The Mathematics of Machine Learning 2025

Lecture 8

Tim van Erven

University of Amsterdam

Outline

Support Vector Machines

1. Intro
2. Optimal Separating Hyperplane
3. SVMs
4. Interpretation as Penalized ERM with Hinge Loss
5. Dual Formulation
6. Kernel Trick

History

Vapnik & Chervonenkis: foundational work on statistical learning starting in sixties and seventies, leading to SVMs around 1990

Vapnik, 1998: “Solve the problem directly and never solve a more general problem as an intermediate step.”

Intro

Plug-in Estimators:

	Estimate	Examples
Generative	$P^*(X, Y)$	Naive Bayes
Discriminative	$P^*(Y \mid X)$	LDA, logistic regression
Now	decision boundary directly	SVMs

Intro

Plug-in Estimators:

	Estimate	Examples
Generative	$P^*(X, Y)$	Naive Bayes
Discriminative	$P^*(Y \mid X)$	LDA, logistic regression
Now	decision boundary directly	SVMs

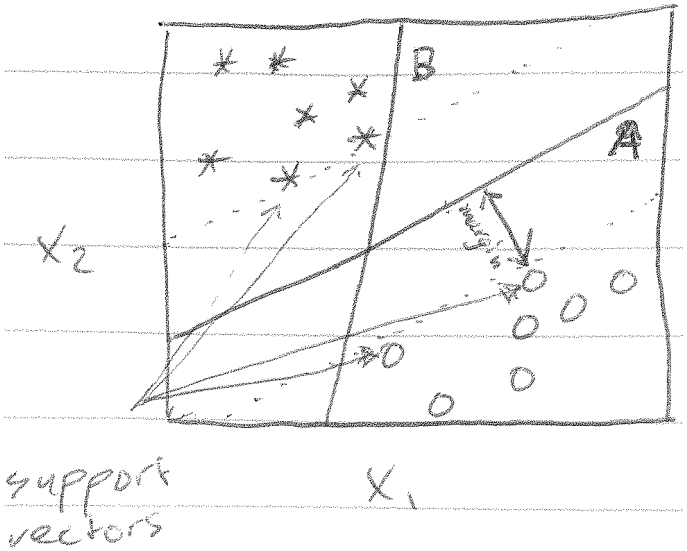
Binary Classification:

$$\begin{pmatrix} Y_1 \\ X_1 \end{pmatrix}, \ldots, \begin{pmatrix} Y_N \\ X_N \end{pmatrix} \qquad Y \in \{-1, +1\}$$

Linear Classifiers:

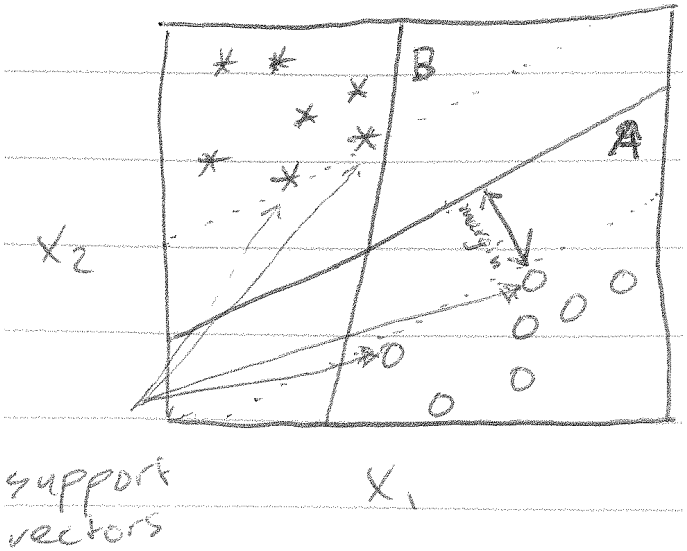
$$f_{\beta_0,\beta}(X) = X^\top \beta + \beta_0 \qquad \beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p$$

Optimal Separating Hyperplane I



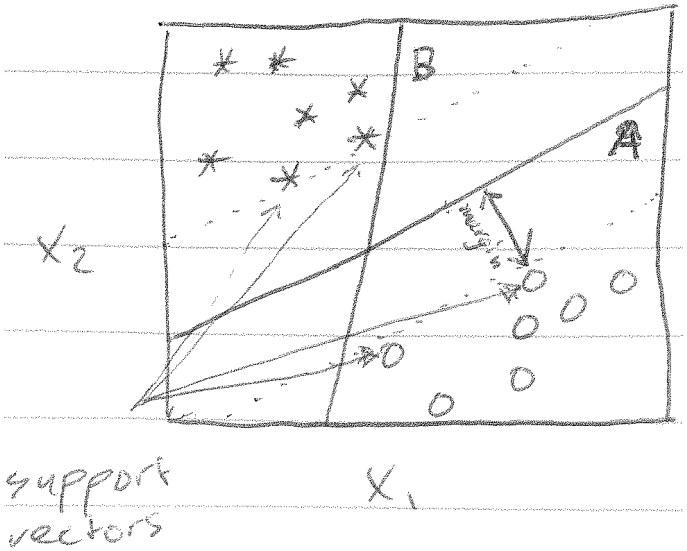
Which decision boundary do you like best? A or B?

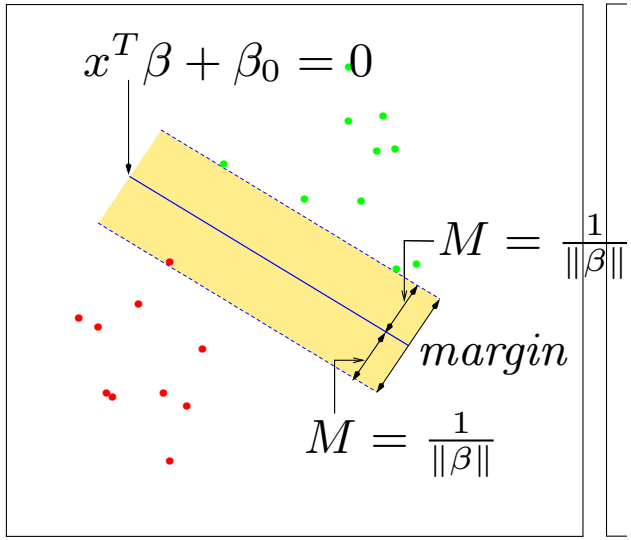
Optimal Separating Hyperplane I



Which decision boundary do you like best? A or B?
A has larger "margin" = distance to nearest point

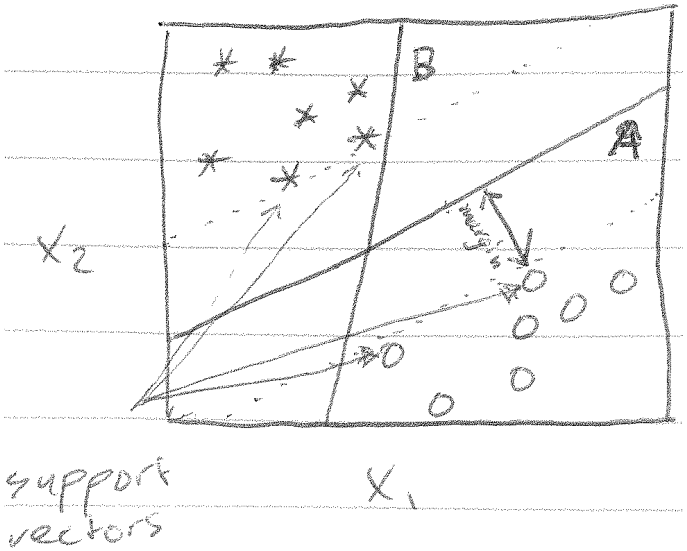
Optimal Separating Hyperplane I

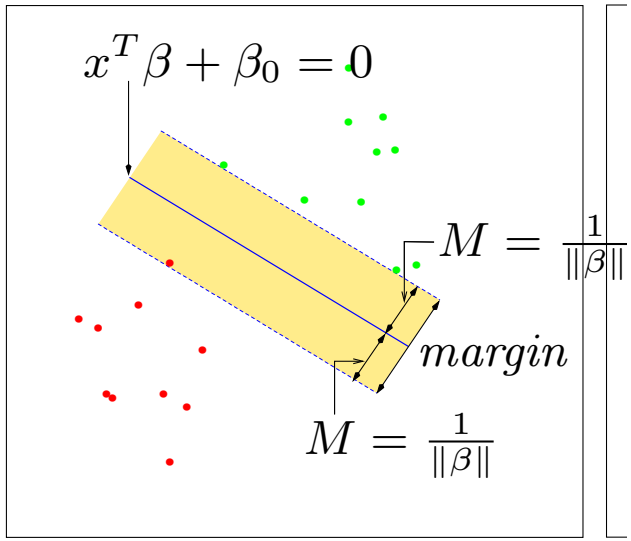




Which decision boundary do you like best? A or B?
A has larger "margin" = distance to nearest point

Optimal Separating Hyperplane I

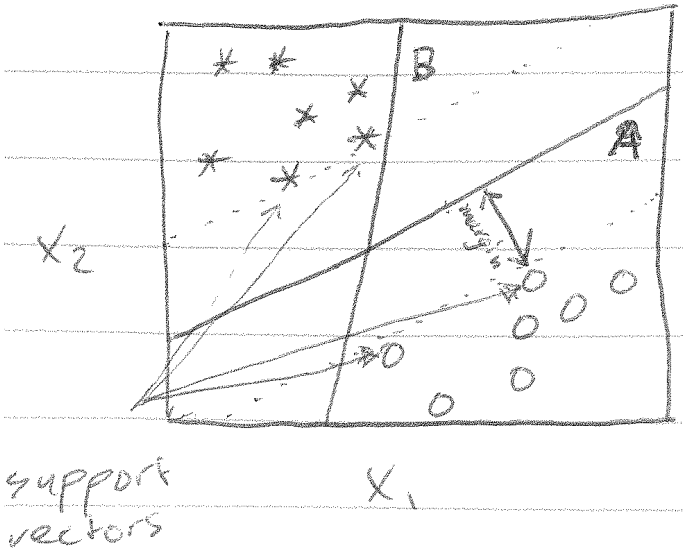


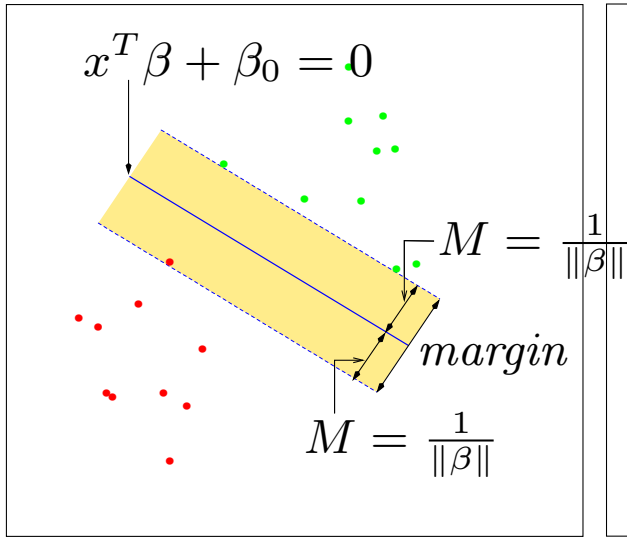


Which decision boundary do you like best? A or B?
A has larger "margin" = distance to nearest point

- **Optimal Separating Hyperplane:** maximizes the margin
- **Support Vectors:** points exactly at the margin
- **Decision Boundary:** $\{x \mid x^\top \beta + \beta_0 = 0\}$

Optimal Separating Hyperplane I





Which decision boundary do you like best? A or B?
A has larger "margin" = distance to nearest point

- **Optimal Separating Hyperplane:** maximizes the margin
- **Support Vectors:** points exactly at the margin
- **Decision Boundary:** $\{x \mid x^\top \beta + \beta_0 = 0\}$

Distance of X_i to decision boundary: $\frac{|X_i^\top \beta + \beta_0|}{\|\beta\|} = \frac{|f_{\beta_0,\beta}(X_i)|}{\|\beta\|}$

Margin: $\frac{Y_i f_{\beta_0,\beta}(X_i)}{\|\beta\|} = \begin{cases} \text{distance if } X_i \text{ classified correctly} \\ \text{negative distance if } X_i \text{ classified incorrectly} \end{cases}$

Optimal Separating Hyperplane II

How do we express optimal separating hyperplane mathematically?

Optimal Separating Hyperplane II

How do we express optimal separating hyperplane mathematically?

$$\max_{\beta_0, \beta, M} \quad M$$

margin

$$\text{subject to:} \quad \frac{Y_i(X_i^\top \beta + \beta_0)}{\|\beta\|} \geq M \qquad \text{for } i = 1, \ldots, N.$$

Optimal Separating Hyperplane II

How do we express optimal separating hyperplane mathematically?

$$\max_{\beta_0,\beta,M} \quad M$$

margin

$$\text{subject to:} \quad \frac{Y_i(X_i^\top \beta + \beta_0)}{\|\beta\|} \geq M \qquad \text{for } i = 1, \ldots, N.$$

Same decision boundary and margin if we multiply β_0, β by a constant $c > 0$, so we can always choose c such that $\|\beta\| = 1/M$:

$$\max_{\beta_0,\beta,M} \quad M$$

$$\|\beta\| = \frac{1}{M}$$

$$\text{subject to:} \quad Y_i(X_i^\top \beta + \beta_0) \geq M\|\beta\| \qquad \text{for } i = 1, \ldots, N.$$

Optimal Separating Hyperplane II

How do we express optimal separating hyperplane mathematically?

$$\max_{\beta_0,\beta,M} \quad M$$

margin

$$\text{subject to:} \quad \frac{Y_i(X_i^\top \beta + \beta_0)}{\|\beta\|} \geq M \qquad \text{for } i = 1, \ldots, N.$$

Same decision boundary and margin if we multiply β_0, β by a constant $c > 0$, so we can always choose c such that $\|\beta\| = 1/M$:

$$\max_{\beta_0,\beta,M} \quad M$$

$$\|\beta\| = \frac{1}{M}$$

$$\text{subject to:} \quad Y_i(X_i^\top \beta + \beta_0) \geq 1 \qquad \text{for } i = 1, \ldots, N.$$

Optimal Separating Hyperplane II

$$\max_{\beta_0,\beta,M} \quad M$$

$$\|\beta\| = \frac{1}{M}$$

subject to: $\quad Y_i(X_i^\top \beta + \beta_0) \geq 1 \quad$ for $i = 1, \ldots, N$.

Optimal Separating Hyperplane II

$$\begin{aligned} &\max_{\beta_0,\beta,M} && M \\ &\|\beta\|=\frac{1}{M} \\ &\text{subject to:} && Y_i(X_i^\top \beta + \beta_0) \geq 1 && \text{for } i = 1, \ldots, N. \end{aligned}$$

$$\begin{aligned} &\max_{\beta_0,\beta} && \frac{1}{\|\beta\|} \\ &\text{subject to:} && Y_i(X_i^\top \beta + \beta_0) \geq 1 && \text{for } i = 1, \ldots, N. \end{aligned}$$

Optimal Separating Hyperplane II

$$\max_{\beta_0,\beta,M} \quad M$$
$$\|\beta\| = \frac{1}{M}$$
$$\text{subject to:} \quad Y_i(X_i^\top \beta + \beta_0) \geq 1 \quad \text{for } i = 1, \ldots, N.$$

---

solution achieved by same parameters

$$\max_{\beta_0,\beta} \quad \frac{1}{\|\beta\|}$$
$$\text{subject to:} \quad Y_i(X_i^\top \beta + \beta_0) \geq 1 \quad \text{for } i = 1, \ldots, N.$$

$$\min_{\beta_0,\beta} \quad \frac{1}{2}\|\beta\|^2$$
$$\text{subject to:} \quad Y_i(X_i^\top \beta + \beta_0) \geq 1 \quad \text{for } i = 1, \ldots, N.$$

Optimal Separating Hyperplane II

$$\max_{\beta_0,\beta,M} \quad M$$
$$\|\beta\| = \frac{1}{M}$$
$$\text{subject to:} \quad Y_i(X_i^\top \beta + \beta_0) \geq 1 \quad \text{for } i = 1, \ldots, N.$$

$$\max_{\beta_0,\beta} \quad \frac{1}{\|\beta\|}$$
$$\text{subject to:} \quad Y_i(X_i^\top \beta + \beta_0) \geq 1 \quad \text{for } i = 1, \ldots, N.$$

---

convex function →

$$\min_{\beta_0,\beta} \quad \frac{1}{2}\|\beta\|^2$$

linear inequality constraints →

$$\text{subject to:} \quad Y_i(X_i^\top \beta + \beta_0) \geq 1 \quad \text{for } i = 1, \ldots, N.$$

Optimal Separating Hyperplane II

$$\max_{\beta_0,\beta,M} \quad M$$
$$\|\beta\| = \frac{1}{M}$$
$$\text{subject to:} \quad Y_i(X_i^\top \beta + \beta_0) \geq 1 \qquad \text{for } i = 1, \ldots, N.$$

$$\max_{\beta_0,\beta} \quad \frac{1}{\|\beta\|}$$
$$\text{subject to:} \quad Y_i(X_i^\top \beta + \beta_0) \geq 1 \qquad \text{for } i = 1, \ldots, N.$$

---

convex function →
$$\min_{\beta_0,\beta} \quad \frac{1}{2}\|\beta\|^2$$
linear inequality constraints →
$$\text{subject to:} \quad Y_i(X_i^\top \beta + \beta_0) \geq 1 \qquad \text{for } i = 1, \ldots, N.$$

Convex optimization problem: **can solve efficiently!**

Optimal Separating Hyperplane II

$$\max_{\beta_0,\beta,M} \quad M$$
$$\|\beta\| = \frac{1}{M}$$
$$\text{subject to:} \quad Y_i(X_i^\top \beta + \beta_0) \geq 1 \qquad \text{for } i = 1, \ldots, N.$$

$$\max_{\beta_0,\beta} \quad \frac{1}{\|\beta\|}$$
$$\text{subject to:} \quad Y_i(X_i^\top \beta + \beta_0) \geq 1 \qquad \text{for } i = 1, \ldots, N.$$

---

convex function →

$$\min_{\beta_0,\beta} \quad \frac{1}{2}\|\beta\|^2$$

linear inequality constraints →

$$\text{subject to:} \quad Y_i(X_i^\top \beta + \beta_0) \geq 1 \qquad \text{for } i = 1, \ldots, N.$$

Convex optimization problem: **can solve efficiently!**

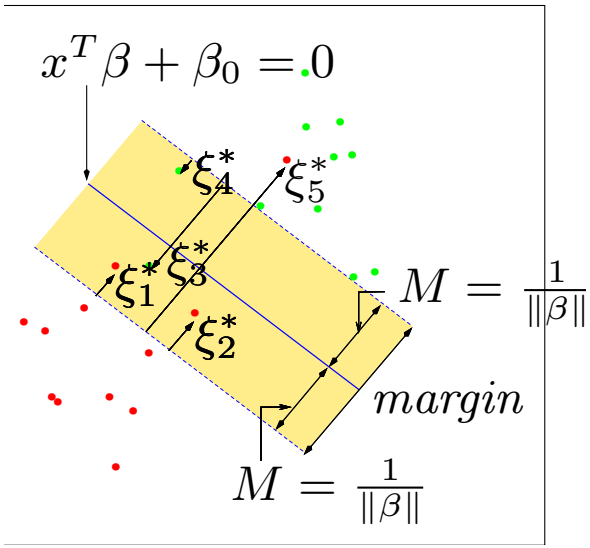
- NB We cannot efficiently compute an ERM solution for 0/1-loss in general, but, for separable data, this gives an efficient way to find one of the ERM solutions!

Support Vector Machines

What if classes are **not** linearly separable?

Support Vector Machines

What if classes are **not** linearly separable?

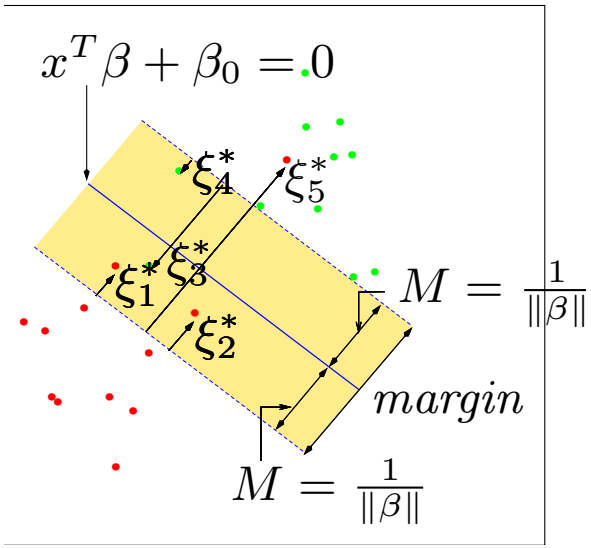


Greek letter "Xi"

Introduce slack variable $\xi_i \geq 0$ for each data point $i = 1, \ldots, N$.

Support Vector Machines

What if classes are **not** linearly separable?



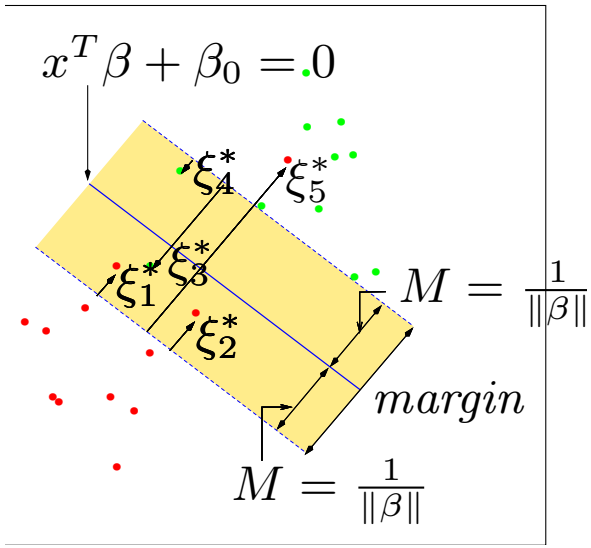
Greek letter "Xi"

Introduce slack variable $\xi_i \geq 0$ for each data point $i = 1, \ldots, N$.

Support vectors: all points **inside** the margin

Support Vector Machines

What if classes are **not** linearly separable?



Greek letter "Xi"

Introduce slack variable $\xi_i \geq 0$ for each data point $i = 1, \ldots, N$.

Support vectors: all points **inside** the margin

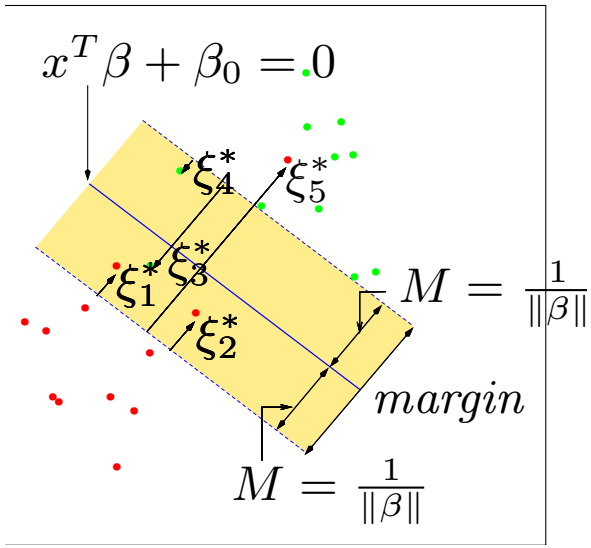
$$\max_{\beta_0, \beta, M, \xi_1, \ldots, \xi_N} \quad M$$

subject to: $\xi_i \geq 0$, $\dfrac{Y_i(X_i^\top \beta + \beta_0)}{\|\beta\|} \geq M(1 - \xi_i)$ for $i = 1, \ldots, N$

$$\sum_{i=1}^{N} \xi_i \leq t$$

Support Vector Machines

What if classes are **not** linearly separable?



Greek letter "Xi"

Introduce slack variable $\xi_i \geq 0$ for each data point $i = 1, \ldots, N$.

Support vectors: all points **inside** the margin

$$\max_{\beta_0, \beta, M, \xi_1, \ldots, \xi_N} \quad M$$

subject to: $\xi_i \geq 0$, $\dfrac{Y_i(X_i^\top \beta + \beta_0)}{\|\beta\|} \geq M(1 - \xi_i)$ for $i = 1, \ldots, N$

$$\sum_{i=1}^{N} \xi_i \leq t$$

t is hyperparameter of the algorithm (assume we take large enough to have a solution)

Support Vector Machines

$$\max_{\beta_0,\beta,M,\xi_1,\ldots,\xi_N} \quad M$$

subject to: $\xi_i \geq 0, \quad \dfrac{Y_i(X_i^\top \beta + \beta_0)}{\|\beta\|} \geq M(1-\xi_i) \qquad$ for $i = 1, \ldots, N$

$$\sum_{i=1}^{N} \xi_i \leq t$$

$$\min_{\beta_0,\beta,\xi_1,\ldots,\xi_N} \quad \frac{1}{2}\|\beta\|^2$$

subject to: $\xi_i \geq 0, \quad Y_i(X_i^\top \beta + \beta_0) \geq 1 - \xi_i \qquad$ for $i = 1, \ldots, N$

$$\sum_{i=1}^{N} \xi_i \leq t$$

Support Vector Machines

$$\max_{\beta_0,\beta,M,\xi_1,\ldots,\xi_N} \quad M$$

subject to: $\xi_i \geq 0, \quad \frac{Y_i(X_i^\top \beta + \beta_0)}{\|\beta\|} \geq M(1-\xi_i) \qquad$ for $i = 1, \ldots, N$

$$\sum_{i=1}^{N} \xi_i \leq t$$

$$\min_{\beta_0,\beta,\xi_1,\ldots,\xi_N} \quad \frac{1}{2}\|\beta\|^2$$

subject to: $\xi_i \geq 0, \quad Y_i(X_i^\top \beta + \beta_0) \geq 1 - \xi_i \qquad$ for $i = 1, \ldots, N$

$$\sum_{i=1}^{N} \xi_i \leq t$$

equivalent: varying t above gives same set of solutions as varying C below

---

$$\min_{\beta_0,\beta,\xi_1,\ldots,\xi_N} \quad \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N} \xi_i$$

subject to: $\xi_i \geq 0, \quad Y_i(X_i^\top \beta + \beta_0) \geq 1 - \xi_i \qquad$ for $i = 1, \ldots, N$

Support Vector Machines

$$\max_{\beta_0,\beta,M,\xi_1,\ldots,\xi_N} \quad M$$

subject to: $\xi_i \geq 0, \quad \frac{Y_i(X_i^\top \beta + \beta_0)}{\|\beta\|} \geq M(1-\xi_i) \qquad$ for $i = 1, \ldots, N$

$$\sum_{i=1}^{N} \xi_i \leq t$$

$$\min_{\beta_0,\beta,\xi_1,\ldots,\xi_N} \quad \frac{1}{2}\|\beta\|^2$$

subject to: $\xi_i \geq 0, \quad Y_i(X_i^\top \beta + \beta_0) \geq 1 - \xi_i \qquad$ for $i = 1, \ldots, N$

$$\sum_{i=1}^{N} \xi_i \leq t$$

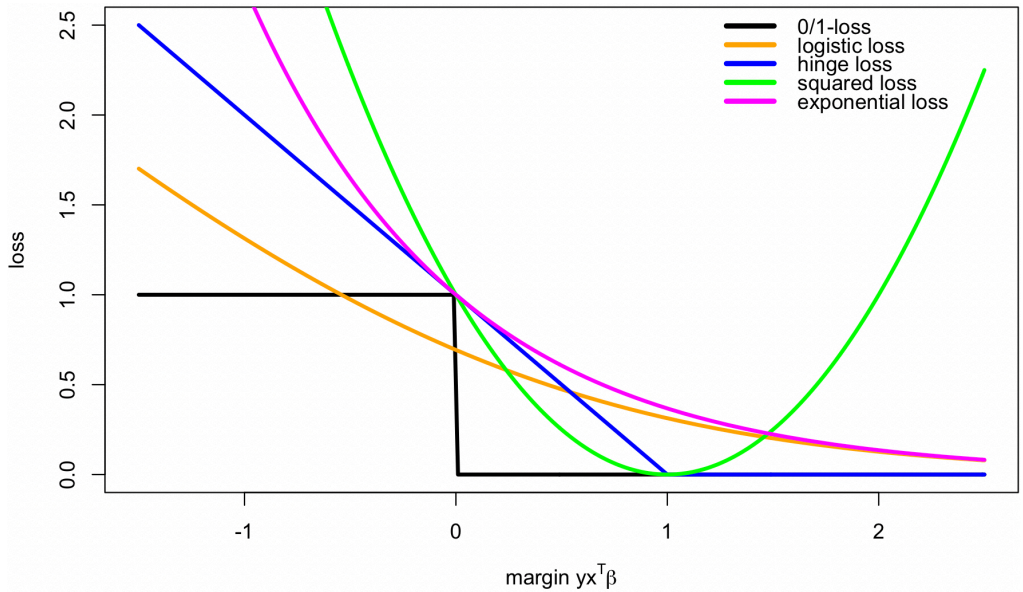
equivalent: varying t above gives same set of solutions as varying C below

---

$$\min_{\beta_0,\beta,\xi_1,\ldots,\xi_N} \quad \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N} \xi_i$$

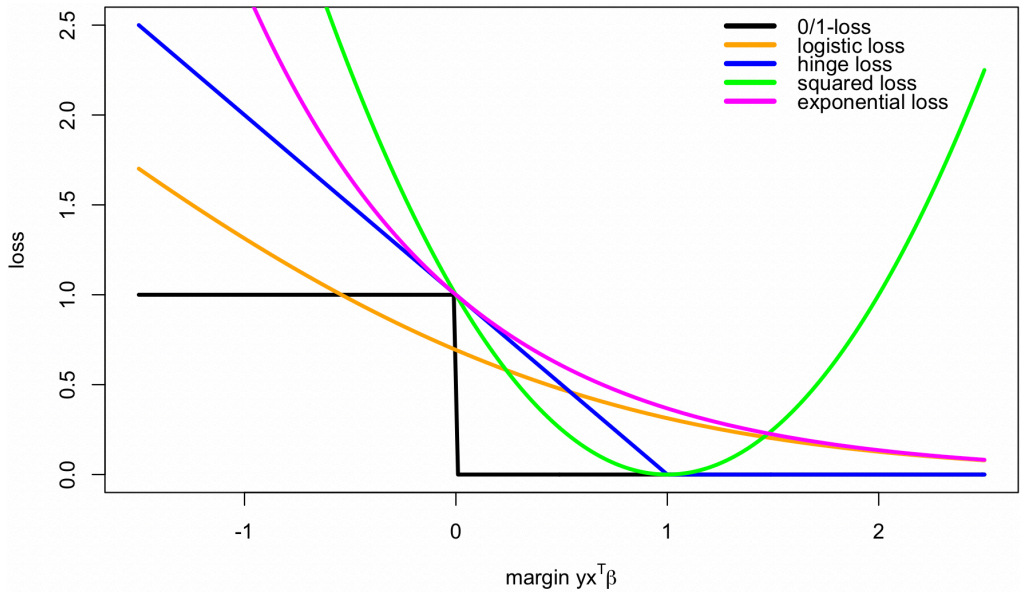
subject to: $\xi_i \geq 0, \quad Y_i(X_i^\top \beta + \beta_0) \geq 1 - \xi_i \qquad$ for $i = 1, \ldots, N$

Interpretation as Penalized ERM with Hinge Loss



Hinge loss: $L(Y_i, f(X_i)) = \max\{0, 1 - Y_i f(X_i)\}$

Interpretation as Penalized ERM with Hinge Loss

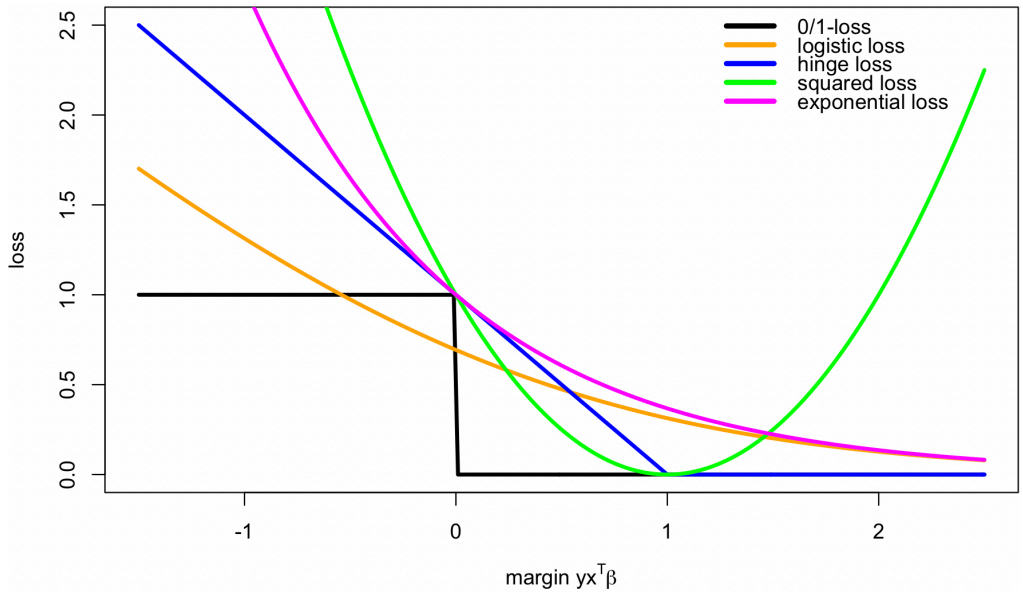


Hinge loss: $L(Y_i, f(X_i)) = \max\{0, 1 - Y_i f(X_i)\}$

$$\min_{\beta_0,\beta,\xi_1,\ldots,\xi_N} \quad \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i$$

subject to: $\xi_i \geq 0, \quad Y_i(X_i^\top \beta + \beta_0) \geq 1 - \xi_i \qquad$ for $i = 1, \ldots, N$

Interpretation as Penalized ERM with Hinge Loss



Hinge loss: $L(Y_i, f(X_i)) = \max\{0, 1 - Y_i f(X_i)\}$

$$\min_{\beta_0,\beta,\xi_1,\ldots,\xi_N} \quad \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i$$

subject to: $\underbrace{\xi_i \geq 0, \quad Y_i(X_i^\top \beta + \beta_0) \geq 1 - \xi_i}_{\xi_i \geq L(Y_i, X_i^\top \beta + \beta_0)}$ for $i = 1, \ldots, N$

Interpretation as Penalized ERM with Hinge Loss

$$\min_{\beta_0,\beta,\xi_1,\dots,\xi_N} \quad \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i$$

$$\text{subject to:} \quad \underbrace{\xi_i \geq 0, \quad Y_i(X_i^\top\beta + \beta_0) \geq 1 - \xi_i}_{\xi_i \geq L(Y_i, X_i^\top\beta + \beta_0)} \qquad \text{for } i = 1, \dots, N$$

$$\min_{\beta_0,\beta} \quad \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N} L(Y_i, X_i^\top\beta + \beta_0)$$

Interpretation as Penalized ERM with Hinge Loss

$$\min_{\beta_0,\beta,\xi_1,\dots,\xi_N} \quad \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i$$

subject to: $\underbrace{\xi_i \geq 0, \quad Y_i(X_i^\top\beta + \beta_0) \geq 1 - \xi_i}_{\xi_i \geq L(Y_i, X_i^\top\beta + \beta_0)}$ for $i = 1, \dots, N$

$$\min_{\beta_0,\beta} \quad \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N} L(Y_i, X_i^\top\beta + \beta_0)$$

$$\min_{\beta_0,\beta} \quad \sum_{i=1}^{N} L(Y_i, X_i^\top\beta + \beta_0) + \frac{1}{2C}\|\beta\|^2$$

ERM for hinge loss with ℓ_2-penalty, $\lambda = \frac{1}{2C}$.

Dual Formulation

$$\min_{\alpha_1,\ldots,\alpha_N} \quad \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} \alpha_i \alpha_k Y_i Y_k \langle X_i, X_k \rangle$$

$X_i^\top X_k$

subject to: $\quad 0 \leq \alpha_i \leq C \qquad$ for $i = 1, \ldots, N$

$$\sum_{i=1}^{N} \alpha_i Y_i = 0$$

[1] Proof in next lecture

Dual Formulation

$X_i^\top X_k$

$$\min_{\alpha_1,\ldots,\alpha_N} \quad \sum_{i=1}^N \alpha_i - \frac{1}{2}\sum_{i=1}^N\sum_{k=1}^N \alpha_i\alpha_k Y_i Y_k \langle X_i, X_k\rangle$$

subject to: $\quad 0 \le \alpha_i \le C \qquad$ for $i = 1, \ldots, N$

$$\sum_{i=1}^N \alpha_i Y_i = 0$$

Then solution to ★ is:[1]

$$\hat{\beta} = \sum_{i=1}^N \hat{\alpha}_i Y_i X_i$$

$$\hat{\beta}_0 = Y_i - X_i^\top \hat{\beta} \qquad \text{for any } i \text{ s.t. } 0 < \hat{\alpha}_i < C.$$

[1] Proof in next lecture

Dual Formulation

$X_i^\top X_k$

$$\min_{\alpha_1,\ldots,\alpha_N} \quad \sum_{i=1}^N \alpha_i - \frac{1}{2}\sum_{i=1}^N\sum_{k=1}^N \alpha_i\alpha_k Y_i Y_k \langle X_i, X_k\rangle$$

subject to: $\quad 0 \le \alpha_i \le C \qquad$ for $i = 1, \ldots, N$

$$\sum_{i=1}^N \alpha_i Y_i = 0$$

Then solution to ★ is:[1]

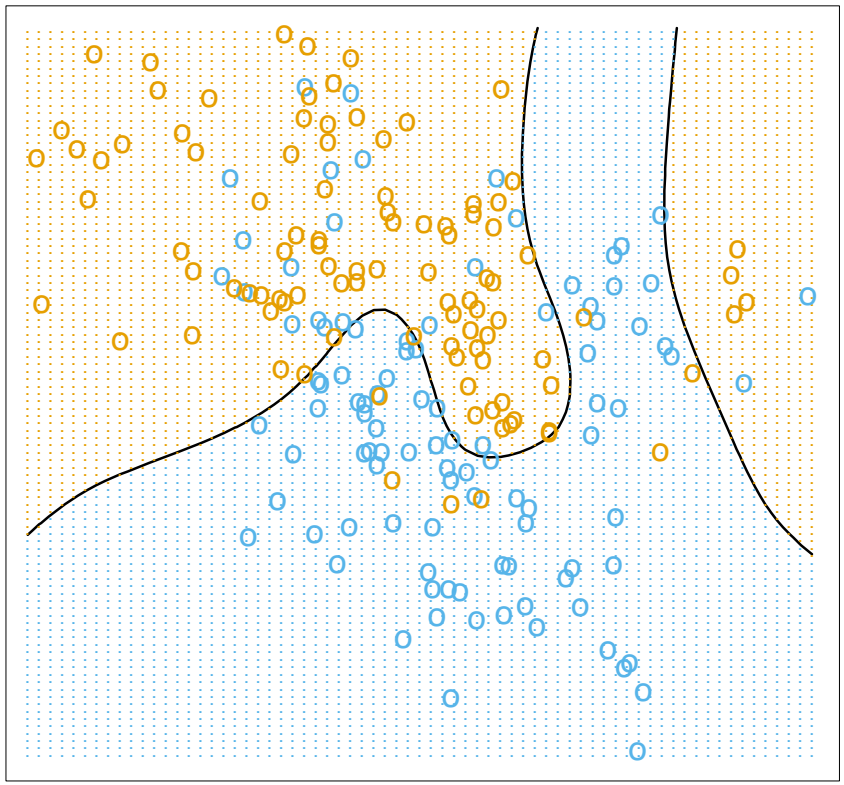
$$\hat{\beta} = \sum_{i=1}^N \hat{\alpha}_i Y_i X_i$$

$$\hat{\beta}_0 = Y_i - X_i^\top \hat{\beta} \qquad \text{for any } i \text{ s.t. } 0 < \hat{\alpha}_i < C.$$

$\hat{\alpha}_i = 0$	for X_i outside margin and correctly classified
$0 < \hat{\alpha}_i < C$	for X_i exactly on margin and correctly classified
$\hat{\alpha}_i = C$	for X_i strictly inside margin or misclassified

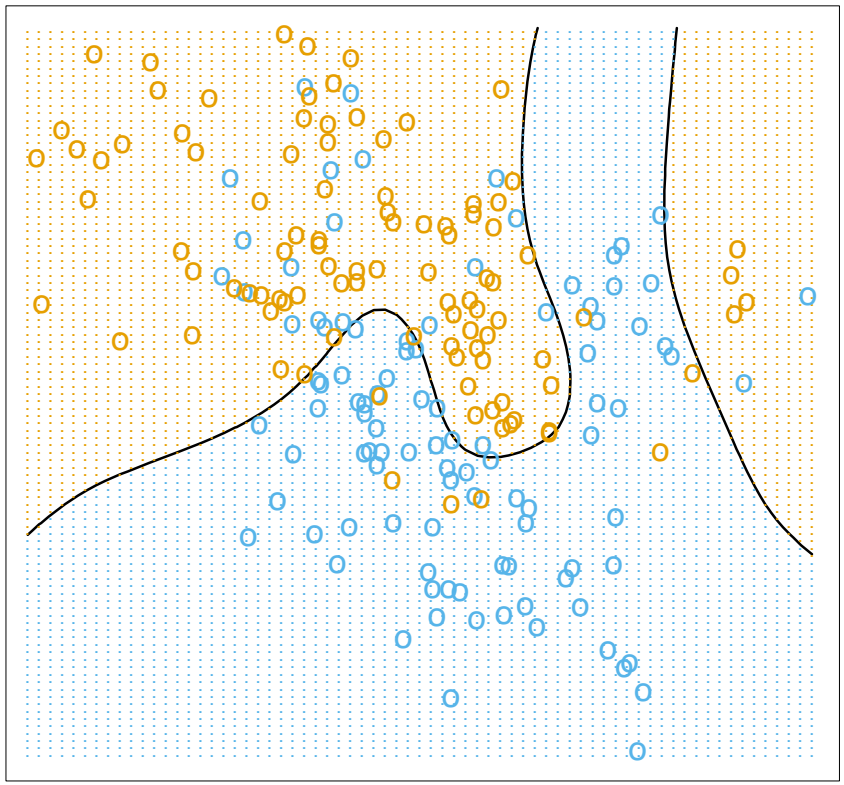
[1] Proof in next lecture

Kernel Trick I



How to learn something like this with **linear classifier?**

Kernel Trick I

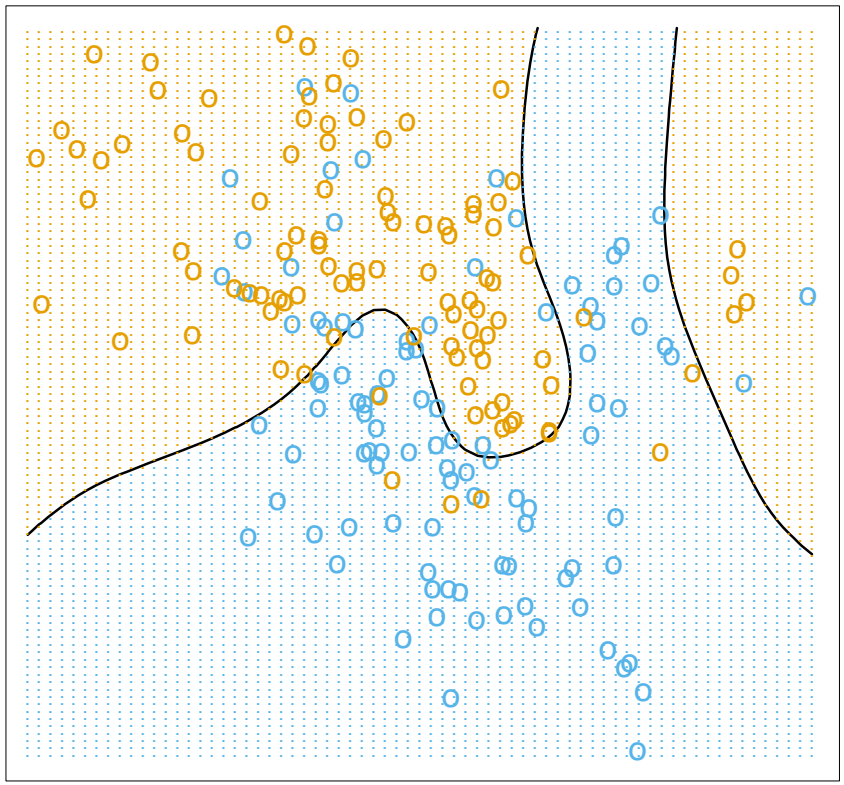


How to learn something like this with **linear classifier?**

Map features X to a larger set of features $h(X)$! For instance,

$$h\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \\ X_1 X_2 \\ X_1^2 \\ X_2^2 \end{pmatrix}$$

Kernel Trick I



How to learn something like this with **linear classifier?**

Map features X to a larger set of features $h(X)$! For instance,

$$h\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \\ X_1 X_2 \\ X_1^2 \\ X_2^2 \end{pmatrix}$$

Then $\langle X_i, X_k \rangle$ in dual formulation becomes $\langle h(X_i), h(X_k) \rangle$.

Kernel Trick II

$$h\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \\ X_1 X_2 \\ X_1^2 \\ X_2^2 \end{pmatrix}$$

Then $\langle X_i, X_k \rangle$ in dual formulation becomes $\langle h(X_i), h(X_k) \rangle$.

Kernel Trick II

$$h\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \\ X_1X_2 \\ X_1^2 \\ X_2^2 \end{pmatrix}$$

Then $\langle X_i, X_k \rangle$ in dual formulation becomes $\langle h(X_i), h(X_k) \rangle$.

Kernel Trick: Do not need to specify h, only need to know the **kernel function**

$$K(X_i, X_k) = \langle h(X_i), h(X_k) \rangle \tag{1}$$

$h(X)$ may even be infinite-dimensional!

Kernel Trick II

$$h\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \\ X_1X_2 \\ X_1^2 \\ X_2^2 \end{pmatrix}$$

Then $\langle X_i, X_k \rangle$ in dual formulation becomes $\langle h(X_i), h(X_k) \rangle$.

Kernel Trick: Do not need to specify h, only need to know the **kernel function**

$$K(X_i, X_k) = \langle h(X_i), h(X_k) \rangle \tag{1}$$

$h(X)$ may even be infinite-dimensional!

- Measure of similarity: larger if X_i and X_k more similar

Kernel Trick II

$$h\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \\ X_1 X_2 \\ X_1^2 \\ X_2^2 \end{pmatrix}$$

Then $\langle X_i, X_k \rangle$ in dual formulation becomes $\langle h(X_i), h(X_k) \rangle$.

Kernel Trick: Do not need to specify h, only need to know the **kernel function**

$$K(X_i, X_k) = \langle h(X_i), h(X_k) \rangle \tag{1}$$

$h(X)$ may even be infinite-dimensional!

- Measure of similarity: larger if X_i and X_k more similar
- Would be computationally nice if this were a simple function...
- ... so turn things around and **start by defining** $K(X_i, X_k)$; then find h to satisfy (1)!

Kernel Function

Examples:

- Degree d polynomial: $K(X, X') = \left(1 + \langle X, X' \rangle\right)^d$
  - E.g. the case of $d = 2$ and $X \in \mathbb{R}^2$ corresponds to:

$$h(X) = \begin{pmatrix} 1 \\ \sqrt{2}X_1 \\ \sqrt{2}X_2 \\ X_1^2 \\ X_2^2 \\ \sqrt{2}X_1X_2 \end{pmatrix}$$

Kernel Function

Examples:

- Degree d polynomial: $K(X, X') = \left(1 + \langle X, X' \rangle\right)^d$
- Radial basis: $K(X, X') = e^{-\gamma \|X - X'\|^2}$ for $\gamma > 0$
- Neural network: $K(X, X') = \tanh\left(a\langle X, X' \rangle + b\right)$ for $a > 0, b \in \mathbb{R}$, $\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$

Kernel Function

Examples:

- Degree d polynomial: $K(X, X') = \left(1 + \langle X, X' \rangle\right)^d$
- Radial basis: $K(X, X') = e^{-\gamma \|X - X'\|^2}$ for $\gamma > 0$
- Neural network: $K(X, X') = \tanh\left(a\langle X, X' \rangle + b\right)$ for $a > 0, b \in \mathbb{R}$, $\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$

If K satisfies certain **technical conditions** (symmetric, positive definite), then there always exists some mapping h s.t.

$$K(X_i, X_k) = \langle h(X_i), h(X_k) \rangle.$$

Kernel Function

Examples:

- Degree d polynomial: $K(X, X') = \left(1 + \langle X, X' \rangle\right)^d$
- Radial basis: $K(X, X') = e^{-\gamma \|X - X'\|^2}$ for $\gamma > 0$
- Neural network: $K(X, X') = \tanh\left(a\langle X, X' \rangle + b\right)$ for $a > 0, b \in \mathbb{R}$, $\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$

If K satisfies certain **technical conditions** (symmetric, positive definite), then there always exists some mapping h s.t.

$$K(X_i, X_k) = \langle h(X_i), h(X_k) \rangle.$$

- In general this requires h to map to "infinite-dimensional vectors"
- Details in the homework, where we will represent such vectors as elements of a so-called **reproducing kernel Hilbert space**

Kernel Trick: Classifying a New X

- If h maps X to an "infinite-dimensional vector", then β also becomes infinite-dimensional.
- How do we then compute $f_{\hat{\beta}_0,\hat{\beta}}(X)$ for a new input X?

Kernel Trick: Classifying a New X

- If h maps X to an "infinite-dimensional vector", then β also becomes infinite-dimensional.
- How do we then compute $f_{\hat{\beta}_0,\hat{\beta}}(X)$ for a new input X?

Classifying a new X:

$$\begin{aligned} f_{\hat{\beta}_0,\hat{\beta}}(X) &= \langle h(X), \hat{\beta} \rangle + \hat{\beta}_0 \\ &= \langle h(X), \sum_{i=1}^{N} \hat{\alpha}_i Y_i h(X_i) \rangle + \hat{\beta}_0 \\ &= \sum_{i=1}^{N} \hat{\alpha}_i Y_i \langle h(X), h(X_i) \rangle + \hat{\beta}_0 \\ &= \sum_{i=1}^{N} \hat{\alpha}_i Y_i K(X, X_i) + \hat{\beta}_0 \end{aligned}$$

Again no need to compute h, only need to know kernel function!

Support Vector Machines in Action

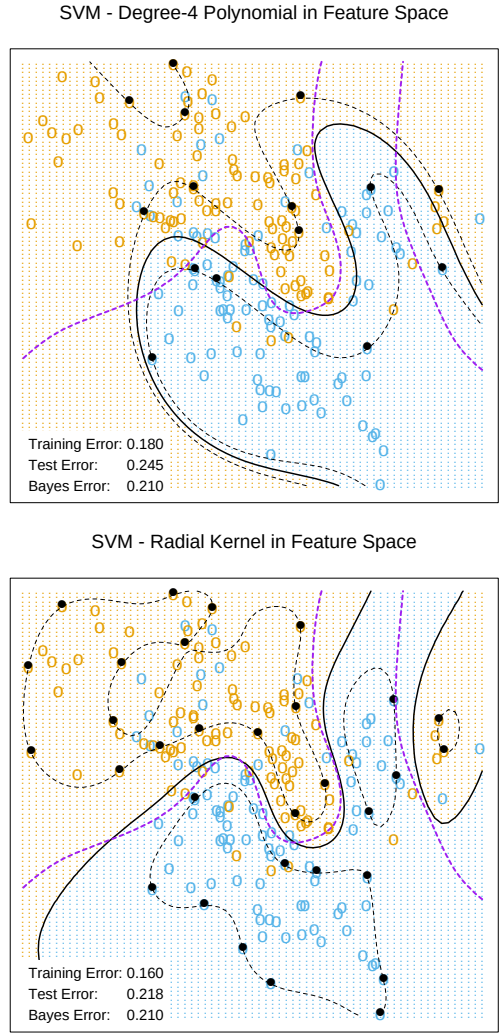


FIGURE 12.3. *Two nonlinear SVMs for the mixture data. The upper plot uses a 4th degree polynomial*

To be continued next week!

- I still owe you a derivation for the relation between $\star$ and the dual formulation
- This will require additional background on convex optimization