Data Mining Techniques

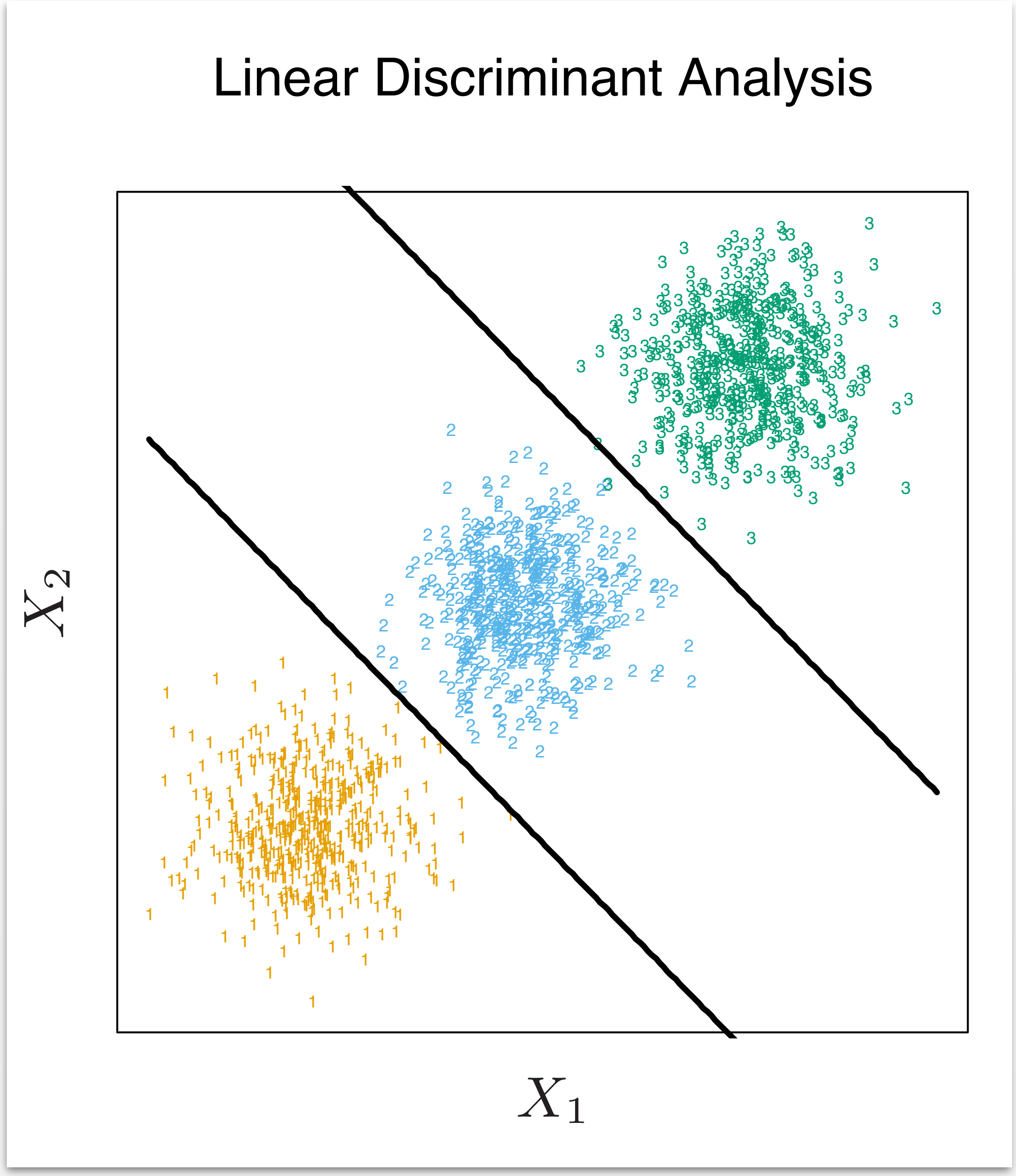
CS 6220 - Section 3 - Fall 2016

Lecture 5: Classification 2

Jan-Willem van de Meent
(*credit*: Zhao, CS 229, Bishop)

Generative Learning Algorithms

Linear Discriminant Analysis



Algorithm

- Mean for each class

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n:y_n=k} \boldsymbol{x}_n$$

- Covariance for each class

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n:y_n=k} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top (\boldsymbol{x}_n - \boldsymbol{\mu}_k)$$

- Average covariance

$$\boldsymbol{\Sigma} = \frac{1}{N} \sum_k N_k \Sigma_k$$

Linear Discriminant Analysis

Predict using likelihood

$$y^* = \underset{k}{\operatorname{argmax}}\, p(\boldsymbol{x}^* \,|\, y = k)$$

Linear Discriminant Analysis

$$p(\boldsymbol{x} \,|\, y = k) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma})$$

Quadratic Discriminant Analysis

$$p(\boldsymbol{x} \,|\, y = k) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Algorithm

- Mean for each class

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n: y_n = k} \boldsymbol{x}_n$$

- Covariance for each class

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n: y_n = k} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top (\boldsymbol{x}_n - \boldsymbol{\mu}_k)$$

- Average covariance

$$\boldsymbol{\Sigma} = \frac{1}{N} \sum_k N_k \Sigma_k$$

Linear Discriminant Analysis

Predict using likelihood

$$y^* = \underset{k}{\operatorname{argmax}}\, p(\boldsymbol{x}^* \mid y = k)$$

Linear Discriminant Analysis

Predict using likelihood

$$y^* = \underset{k}{\operatorname{argmax}}\, p(\boldsymbol{x}^* \mid y = k)$$

Predict using posterior

$$y^* = \underset{k}{\operatorname{argmax}}\, p(y = k \mid \boldsymbol{x}^*)$$

Linear Discriminant Analysis

Predict using likelihood

$$y^* = \underset{k}{\operatorname{argmax}}\, p(\boldsymbol{x}^* \mid y = k)$$

Predict using posterior

$$y^* = \underset{k}{\operatorname{argmax}}\, p(y = k \mid \boldsymbol{x}^*)$$

Generative Model

$$\begin{aligned} y_n &\sim \text{Discrete}(\boldsymbol{\pi}) \\ x_n | y_n = k &\sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \end{aligned}$$

Bayes Rule

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})}$$

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \boldsymbol{y})p(\boldsymbol{y})}{p(\boldsymbol{x})}$$

Linear Discriminant Analysis

Predict using likelihood

$$y^* = \underset{k}{\operatorname{argmax}}\, p(\boldsymbol{x}^* \,|\, y = k)$$

Predict using posterior

$$y^* = \underset{k}{\operatorname{argmax}}\, p(y = k \,|\, \boldsymbol{x}^*)$$

Generative Model

$$y_n \sim \text{Discrete}(\boldsymbol{\pi})$$
$$x_n | y_n = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Generative Learning

- Treat features as *"observations"*
- Treat class labels as *"latent variables"*
- Calculate ML estimates of parameters
- Predict according to MAP value

Linear Discriminant Analysis

Predict using likelihood

$$y^* = \underset{k}{\operatorname{argmax}}\, p(\boldsymbol{x}^* \,|\, y = k)$$

Predict using posterior

$$y^* = \underset{k}{\operatorname{argmax}}\, p(y = k \,|\, \boldsymbol{x}^*)$$

Generative Model

$$\begin{aligned} y_n &\sim \mathrm{Discrete}(\boldsymbol{\pi}) \\ x_n | y_n = k &\sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \end{aligned}$$

Maximum Likelihood Estimates

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n: y_n = k} \boldsymbol{x}_n$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n: y_n = k} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top (\boldsymbol{x}_n - \boldsymbol{\mu}_k)$$

$$\pi_k = \textbf{?}$$

Linear Discriminant Analysis

Predict using likelihood

$$y^* = \underset{k}{\operatorname{argmax}}\, p(\boldsymbol{x}^* \,|\, y = k)$$

Predict using posterior

$$y^* = \underset{k}{\operatorname{argmax}}\, p(y = k \,|\, \boldsymbol{x}^*)$$

Generative Model

$$\begin{aligned} y_n &\sim \text{Discrete}(\boldsymbol{\pi}) \\ x_n | y_n = k &\sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \end{aligned}$$

Maximum Likelihood Estimates

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n:y_n=k} \boldsymbol{x}_n$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n:y_n=k} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top (\boldsymbol{x}_n - \boldsymbol{\mu}_k)$$

$$\pi_k = \frac{N_k}{N}$$

Naive Bayes

Example: Spam Filtering

Features: Words in E-mail

$$x = \begin{array}{ll} \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} & \begin{array}{l} \text{a} \\ \text{aardvark} \\ \text{aardwolf} \\ \vdots \\ \text{buy} \\ \vdots \\ \text{zygmurgy} \end{array} \end{array}$$

Labels: Spam or not Spam

$$y_n \in \{0, 1\}$$

Naive Bayes

Features: Words in E-mail

$$x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \quad \begin{matrix} \text{a} \\ \text{aardvark} \\ \text{aardwolf} \\ \vdots \\ \text{buy} \\ \vdots \\ \text{zygmurgy} \end{matrix}$$

Generative Model

$$\begin{aligned} y_n &\sim \text{Bernoulli}(\mu) \\ \boldsymbol{x}_{nd} \,|\, y_n = k &\sim \text{Bernoulli}(\phi_{kd}) \end{aligned}$$

Conditional Independence

$$p(\boldsymbol{x}_n | y_n) = \prod_{d=1}^{D} p(x_{nd} | y_n)$$

Naive Bayes

Features: Words in E-mail

$$x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \text{a} \\ \text{aardvark} \\ \text{aardwolf} \\ \vdots \\ \text{buy} \\ \vdots \\ \text{zygmurgy} \end{matrix}$$

Generative Model

$$y_n \sim \text{Bernoulli}(\mu)$$
$$\boldsymbol{x}_{nd} \,|\, y_n = k \sim \text{Bernoulli}(\phi_{kd})$$

Maximum Likelihood

$$\mu = \frac{1}{N} \sum_{n=1}^{N} I[y_n = 1]$$
$$\phi_{kd} = \frac{1}{N_k} \sum_{n:y_n=k} I[x_{nd} = 1]$$

Online Estimation and Smoothing

Features: Words in E-mail

$$x = \begin{array}{c} \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \end{array} \quad \begin{array}{l} \text{a} \\ \text{aardvark} \\ \text{aardwolf} \\ \vdots \\ \text{buy} \\ \vdots \\ \text{zygmurgy} \end{array}$$

Suppose word *d* not in training set

$$\phi_{0d} = \phi_{1d} = 0$$

$$p(\boldsymbol{x}|y=0) = p(\boldsymbol{x}|y=1) = 0$$

Bayes Rule

$$p(\boldsymbol{y}\,|\,\boldsymbol{x}) = \frac{p(\boldsymbol{x}\,|\,\boldsymbol{y})p(\boldsymbol{y})}{p(\boldsymbol{x})} = \frac{0}{0}$$

Online Estimation and Smoothing

Generative model with prior

$$\begin{aligned} \mu &\sim \mathrm{Beta}(1,1) \\ \phi_{kd} &\sim \mathrm{Beta}(1,1) \\ y_n &\sim \mathrm{Bernoulli}(\mu) \\ x_{nd} \mid y_n = k &\sim \mathrm{Bernoulli}(\phi_{kd}) \end{aligned}$$

Posterior Mean

$$\mu^*, \boldsymbol{\phi}^* = \mathbb{E}_{p(\mu, \boldsymbol{\phi} \mid \boldsymbol{x}_{1:N}, y_{1:N})}[\mu, \boldsymbol{\phi}]$$

Conjugacy

$$\mathrm{Bin}(m|N,\mu) \quad = \quad \binom{N}{m}\mu^m(1-\mu)^{N-m}$$

$$\mathrm{Beta}(\mu|a,b) \quad = \quad \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

$$\begin{aligned} p(\mu\,|\,m) &= \frac{p(m,\mu)}{p(m)} \\ &\propto \mathrm{Bin}(m\,|\,N,\mu)\mathrm{Beta}(\mu\,|\,a,b) \\ &\propto \mu^{m+(a-1)}(1-\mu)^{(N-m)+(b-1)} \end{aligned}$$

Online Estimation and Smoothing

Generative model with prior

$$
\begin{aligned}
\mu &\sim \mathrm{Beta}(1,1) \\
\phi_{kd} &\sim \mathrm{Beta}(1,1) \\
y_n &\sim \mathrm{Bernoulli}(\mu) \\
x_{nd} \mid y_n = k &\sim \mathrm{Bernoulli}(\phi_{kd})
\end{aligned}
$$

Posterior Mean

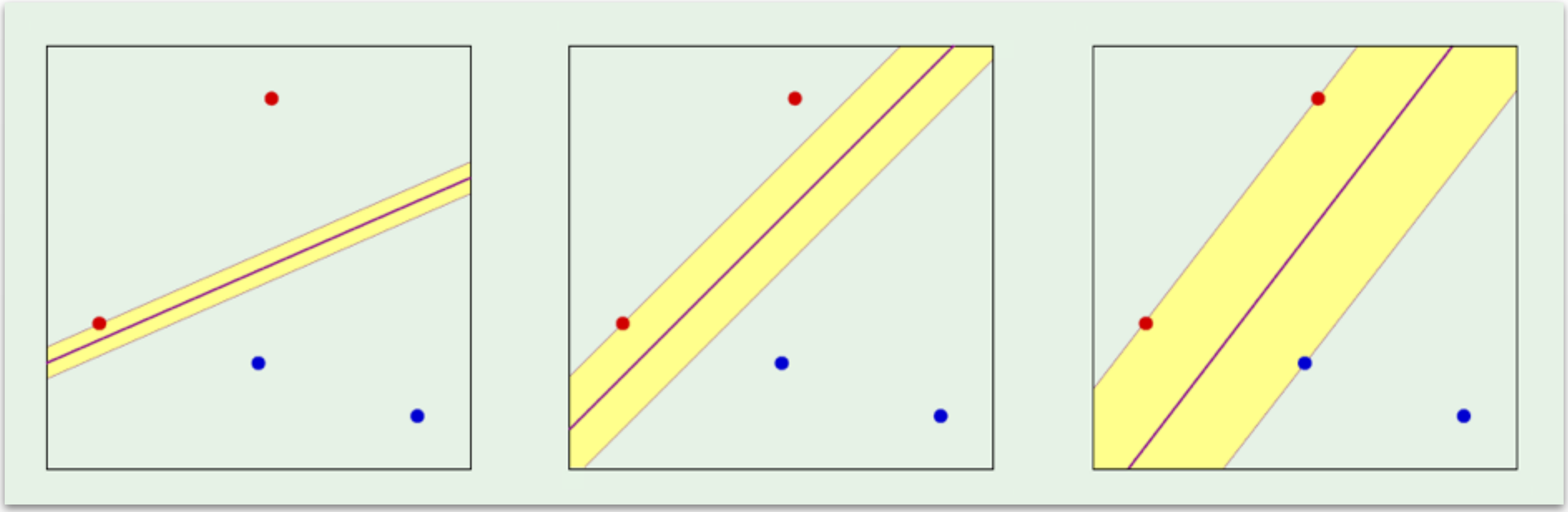
$$\mu^*, \boldsymbol{\phi}^* = \mathbb{E}_{p(\mu, \boldsymbol{\phi} \mid \boldsymbol{x}_{1:N}, y_{1:N})}[\mu, \boldsymbol{\phi}]$$

$$\mu^* = \frac{N_1 + 1}{N + 2}$$

$$\phi^*_{kd} = \frac{N_{kd} + 1}{N_k + 2}$$

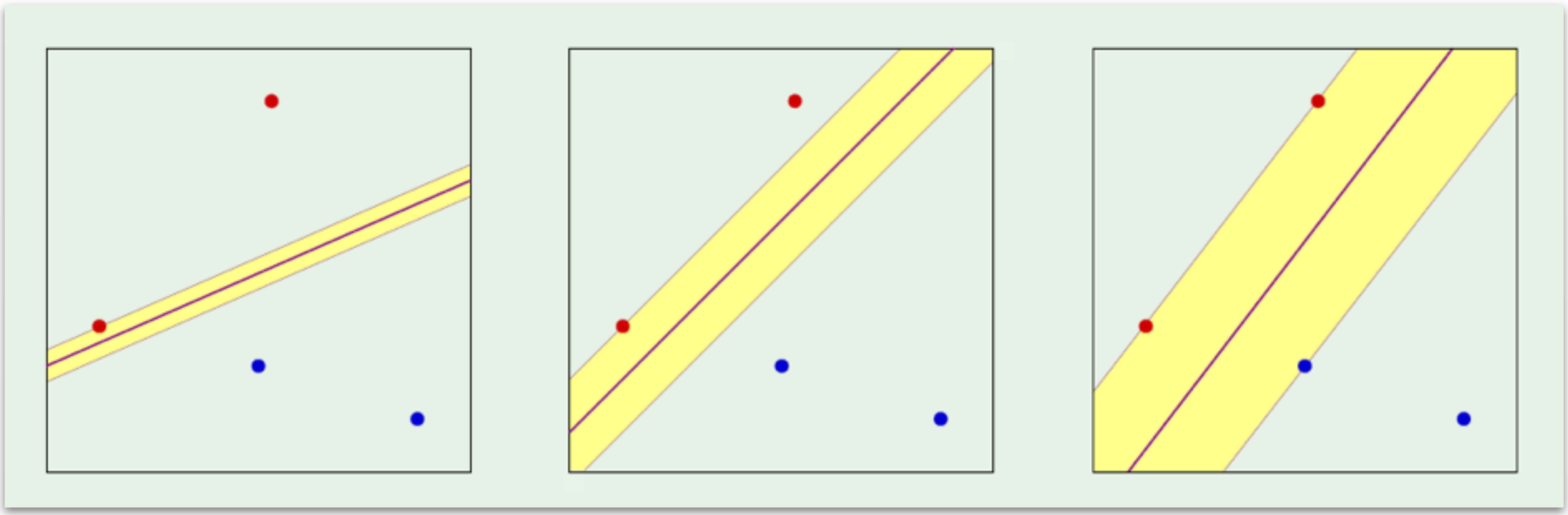
Support Vector Machines

Intuition



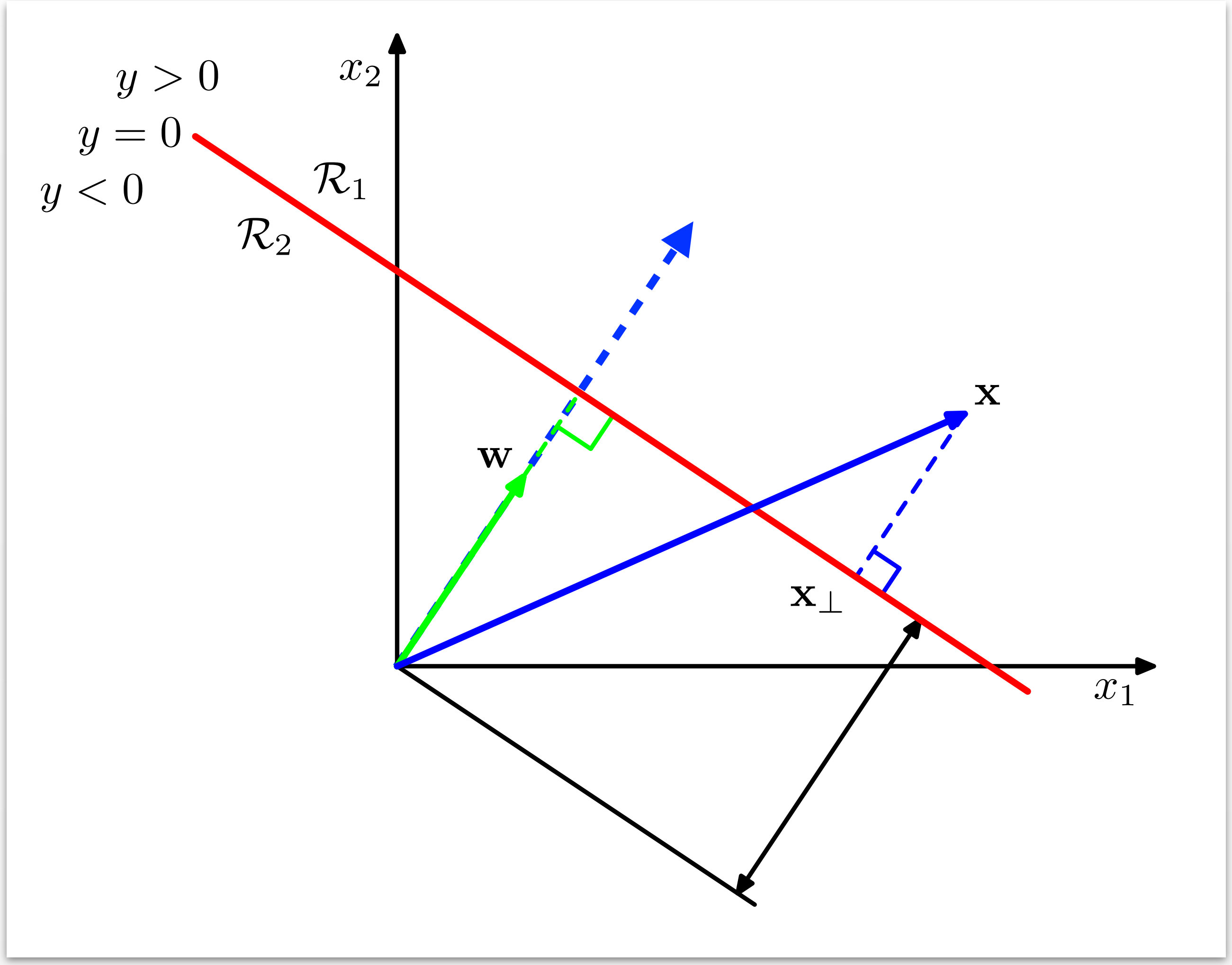
Which of these linear classifiers is the best?

Max Margin Classifiers



Idea: Maximize the *margin* between two *separable* classes

Max Margin Classifiers



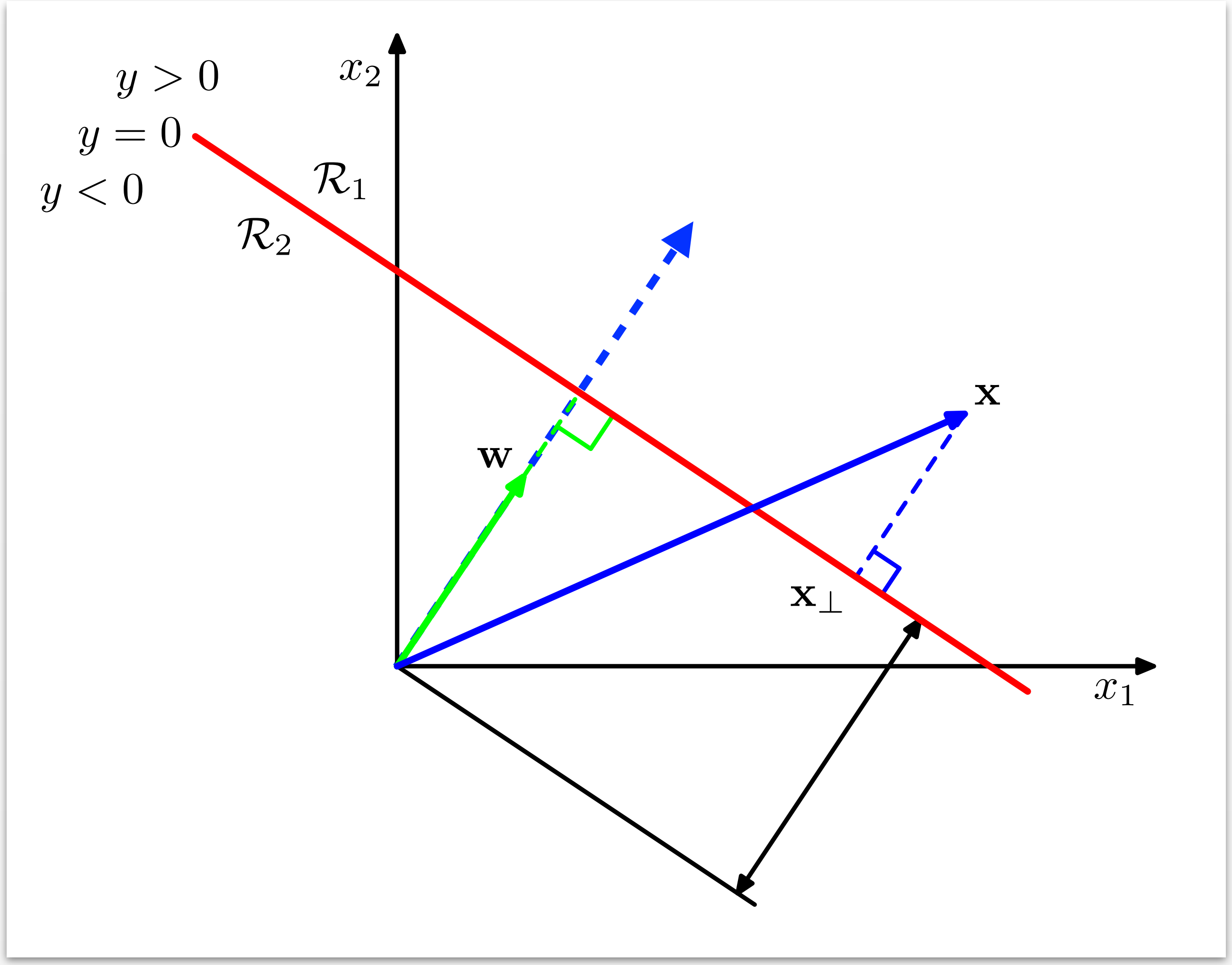
$$h(\boldsymbol{x}; \boldsymbol{w}, b) = \mathrm{Sign}(\boldsymbol{w}^\top \boldsymbol{x} + b)$$

$$y_n \in \{-1, 1\}$$

$$\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_D)$$

$$\boldsymbol{w} = (w_1, \ldots, w_D)$$

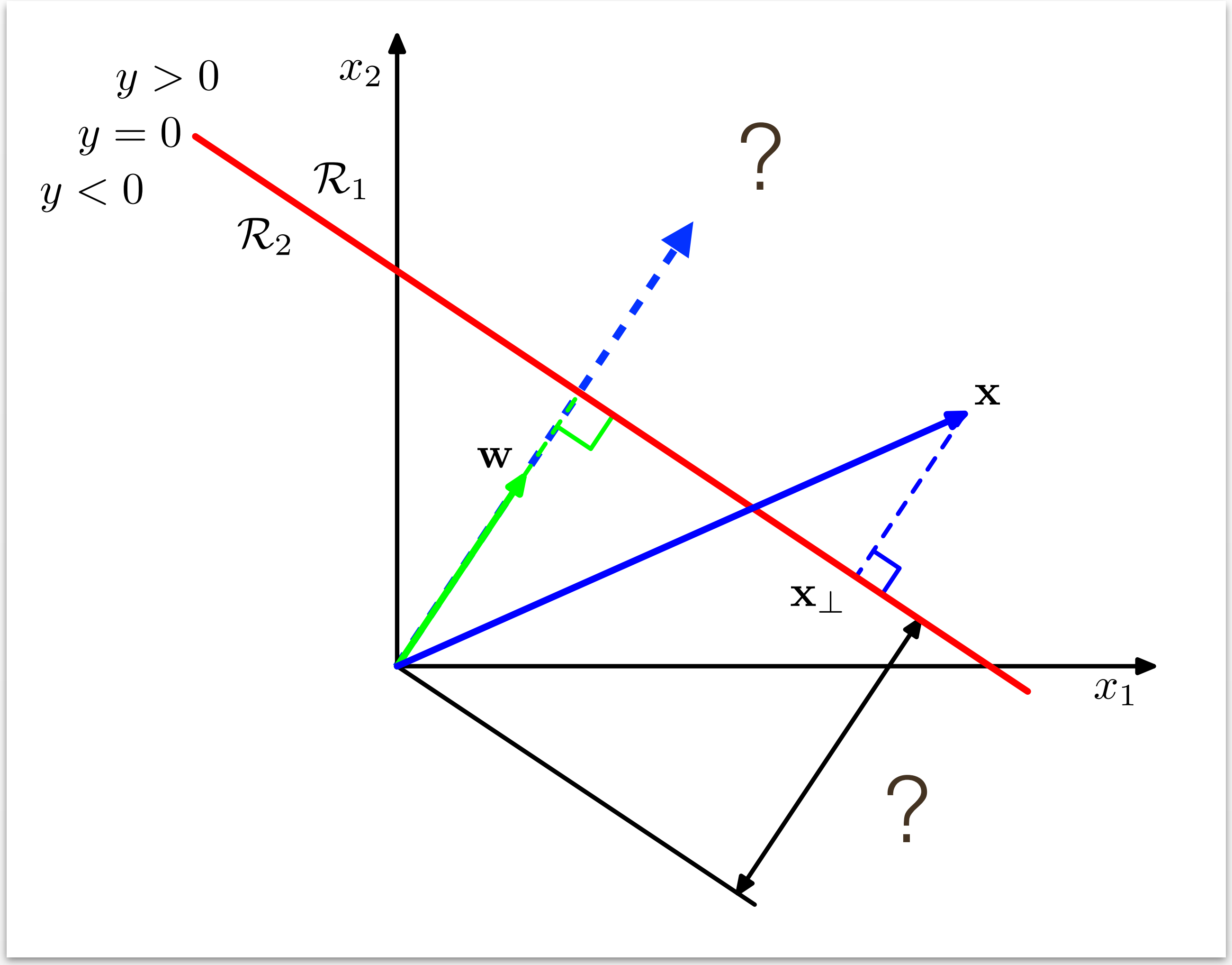
Max Margin Classifiers



$$\boldsymbol{w}^\top \boldsymbol{x} + b =$$

$$||\boldsymbol{w}|| \left(\frac{\boldsymbol{w}^\top \boldsymbol{x}}{||\boldsymbol{w}||} + \frac{b}{||\boldsymbol{w}||} \right)$$

Max Margin Classifiers

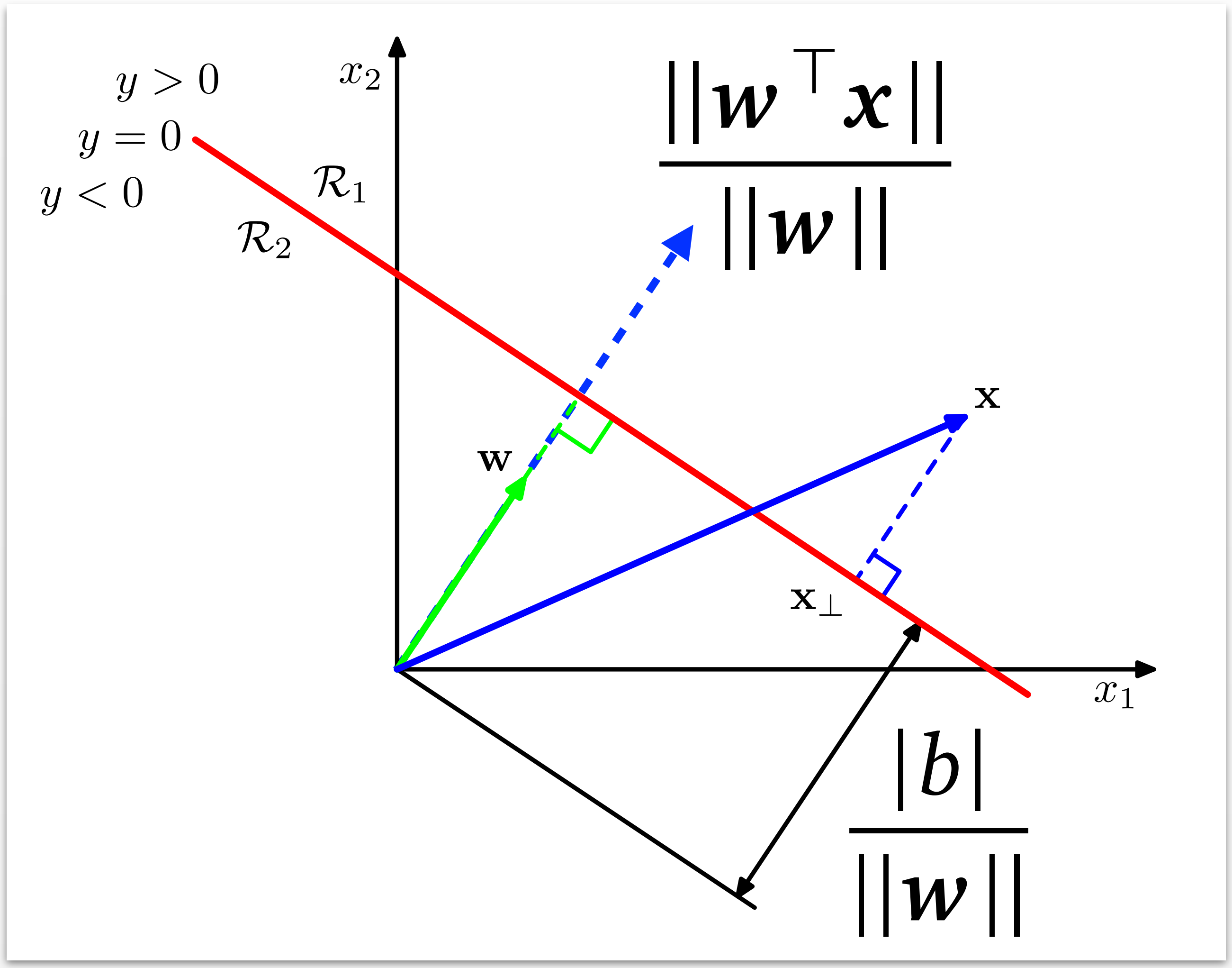


$$\boldsymbol{w}^\top \boldsymbol{x} + b =$$

$$||\boldsymbol{w}||\left(\frac{\boldsymbol{w}^\top \boldsymbol{x}}{||\boldsymbol{w}||} + \frac{b}{||\boldsymbol{w}||}\right)$$

What are the lengths of these vectors?

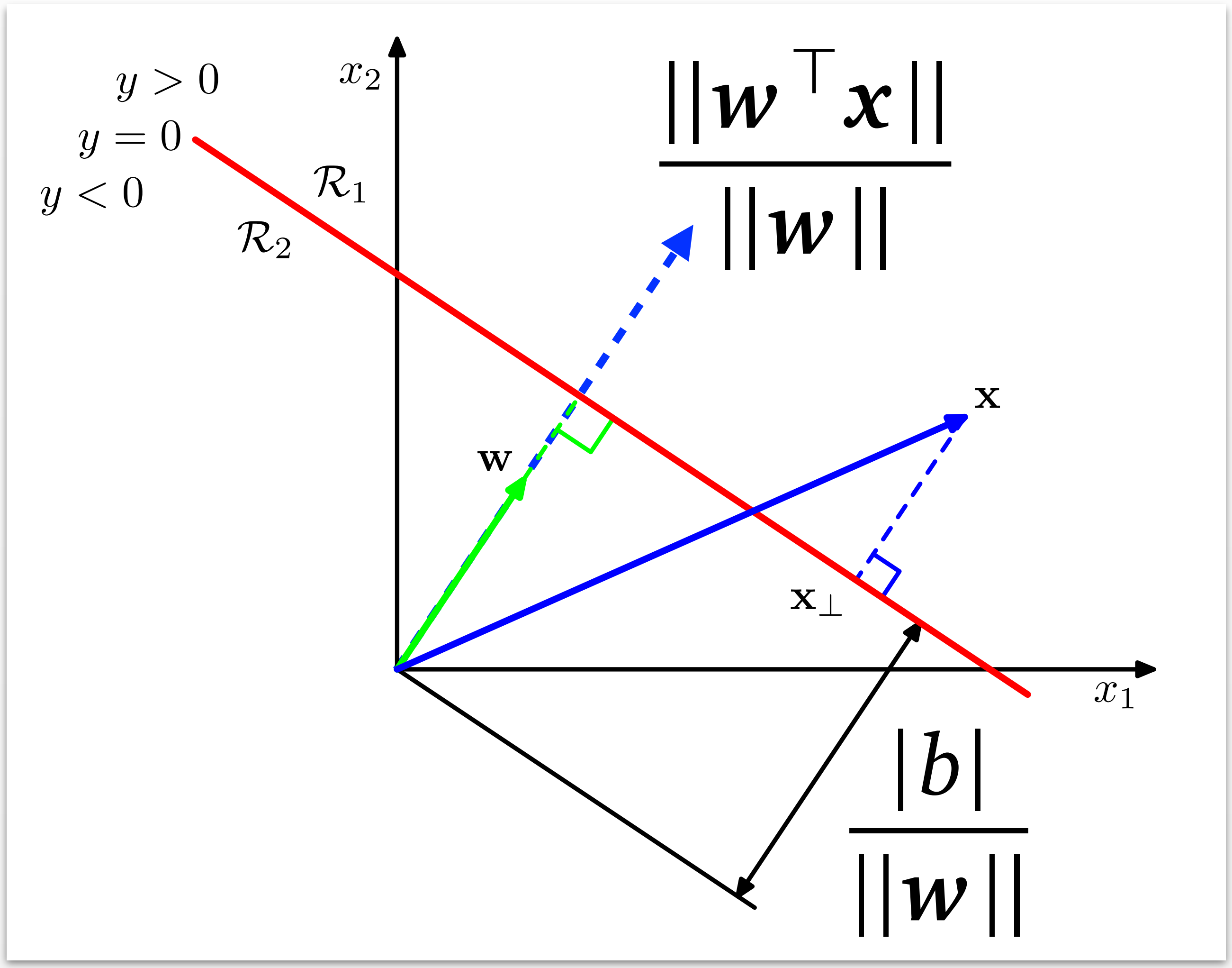
Max Margin Classifiers



$$\boldsymbol{w}^\top \boldsymbol{x} + b =$$

$$||\boldsymbol{w}|| \left(\frac{\boldsymbol{w}^\top \boldsymbol{x}}{||\boldsymbol{w}||} + \frac{b}{||\boldsymbol{w}||} \right)$$

Max Margin Classifiers



$$\boldsymbol{w}^\top \boldsymbol{x} + b = ||\boldsymbol{w}|| \left(\frac{\boldsymbol{w}^\top \boldsymbol{x}}{||\boldsymbol{w}||} + \frac{b}{||\boldsymbol{w}||} \right)$$

Distance from plane: $\frac{1}{||\boldsymbol{w}||}(\boldsymbol{w}^\top \boldsymbol{x} + b)$

Equivalent Optimization Problems

$$\max_{\boldsymbol{w},b} \hat{\gamma} \qquad y_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b) \geq \hat{\gamma} \qquad n = 1, \ldots, N$$

$$||\boldsymbol{w}|| = 1$$

Distance from plane: $\dfrac{1}{||\boldsymbol{w}||}\left(\boldsymbol{w}^\top \boldsymbol{x} + b\right)$

Equivalent Optimization Problems

$$\max_{\boldsymbol{w},b} \hat{\gamma} \qquad y_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b) \geq \hat{\gamma} \qquad n = 1, \ldots, N$$

$$||\boldsymbol{w}|| = 1$$

$$\max_{\boldsymbol{w},b,\gamma} \frac{\gamma}{||\boldsymbol{w}||} \qquad y_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b) \geq \gamma \qquad n = 1, \ldots, N$$

Distance from plane: $\dfrac{1}{||\boldsymbol{w}||}\left(\boldsymbol{w}^\top \boldsymbol{x} + b\right)$

Equivalent Optimization Problems

$$\max_{\boldsymbol{w},b} \hat{\gamma} \qquad y_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b) \geq \hat{\gamma} \qquad n = 1, \ldots, N$$

$$||\boldsymbol{w}|| = 1$$

$$\max_{\boldsymbol{w},b,\gamma} \frac{\gamma}{||\boldsymbol{w}||} \qquad y_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b) \geq \gamma \qquad n = 1, \ldots, N$$

$$\max_{\boldsymbol{w},b} \frac{1}{||\boldsymbol{w}||} \qquad y_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b) \geq 1 \qquad n = 1, \ldots, N$$

Distance from plane: $\frac{1}{||\boldsymbol{w}||}\left(\boldsymbol{w}^\top \boldsymbol{x} + b\right)$

Equivalent Optimization Problems

$$\max_{\boldsymbol{w},b} \hat{\gamma} \qquad y_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b) \geq \hat{\gamma} \qquad n = 1, \dots, N$$

$$||\boldsymbol{w}|| = 1$$

$$\max_{\boldsymbol{w},b,\gamma} \frac{\gamma}{||\boldsymbol{w}||} \qquad y_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b) \geq \gamma \qquad n = 1, \dots, N$$

$$\max_{\boldsymbol{w},b} \frac{1}{||\boldsymbol{w}||} \qquad y_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b) \geq 1 \qquad n = 1, \dots, N$$

$$\min_{\boldsymbol{w},b} \frac{1}{2}||\boldsymbol{w}||^2 \qquad y_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b) \geq 1 \qquad n = 1, \dots, N$$

Distance from plane: $\frac{1}{||\boldsymbol{w}||}\left(\boldsymbol{w}^\top \boldsymbol{x} + b\right)$

Equivalent Optimization Problems

$$\max_{\boldsymbol{w},b} \hat{\gamma} \qquad y_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b) \geq \hat{\gamma} \qquad n = 1, \ldots, N$$

$$||\boldsymbol{w}|| = 1$$

$$\max_{\boldsymbol{w},b,\gamma} \frac{\gamma}{||\boldsymbol{w}||} \qquad y_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b) \geq \gamma \qquad n = 1, \ldots, N$$

$$\max_{\boldsymbol{w},b} \frac{1}{||\boldsymbol{w}||} \qquad y_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b) \geq 1 \qquad n = 1, \ldots, N$$

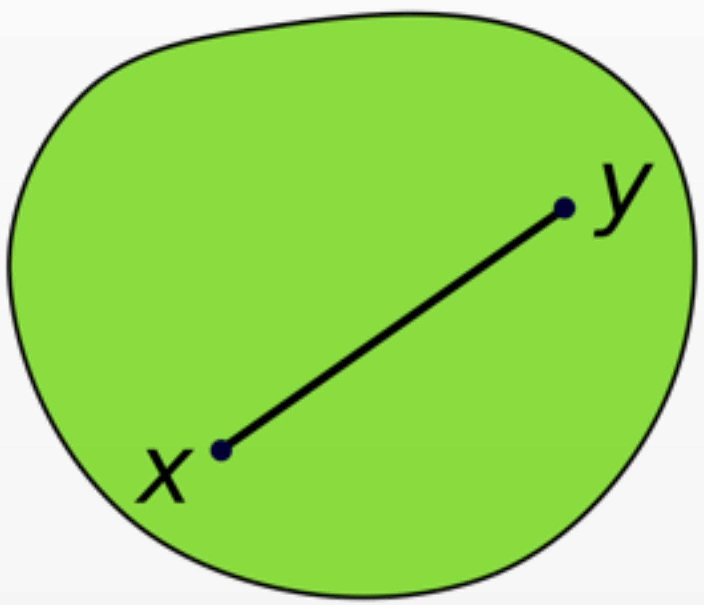
$$\min_{\boldsymbol{w},b} \frac{1}{2}||\boldsymbol{w}||^2 \qquad y_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b) \geq 1 \qquad n = 1, \ldots, N$$

Distance from plane: $\frac{1}{||\boldsymbol{w}||}\left(\boldsymbol{w}^\top \boldsymbol{x} + b\right)$

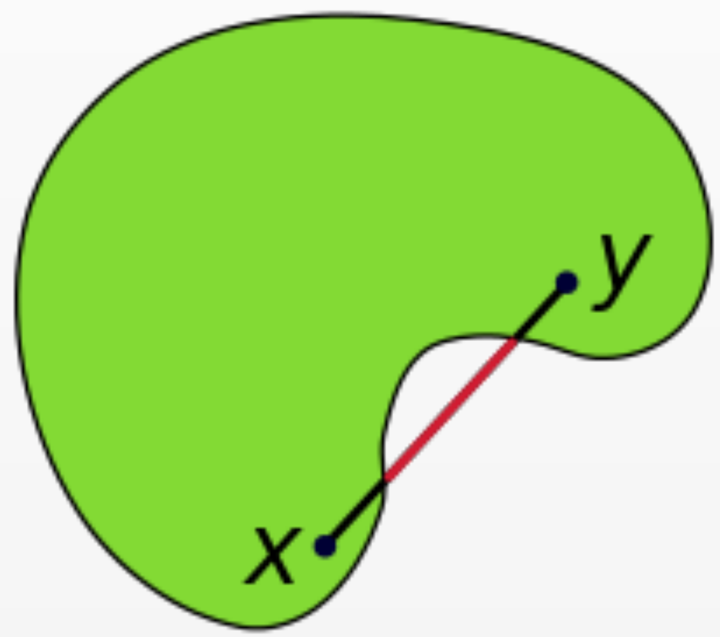
Intermezzo

Convex Optimization

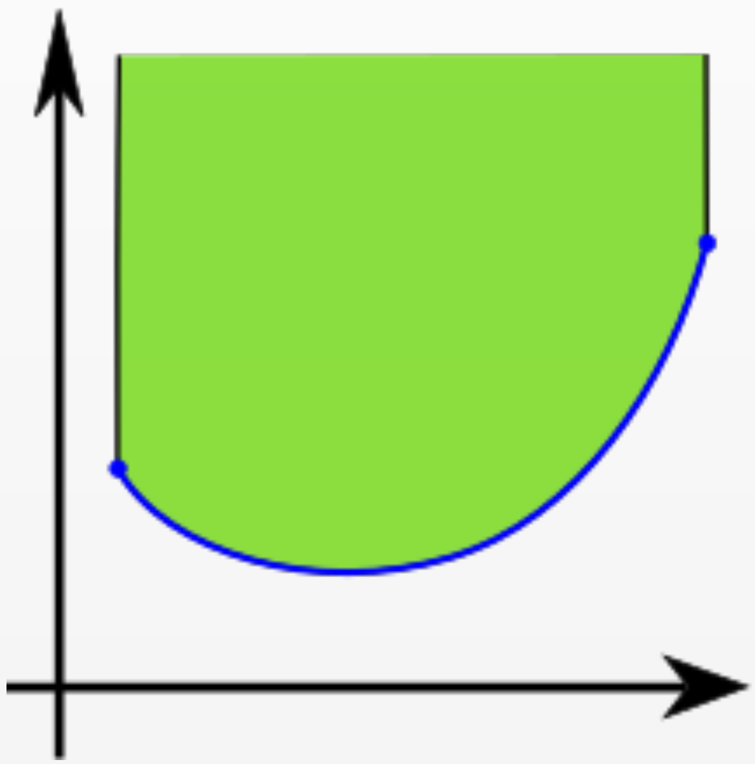
Convex Sets and Functions



Convex
Set



Non-convex
Set



Convex
Function

Lagrange Duality

Constrained Optimization Problem

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & h_i(w) = 0, \quad i = 1, \ldots, l. \end{aligned}$$

Lagrangian

$$\mathcal{L}(w, \beta) = f(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

Optimum

$$\frac{\partial \mathcal{L}}{\partial w_i} = 0; \quad \frac{\partial \mathcal{L}}{\partial \beta_i} = 0,$$

Lagrange Duality

Primal Optimization Problem

$$
\begin{aligned}
\min_w \quad & f(w) \\
\text{s.t.} \quad & g_i(w) \le 0, \quad i = 1, \ldots, k \\
& h_i(w) = 0, \quad i = 1, \ldots, l.
\end{aligned}
$$

Generalized Lagrangian

$$
\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w).
$$

Lagrange Duality

Primal Optimization Problem

$$\theta_{\mathcal{P}}(w) = \max_{\alpha, \beta \,:\, \alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta).$$

Generalized Lagrangian

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w).$$

Lagrange Duality

Primal Optimization Problem

$$\theta_{\mathcal{P}}(w) = \begin{cases} f(w) & \text{if } w \text{ satisfies primal constraints} \\ \infty & \text{otherwise.} \end{cases}$$

Generalized Lagrangian

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w).$$

Lagrange Duality

Primal Optimization Problem

$$\min_{w} \theta_{\mathcal{P}}(w) = \min_{w} \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta),$$

Generalized Lagrangian

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w).$$

Lagrange Duality

Dual Optimization Problem

$$\theta_{\mathcal{D}}(\alpha, \beta) = \min_{w} \mathcal{L}(w, \alpha, \beta).$$

Generalized Lagrangian

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w).$$

Lagrange Duality

Dual Optimization Problem

$$\max_{\alpha,\beta\,:\,\alpha_i \geq 0} \theta_{\mathcal{D}}(\alpha, \beta) = \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \min_{w} \mathcal{L}(w, \alpha, \beta).$$

Generalized Lagrangian

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w).$$

Lagrange Duality

Relationship between Primal and Dual

$$d^* = \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \min_w \mathcal{L}(w, \alpha, \beta) \;\; \textbf{?} \;\; \min_w \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta) = p^*.$$

Generalized Lagrangian

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w).$$

Lagrange Duality

Relationship between Primal and Dual

$$d^* = \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \min_w \mathcal{L}(w, \alpha, \beta) \leq \min_w \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta) = p^*.$$

Generalized Lagrangian

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w).$$

Lagrange Duality

Duality gap p-d* is zero when*

$$f \quad \text{convex}, \qquad g_i \quad \text{convex}, \qquad h_i \quad \text{affine}$$

Generalized Lagrangian

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w).$$

Lagrange Duality

Karush-Kuhn-Tucker (KKT) conditions at optimum

$$
\begin{aligned}
\frac{\partial}{\partial w_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, n \\
\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, l \\
\alpha_i^* g_i(w^*) &= 0, \quad i = 1, \ldots, k \\
g_i(w^*) &\leq 0, \quad i = 1, \ldots, k \\
\alpha^* &\geq 0, \quad i = 1, \ldots, k
\end{aligned}
$$

Dual complementarity

$g_i(w) = 0$ when $\alpha_i > 0$, $\alpha_i = 0$ when $g_i(w) < 0$

(back to SVMs)

Equivalent Optimization Problems

$$\max_{\boldsymbol{w},b} \hat{\gamma} \qquad y_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b) \geq \hat{\gamma} \qquad n = 1, \ldots, N$$

$$||\boldsymbol{w}|| = 1$$

$$\max_{\boldsymbol{w},b,\gamma} \frac{\gamma}{||\boldsymbol{w}||} \qquad y_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b) \geq \gamma \qquad n = 1, \ldots, N$$

$$\max_{\boldsymbol{w},b} \frac{1}{||\boldsymbol{w}||} \qquad y_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b) \geq 1 \qquad n = 1, \ldots, N$$

$$\min_{\boldsymbol{w},b} \frac{1}{2}||\boldsymbol{w}||^2 \qquad y_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b) \geq 1 \qquad n = 1, \ldots, N$$

SVMs as Convex Optimization

$$\min_{\boldsymbol{w},b} \frac{1}{2}||\boldsymbol{w}||^2 \qquad y_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b) \geq 1 \qquad n = 1, \ldots, N$$

SVMs as Convex Optimization

$$\min_{\boldsymbol{w},b} \frac{1}{2}||\boldsymbol{w}||^2 \qquad y_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b) \geq 1 \qquad n = 1, \ldots, N$$

Write as Convex Optimization Problem

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w).$$

SVMs as Convex Optimization

$$\min_{\boldsymbol{w},b} \frac{1}{2}||\boldsymbol{w}||^2 \qquad y_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b) \geq 1 \qquad n = 1, \ldots, N$$

Write as Convex Optimization Problem

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w).$$

$$f(\boldsymbol{w}) = \frac{1}{2}||\boldsymbol{w}||^2$$

$$g_i(w) = -y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) + 1 \leq 0$$

SVMs as Convex Optimization

$$\min_{\boldsymbol{w},b} \frac{1}{2}||\boldsymbol{w}||^2 \qquad y_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b) \geq 1 \qquad n = 1, \ldots, N$$

Generalized Lagrangian

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)}(w^T x^{(i)} + b) - 1 \right].$$

(note: no equality constraints)

Dual Form

Generalized Lagrangian

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)}(w^T x^{(i)} + b) - 1 \right].$$

Dual problem

$$\theta_D(\alpha) = \min_{\boldsymbol{w}, b} \mathscr{L}(w, b, \alpha)$$

Solve for w

$$\nabla_w \mathcal{L}(w, b, \alpha) = w - \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} = 0$$

Solve for b

$$\frac{\partial}{\partial b} \mathcal{L}(w, b, \alpha) = \sum_{i=1}^{m} \alpha_i y^{(i)} = 0.$$

Dual Form

Generalized Lagrangian

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)}(w^T x^{(i)} + b) - 1 \right].$$

Dual problem

$$\theta_D(\alpha) = \textbf{?}$$

Solve for w

$$\nabla_w \mathcal{L}(w, b, \alpha) = w - \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} = 0$$

Solve for b

$$\frac{\partial}{\partial b} \mathcal{L}(w, b, \alpha) = \sum_{i=1}^{m} \alpha_i y^{(i)} = 0.$$

Dual Form

Generalized Lagrangian

$$\mathcal{L}(w,b,\alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)}(w^T x^{(i)} + b) - 1 \right].$$

Dual problem

$$\theta_D(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)} - b \sum_{i=1}^{m} \alpha_i y^{(i)}.$$

Solve for w

$$\nabla_w \mathcal{L}(w,b,\alpha) = w - \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} = 0$$

Solve for b

$$\frac{\partial}{\partial b}\mathcal{L}(w,b,\alpha) = \sum_{i=1}^{m} \alpha_i y^{(i)} = 0.$$

Dual Form

Generalized Lagrangian

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)}(w^T x^{(i)} + b) - 1 \right].$$

Dual problem

$$\theta_D(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)}$$

Dual Form

Generalized Lagrangian

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)}(w^T x^{(i)} + b) - 1 \right].$$

Dual problem

$$\theta_D(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)}$$

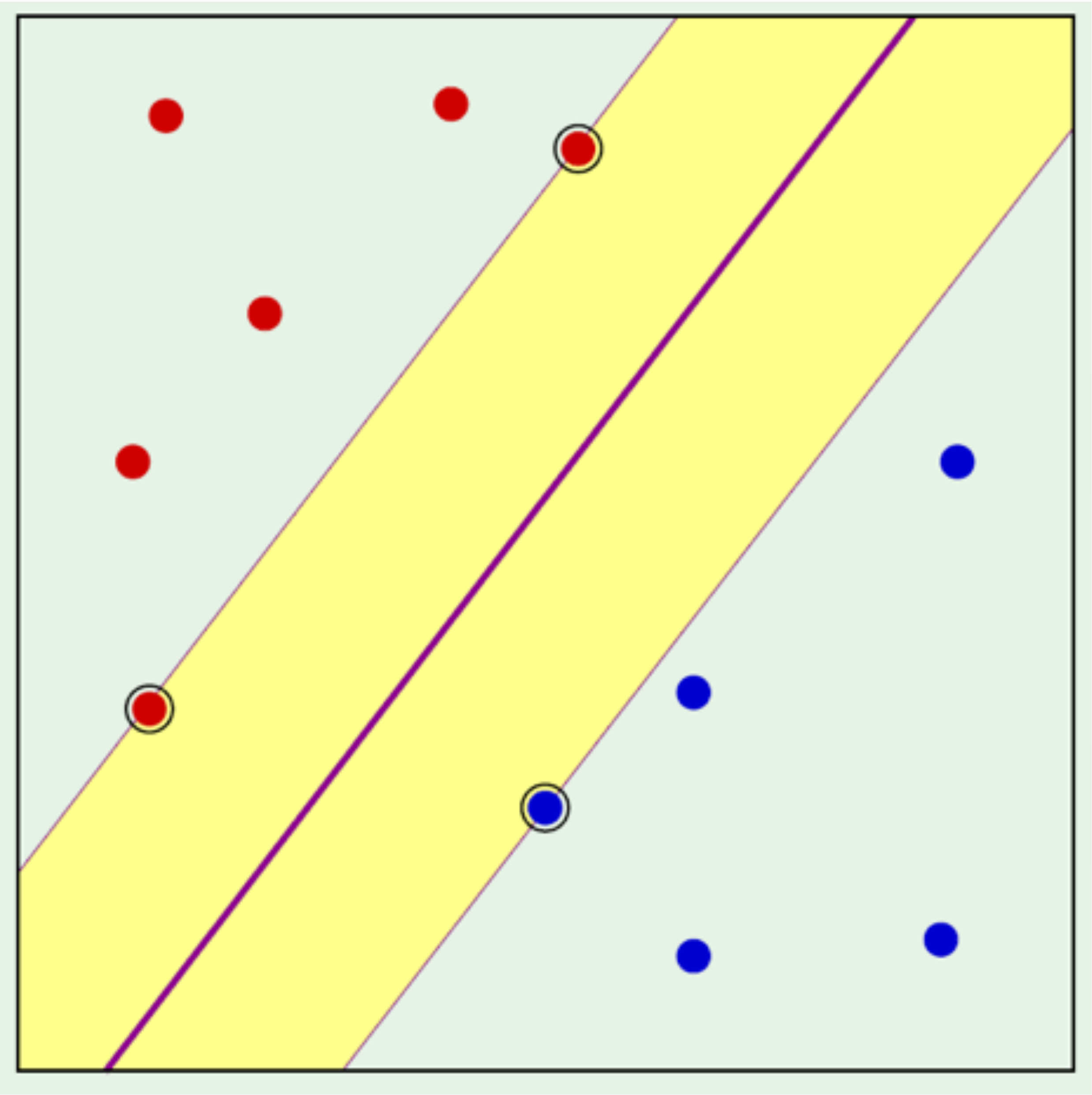
Compute w

$$w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}.$$

Compute b

$$b^* = -\frac{\max_{i:y^{(i)}=-1} w^{*T} x^{(i)} + \min_{i:y^{(i)}=1} w^{*T} x^{(i)}}{2}.$$

Support Vectors



Compute w

$$w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}.$$

Compute b

$$b^* = -\frac{\max_{i:y^{(i)}=-1} w^{*T} x^{(i)} + \min_{i:y^{(i)}=1} w^{*T} x^{(i)}}{2}.$$

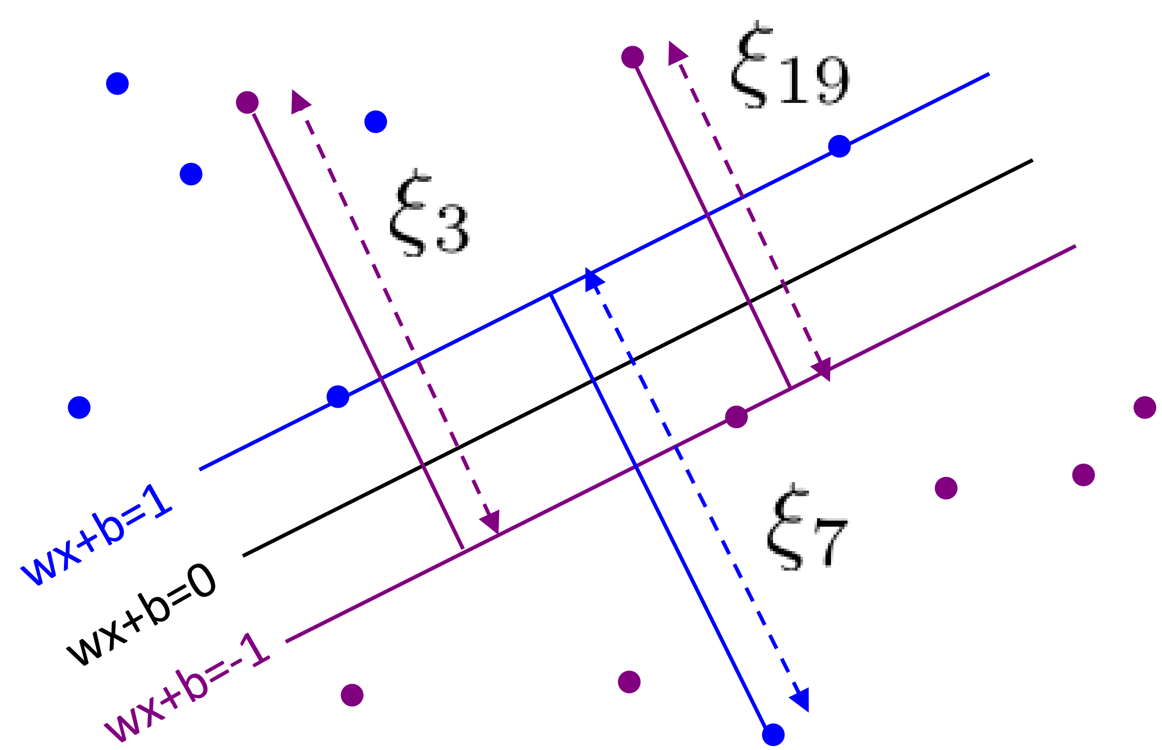
Dual complementarity

- $a_i = 0$ when $g_i(w) < 0$
- $g_i(w) = 0$ when $a_i > 0$

SVM with Non-Separable Data

$$\arg\min_{\mathbf{w},b,\xi\geq 0} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{m}\xi_i$$

$$\text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m$$



SVM with Non-Separable Data

Generalized Lagrangian

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\alpha_i(1 - \xi_i - y_i(\mathbf{w}^T\mathbf{x}_i + b)) - \sum_{i=1}^{m}\lambda_i\xi_i$$

Solve for Dual Form

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{m} y_i\alpha_i = 0$$

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i \Rightarrow \mathbf{w} = \sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i$$

$$\frac{\partial L}{\partial \xi} = C - \alpha_i - \lambda_i = 0$$

SVM with Non-Separable Data

Dual Optimization Problem

$$\arg\max_{\alpha \geq 0} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\text{s.t.} \quad \sum_{i=1}^{m} \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C, i = 1, \ldots, m$$

Inner Products

Dual Optimization Problem

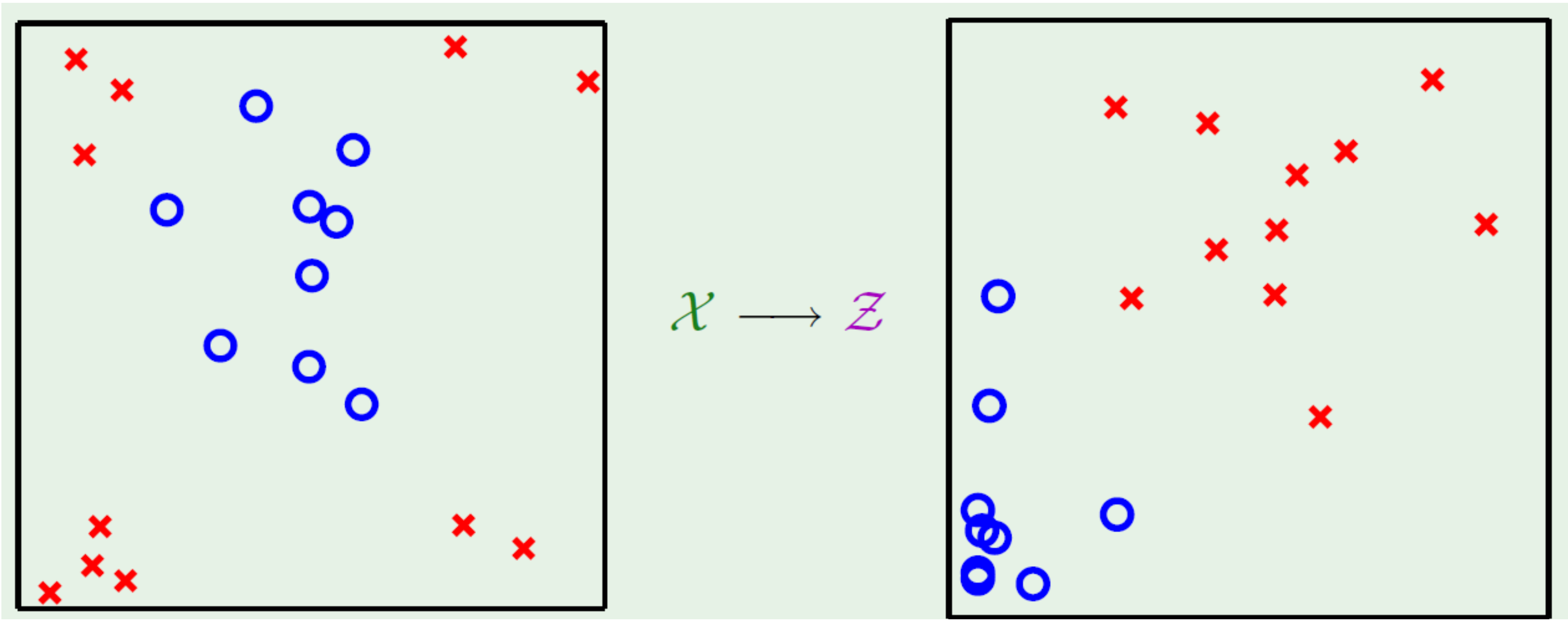
$$\begin{aligned} \max_\alpha \quad & W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle. \\ \text{s.t.} \quad & \alpha_i \geq 0, \quad i = 1, \ldots, m \\ & \sum_{i=1}^{m} \alpha_i y^{(i)} = 0, \end{aligned}$$

Prediction

$$\begin{aligned} w^T x + b &= \left(\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \right)^T x + b \\ &= \sum_{i=1}^{m} \alpha_i y^{(i)} \langle x^{(i)}, x \rangle + b. \end{aligned}$$

Next Lecture: Nonlinear SVMs

Let $\mathbf{z} = \Phi(\mathbf{x})$ for some function Φ:



Apply SVM in the $\mathbf{z}$ space by maximizing:

$$L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j \mathbf{z}_i^{\mathbf{T}} \mathbf{z}_j$$