^2 + \lambda \cdot \sum_{i=1}^{m} \psi(1 - y_i(w^T x_i + b))$$

  for some constant $\lambda > 0$.

  – we design an SVM whose number of class-violated data points is small.

  – the problem is also robust against mislabeled data points.

  – the problem has a sparse opt. flavor.

- Problem 3 is nonconvex, owing to ψ (the same problem as in ℓ_0 norm).

- Like sparse opt., a compromise is to approximate ψ by a more manageable function. As an example, consider the hinge loss function:

$$h(x) = \begin{cases} 0, & x \leq 0 \\ x, & x > 0. \end{cases}$$

  h is convex. Also, note that $h(x) = \max\{0, x\}$.

- ℓ_1**-norm-like soft margin SVM:**

$$\min_{w,b} \|w\|_2^2 + \lambda \cdot \sum_{i=1}^{m} \max\{0, 1 - y_i(w^T x_i + b)\}.$$

  The problem above can be reformulated as an SOCP (or convex QP):

$$\begin{aligned} \min_{w,b,\xi} \ & \|w\|_2^2 + \lambda \cdot \sum_{i=1}^{m} \max\{0, \xi_i\} \\ \text{s.t. } & \xi_i \geq 0, \ \xi_i \geq 1 - y_i(w^T x_i + b), \quad i = 1, \ldots, m. \end{aligned}$$

  Note: the above problem is the classical SVM formulation.

Variations of SVM Formulations

- One may consider other approximate functions for ψ (e.g., the logistic regression loss $\log(1 + e^{-y_i(w^T x_i + b)})$).

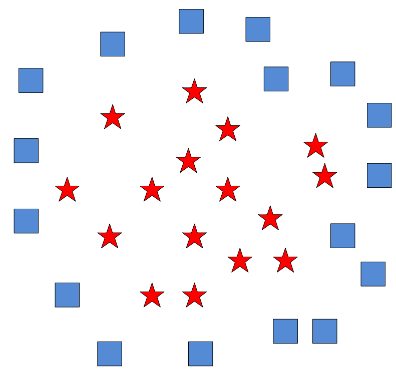
- One may also modify the uncertainty model.
  - For example, consider an interval uncertainty $\|e_i\|_\infty \leq \rho$.
  - The resulting SVM problem (with ℓ_1-norm-like soft margin):

$$\min_{w,b} \|w\|_1 + \lambda \cdot \sum_{i=1}^{m} \max\{0, 1 - y_i(w^T x_i + b)\}.$$

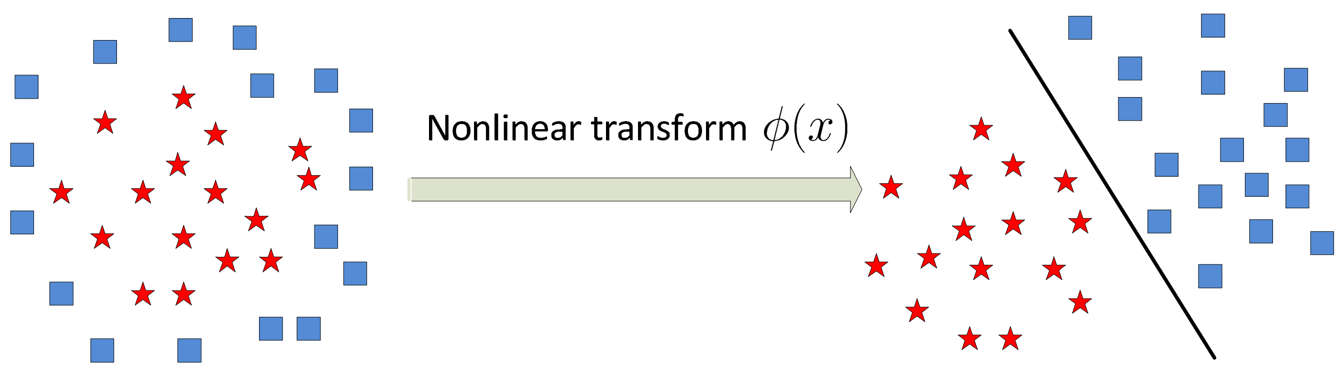
  - Alternative interpretation: Since $\|w\|_1$ approximates $\|w\|_0$, the above SVM problem has a flavor of choosing the smallest of elements (or features) to perform classification.

Nonlinear SVM

- SVM restricts itself to the use of linear decision regions.
  - pros: "easy" to optimize.
  - cons: there are many cases where linear decision regions are not adequate.



- A possible remedy is to introduce a nonlinear mapping $\phi(x)$ to map data into a different space, and then construct a linear classifier in that space.



- Nonlinear SVM problem

$$\min_{w,b,\xi} \ \|w\|_2^2 + \lambda \cdot \sum_{i=1}^{m} \xi_i$$
$$\text{s.t. } y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, m.$$

where $\phi : \mathbf{R}^n \to \mathbf{R}^l$ is a predefined nonlinear mapping.

- This problem is still an SOCP, though a nonlinear mapping is applied to data x_i.

- In practice, the dimension l of $\phi(x)$ can be very large or even infinite. This can cause significant problems in storing data in memory and solving the SOCP.

The Representer Theorem

- The representer theorem [Shawe-Taylor and N. Cristianini'04] states that there is an optimal solution w of the nonlinear SVM problem such that

$$w = \sum_{i=1}^{m} \alpha_i \phi(x_i)$$

  for some $\alpha \in \mathbf{R}^m$.

- The representer thm. suggests that $w \in \mathbf{R}^l$ lies in some low dimensional space spanned by $\{\phi(x_i)\}_{i=1}^m$, though the dimension l could be huge or even infinite.

- The nonlinear SVM problem is transformed to

$$\begin{aligned} \min_{w,b,\alpha,\xi} \ & \|w\|_2^2 + \lambda \cdot \textstyle\sum_{i=1}^m \xi_i \\ \text{s.t. } & y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, m, \\ & w = \textstyle\sum_{i=1}^m \alpha_i \phi(x_i). \end{aligned}$$

The Kernel Trick

- By direct substitution w, the nonlinear SVM problem can further be rewritten as

$$\min_{b,\alpha,\xi} \alpha^T Q\alpha + \lambda \cdot \textstyle\sum_{i=1}^m \xi_i$$
$$\text{s.t. } y_i(\textstyle\sum_{j=1}^m \alpha_j Q_{ij} + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, m,$$

  where $Q \in \mathbf{S}_+^m$ with $Q_{ij} = \phi(x_i)^T \phi(x_j)$.

- To obtain Q, we do not need $\phi(x_i)$ explicitly. We only need the inner products $\phi(x_i)^T \phi(x_j)$.

- There is no need to explicitly define the transform $\phi(x)$. Instead, we specify the so-called kernel function $K(x, x') = \phi(x)^T \phi(x')$.

- Popular choice of kernel:

$$K(x, x') = \exp(-\delta\|x - x'\|^2) \qquad \text{(Radial basis function)}$$
$$K(x, x') = (x^T x'/a + b)^d \qquad \text{(Polynomial kernel)}$$

The Decision Function under the Kernel Trick

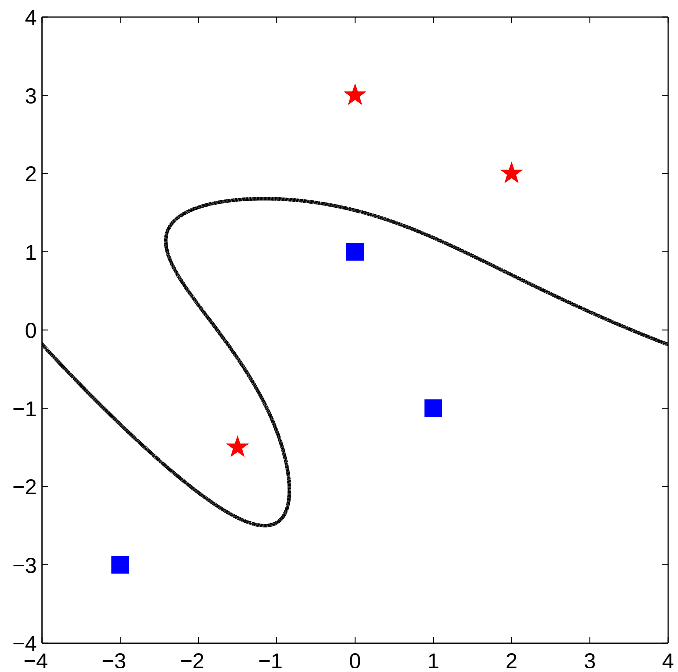
- The decision function is written as

$$
\begin{aligned}
f(x) &= \operatorname{sign}\left((w^\star)^T x + b^\star\right) \\
&= \operatorname{sign}\left(\textstyle\sum_{i=1}^{m} \alpha_i^\star K(x, x_i) + b^\star\right).
\end{aligned}
$$

- The decision function is again specified by the kernel function $K(x, x')$ only.

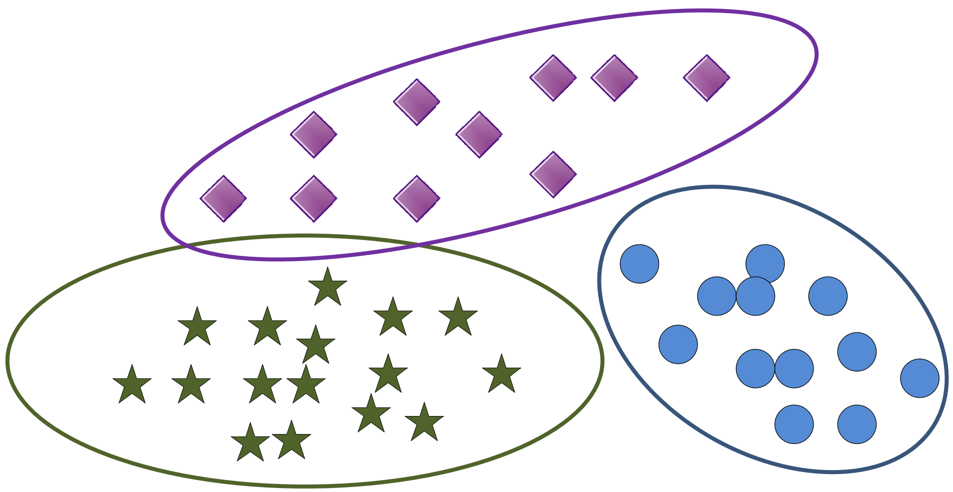
A toy example

- Six data points: $(-3, -3)$, $(0, 1)$, $(1, -1)$ are of class -1, and $(-1.5, -1.5)$, $(2, 2)$, $(0, 3)$ are of class 1.
- Nonlinear SVM with regularization $\lambda = 0.5$.
- Radical basis function with $\delta = 0.2$.
- The black line is the decision boundary.



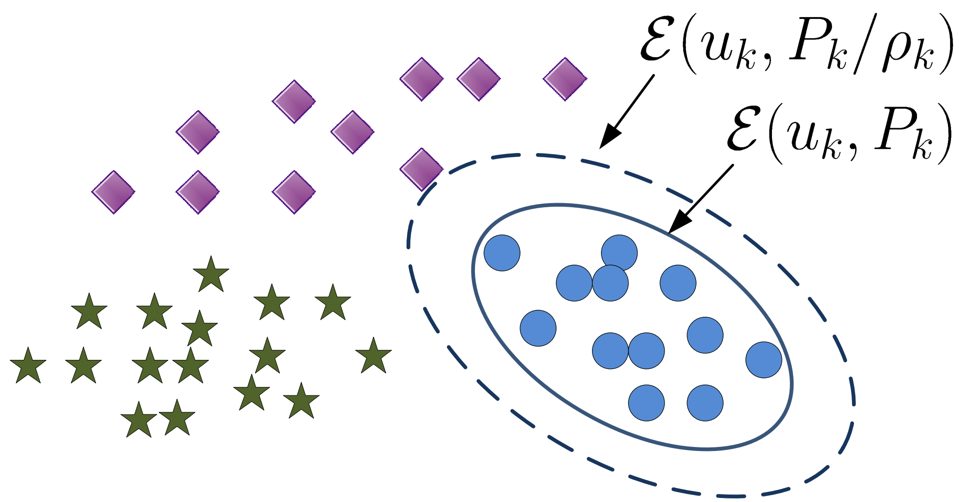
Maximum-Ratio Separating Ellipsoids (MRSEs)

- SVM employs linear decision regions.
- One can also consider ellipsoidal decision regions.



- Consider a K-class classification problem with $\mathcal{Y} = \{1, \ldots, K\}$.
- Define the ellipsoidal set as $\mathcal{E}(u, P) = \{x \mid (x-u)^T P (x-u) \leq 1\}$, where u is the center and $P \succeq 0$.
- For each class $k \in \mathcal{Y}$, the objective is to find an ellipsoid $\mathcal{E}(u_k, P_k)$ and a scaled ellipsoid $\mathcal{E}(u_k, P_k/\rho_k)$ with $\rho_k \geq 1$ such that

$$\begin{cases} x_i \in \mathcal{E}(u_k, P_k), & \text{if } y_i = k, \\ x_i \notin \mathcal{E}(u_k, P_k/\rho_k), & \text{if } y_i \neq k. \end{cases}$$



- The scaling factor ρ_k should be maximized, as ρ_k can be considered as the margin between class k and all other classes.

- The MRSE optimization problem:

$$\begin{aligned}
\max_{P_k, u_k, \rho_k} \quad & \rho_k + \lambda_1 \log\det P_k \\
\text{s.t.} \quad & (x_i - u_k)^T P_k (x_i - u_k) \leq 1, \quad \text{if } y_i = k, \\
& (x_i - u_k)^T P_k (x_i - u_k) \geq \rho_k, \quad \text{if } y_i \neq k, \\
& \rho_k \geq 1, \\
& P_k \succeq 0,
\end{aligned}$$

  for $k = 1, \ldots, K$, where a regularization $\lambda_1 \log\det P_k$ with $\lambda_1 > 0$ is added to the objective function to ensure that the ellipsoid is non-degenerate, i.e., $P_k \succ 0$.

- If the optimal ρ_k satisfies $\rho_k \geq 1$, then the training data of class k can be perfectly separated from those of other classes.

- The MRSE problem is not convex, but can be transformed to a convex problem by a technique called homogeneous embedding.

$$\begin{aligned} \max_{\Phi_k, \rho_k} \ & \rho_k + \lambda_1 \log\det \Phi_{11} \\ \text{s.t.} \ & z_i^T \Phi_k z_i \le 1, \quad \text{if } y_i = k, \\ & z_i^T \Phi_k z_i \ge \rho_k, \quad \text{if } y_i \ne k, \\ & \Phi_k = \begin{bmatrix} \Phi_{11} & \phi_{12} \\ \phi_{12}^T & \phi_{22} \end{bmatrix} \succeq 0, \\ & \Phi_k \succeq 0, \end{aligned}$$

where $z_i = [x_i^T, 1]^T$.

- $P_k^\star$ and $u_k^\star$ can be recovered by

$$P_k^\star = \Phi_{11}^\star/(1-\delta^\star), \quad u_k^\star = -(\Phi_{11}^\star)^{-1}\phi_{12}^\star, \quad \delta^\star = \phi_{22}^\star - (\phi_{12}^\star)^T(\Phi_{11}^\star)^{-1}\phi_{12}^\star.$$

- For the case of non-separate data, the same soft margin formulation in SVM can be used:

$$
\begin{aligned}
\max_{\Phi_k,\rho_k,\xi}\ & \rho_k + \lambda_1 \log\det\Phi_{11} - \lambda_2 \textstyle\sum_i \xi_i \\
\text{s.t. } & z_i^T \Phi_k z_i \le 1 + \xi_i, \quad \text{if } y_i = k, \\
& z_i^T \Phi_k z_i \ge \rho_k - \xi_i, \quad \text{if } y_i \ne k, \\
& \Phi_k = \begin{bmatrix} \Phi_{11} & \phi_{12} \\ \phi_{12}^T & \phi_{22} \end{bmatrix}, \\
& \Phi_k \succeq 0, \\
& \gamma_i \ge 0, \ \text{for all } i,
\end{aligned}
$$

where $\lambda_2 > 0$ is some positive regularization parameter.

Classification Rule

- Suppose in the training phase, we have solved the MRSE problem for each $k \in \{1, \ldots, K\}$.

- Given a new data x, define the score of class k as

$$s_k = \frac{(x - u_k^\star)^T P_k^\star (x - u_k^\star)}{\sqrt{\rho_k^\star}}.$$

- The score s_k measures how closed x is to class k.

- Choose the class that has the minimum score:

$$\hat{k} = \arg \min_{k=1,\ldots,K} \quad s_k.$$

Reference

A. Ben-Tal, L. El-Ghaoui, and A. Nemirovski, "Robust Optimization", Princeton University Press, 2009.

L. Xiao and L. Deng, "A geometric perspective of large-margin training of Gaussian Models", *IEEE Trans. Signal Process. Mag.*, 2010.

J. Shawe-Taylor and N. Cristianini, "Kernel Methods for Pattern Analysis", Cambridge University Press, Cambridge, 2004.

K. Q. Weinberger, F. Sha, and L. K. Saul, "Convex optimization for distance metric learning and pattern classification," *IEEE Signal Process. Mag.*, May. 2010.