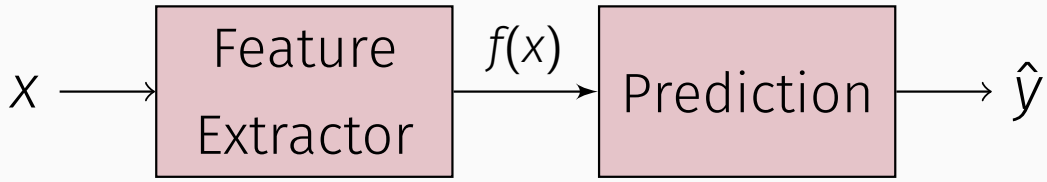
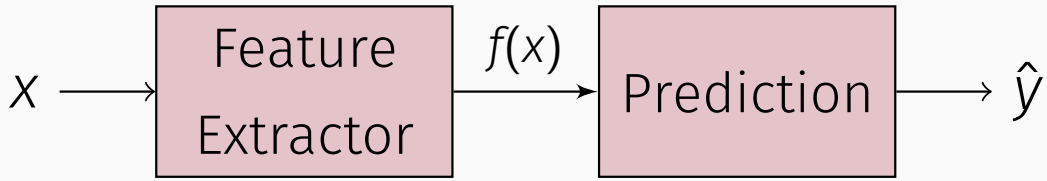
Kernel Subspace and Feature Extraction

Xiangxiang Xu
Joint work with Prof. Lizhong Zheng

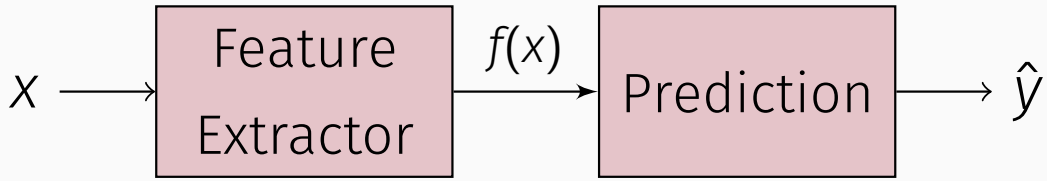
ISIT 2023

Machine Learning and Kernel Methods



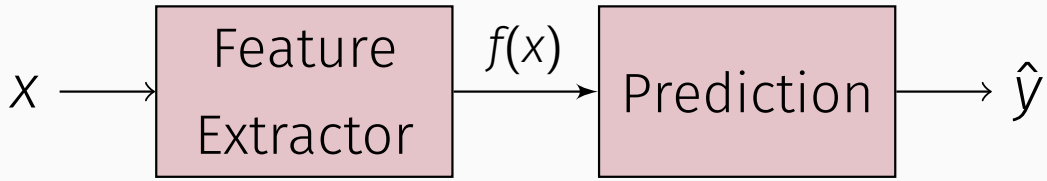


Features: Explicit & Implicit



Features: Explicit & Implicit

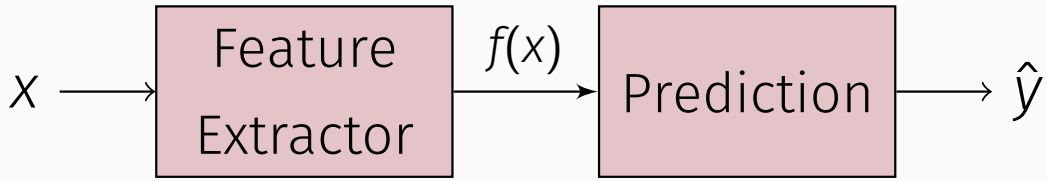
- Parameterized features, e.g., deep neural networks



Features: Explicit & Implicit

- Parameterized features, e.g., deep neural networks
- Kernel methods: features defined by kernel

Machine Learning and Kernel Methods

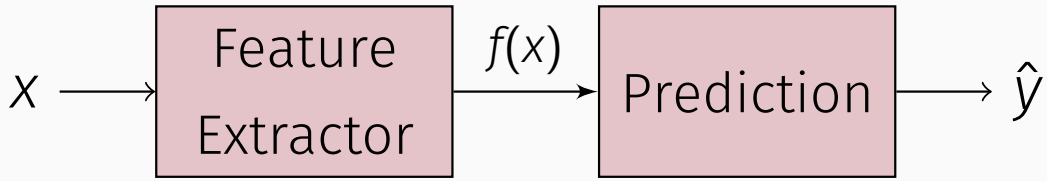


Features: Explicit & Implicit

- Parameterized features, e.g., deep neural networks
- Kernel methods: features defined by kernel

Understand Kernel

Machine Learning and Kernel Methods



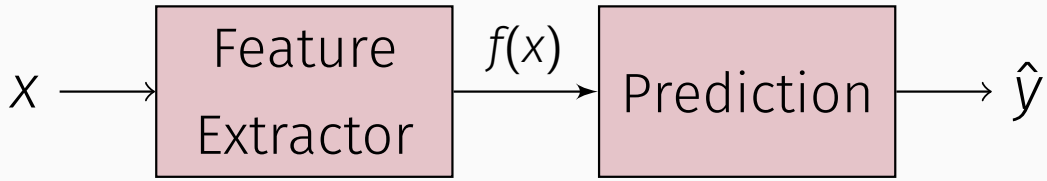
Features: Explicit & Implicit

- Parameterized features, e.g., deep neural networks
- Kernel methods: features defined by kernel

Understand Kernel

- How it worked?

Machine Learning and Kernel Methods



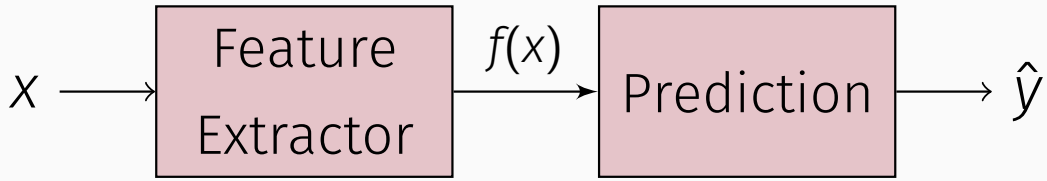
Features: Explicit & Implicit

- Parameterized features, e.g., deep neural networks
- Kernel methods: features defined by kernel

Understand Kernel

- How it worked?
- What makes a kernel good?

Machine Learning and Kernel Methods



Features: Explicit & Implicit

- Parameterized features, e.g., deep neural networks
- Kernel methods: features defined by kernel

Understand Kernel

- How it worked?
- What makes a kernel good?
- How to obtain a good kernel?

Binary Classification: Statistical View

Data X, Label $Y = \pm 1$, balanced

Binary Classification: Statistical View

Data X, Label $Y = \pm 1$, balanced

MAP $\equiv$ ML: $$P_{X|Y=1}(x) \underset{\hat{y}=-1}{\overset{\hat{y}=1}{\gtrless}} P_{X|Y=-1}(x)$$

Binary Classification: Statistical View

Data X, Label $Y = \pm 1$, balanced

MAP $\equiv$ ML: $$P_{X|Y=1}(x) \underset{\hat{y}=-1}{\overset{\hat{y}=1}{\gtrless}} P_{X|Y=-1}(x)$$

A Feature Perspective

Consider $$f^*(x) \propto \frac{P_{X|Y=1}(x) - P_{X|Y=-1}(x)}{P_X(x)}$$

- $\mathbb{E}[f^*(X)] = 0$, $\mathsf{var}(f^*(X)) = 1$ (pick the scaling factor)

Binary Classification: Statistical View

Data X, Label $Y = \pm 1$, balanced

MAP $\equiv$ ML: $$P_{X|Y=1}(x) \underset{\hat{y}=-1}{\overset{\hat{y}=1}{\gtrless}} P_{X|Y=-1}(x)$$

A Feature Perspective

Consider $$f^*(x) \propto \frac{P_{X|Y=1}(x) - P_{X|Y=-1}(x)}{P_X(x)}$$

- $\mathbb{E}[f^*(X)] = 0$, $\mathsf{var}(f^*(X)) = 1$ (pick the scaling factor)
- maximally correlated choice

Binary Classification: Statistical View

Data X, Label $Y = \pm 1$, balanced

MAP $\equiv$ ML: $$P_{X|Y=1}(x) \underset{\hat{y}=-1}{\overset{\hat{y}=1}{\gtrless}} P_{X|Y=-1}(x)$$

A Feature Perspective

Consider $$f^*(x) \propto \frac{P_{X|Y=1}(x) - P_{X|Y=-1}(x)}{P_X(x)}$$

- $\mathbb{E}[f^*(X)] = 0$, $\mathsf{var}(f^*(X)) = 1$ (pick the scaling factor)
- maximally correlated choice
- $\dfrac{P_{Y|X}(y|x)}{P_Y(y)} = 1 + \varrho \cdot y \cdot f^*(x)$ ϱ: maximal correlation

Binary Classification: Statistical View

Data X, Label $Y = \pm 1$, balanced

MAP $\equiv$ ML: $$P_{X|Y=1}(x) \underset{\hat{y}=-1}{\overset{\hat{y}=1}{\gtrless}} P_{X|Y=-1}(x)$$

A Feature Perspective

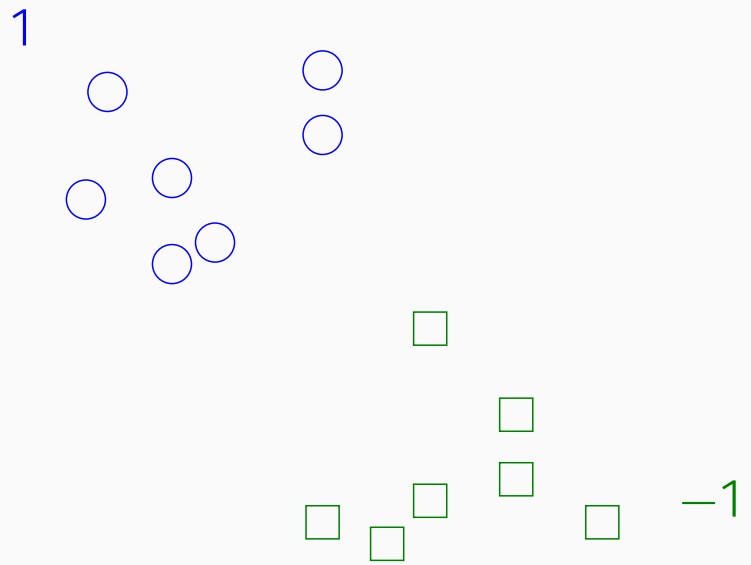
Consider $$f^*(x) \propto \frac{P_{X|Y=1}(x) - P_{X|Y=-1}(x)}{P_X(x)}$$

- $\mathbb{E}[f^*(X)] = 0$, $\mathsf{var}(f^*(X)) = 1$ (pick the scaling factor)
- maximally correlated choice
- $\dfrac{P_{Y|X}(y|x)}{P_Y(y)} = 1 + \varrho \cdot y \cdot f^*(x)$ ϱ: maximal correlation

Optimal Decision

$$\hat{y}_{\mathrm{MAP}}(x) = \mathsf{sgn}(f^*(x))$$ f^*: sufficient statistic

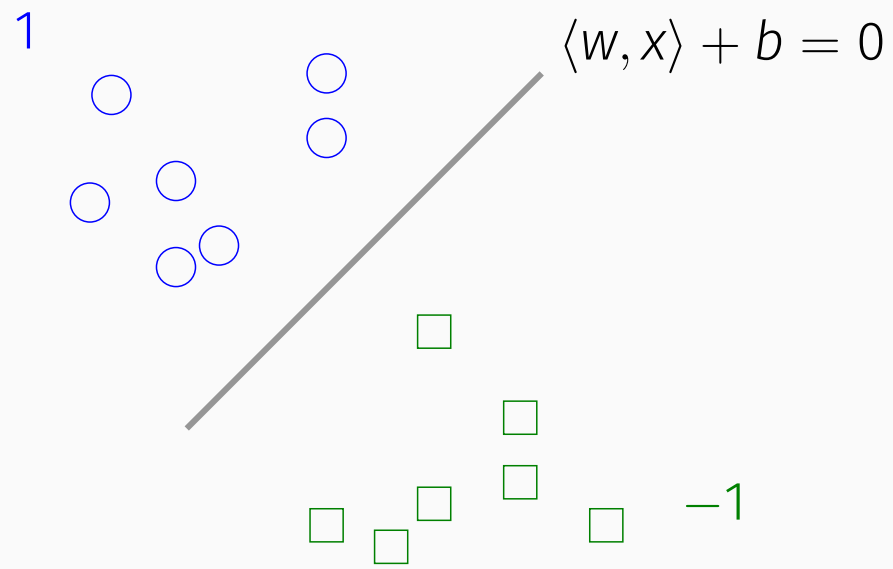
Support Vector Machine



Support Vector Machine

Separating Hyperplane

$$\langle w, x \rangle + b = 0$$

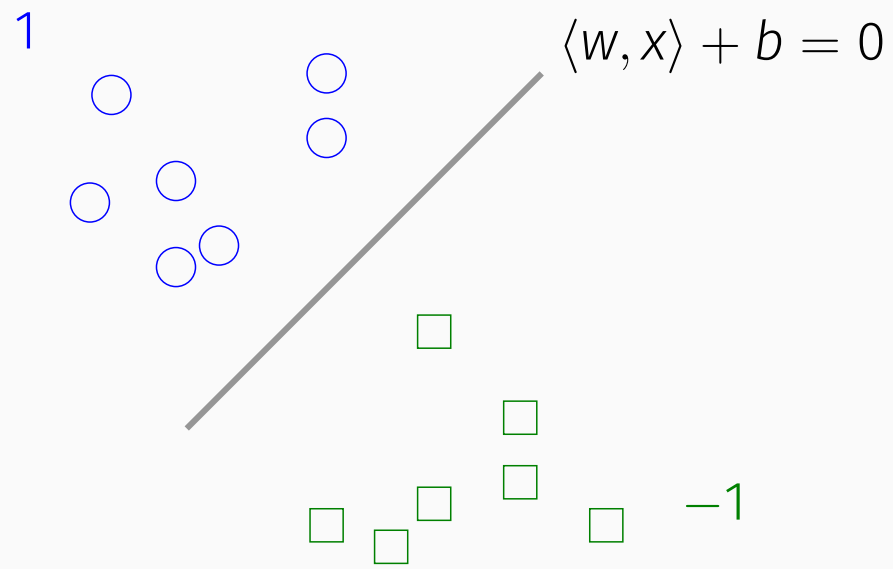


Support Vector Machine

Separating Hyperplane

$$\langle w, x \rangle + b = 0$$

Decide which side x lies in:



Support Vector Machine

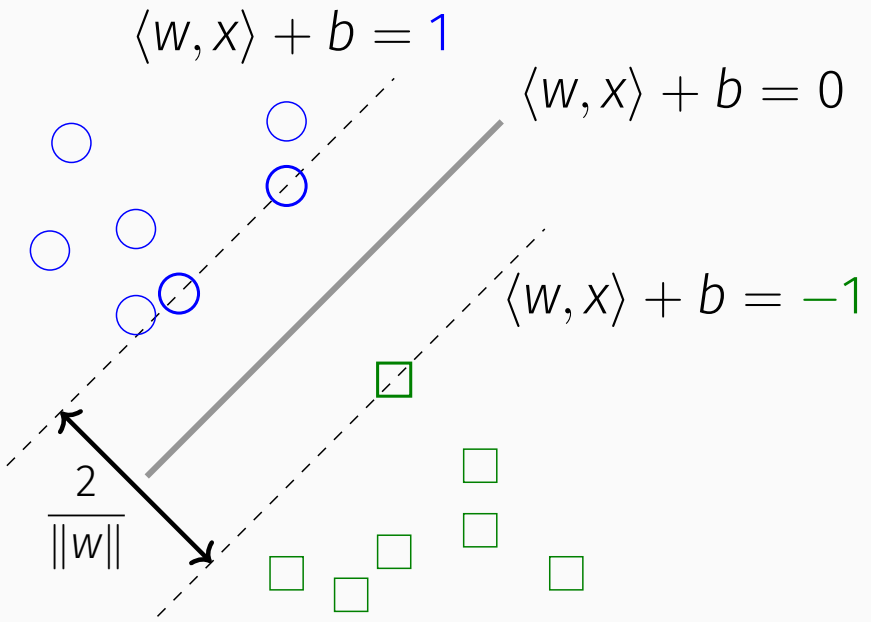
Separating Hyperplane

$$\langle w, x \rangle + b = 0$$

Decide which side x lies in:

$$\hat{y}_{\text{SVM}}(x) \triangleq \text{sgn}(\langle w^*, x \rangle + b^*)$$

w^*, b^*: maximize margin



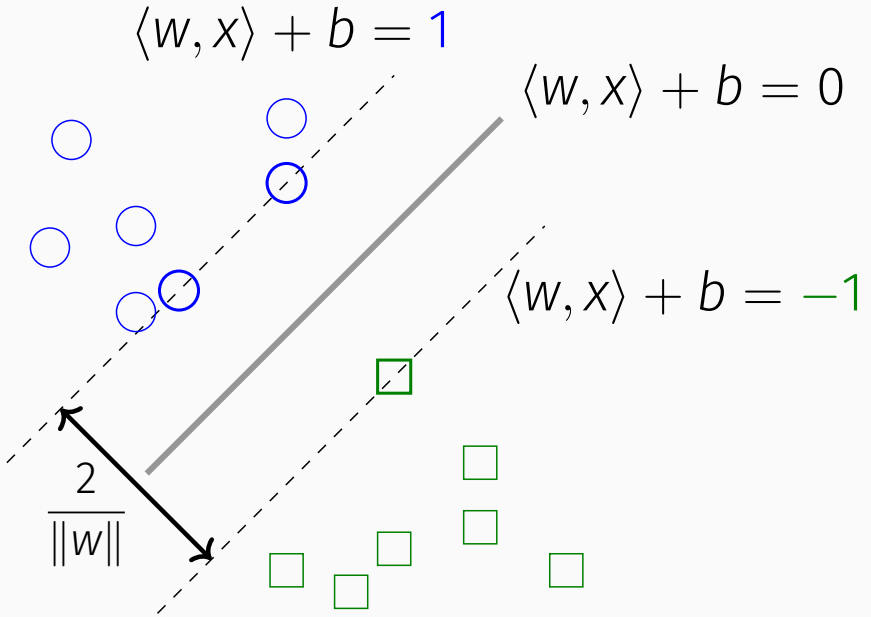
Support Vector Machine

Separating Hyperplane

$$\langle w, x \rangle + b = 0$$

Decide which side x lies in:

$$\hat{y}_{\text{SVM}}(x) \triangleq \text{sgn}(\langle w^*, x \rangle + b^*)$$



w^*, b^*: maximize margin

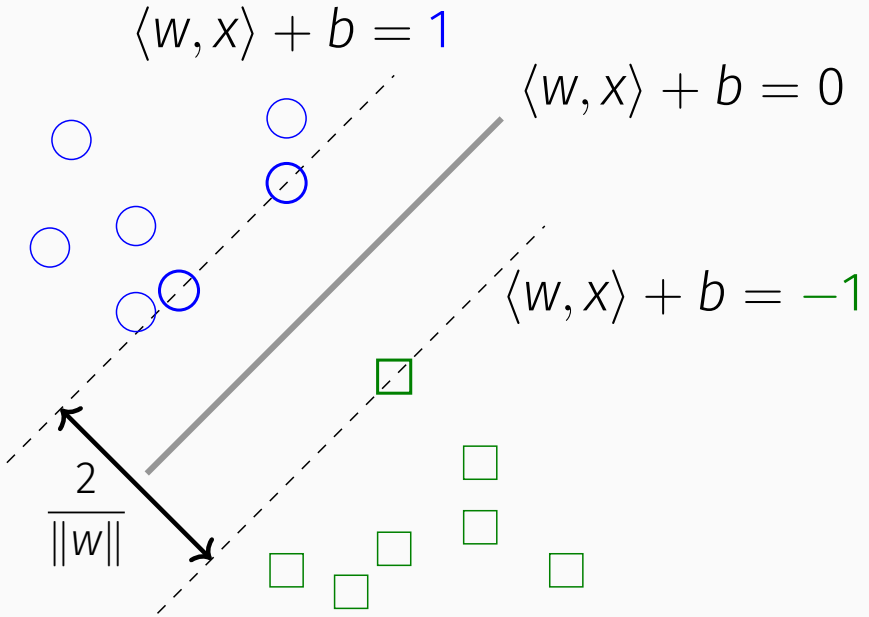
$$\min_{w,b} \quad \frac{1}{2} \cdot \|w\|^2$$

$$\text{s.t.} \quad 1 - y_i \cdot (\langle w, x_i \rangle + b) \leq 0$$

Support Vector Machine

Separating Hyperplane

$$\langle w, x \rangle + b = 0$$



Decide which side x lies in:

$$\hat{y}_{\text{SVM}}(x) \triangleq \text{sgn}(\langle w^*, x \rangle + b^*)$$

w^*, b^*: maximize margin

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

$$\text{s. t.} \quad 1 - y_i \cdot (\langle w, x_i \rangle + b) \leq 0$$

Soft Margin

$$L_{\text{SVM}}(w, b; \lambda) \triangleq \mathbb{E}\left[(1 - Y \cdot (\langle w, X \rangle + b))^+\right] + \frac{\lambda}{2} \cdot \|w\|^2$$

Implementation

- solve the dual optimization problem
- expressed in terms of $\langle x_i, x_j \rangle$

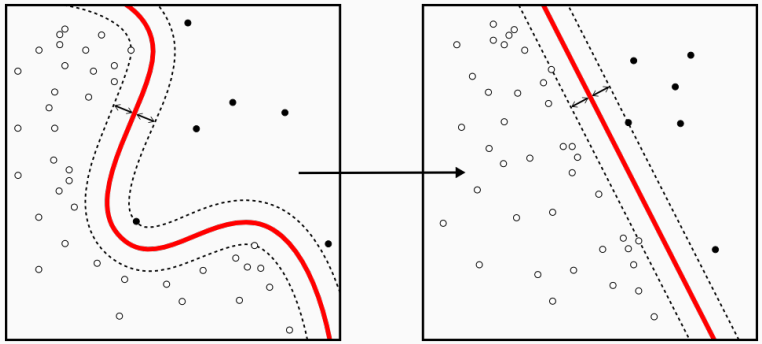
Support Vector Machine

Implementation

- solve the dual optimization problem
- expressed in terms of $\langle x_i, x_j \rangle$

Introducing Features

- use $f(X)$ instead of X
- linear separability
- "flexibility"

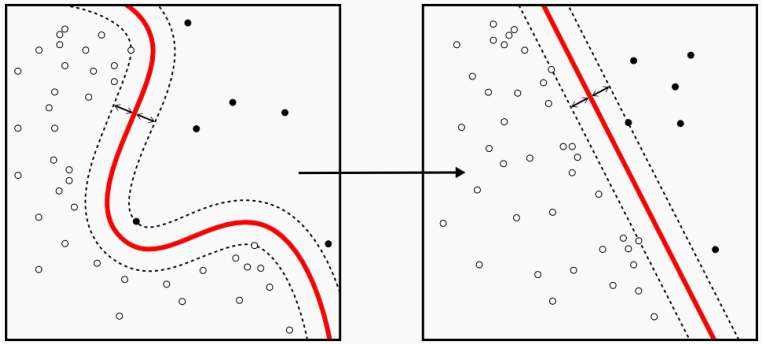


Implementation

- solve the dual optimization problem
- expressed in terms of $\langle x_i, x_j \rangle$

Introducing Features

- use $f(X)$ instead of X
- linear separability
- "flexibility"



Kernel $k(x, x') = \langle f(x), f(x') \rangle$

- resulting f: implicit and high-dimensional
- map data without extra computational cost

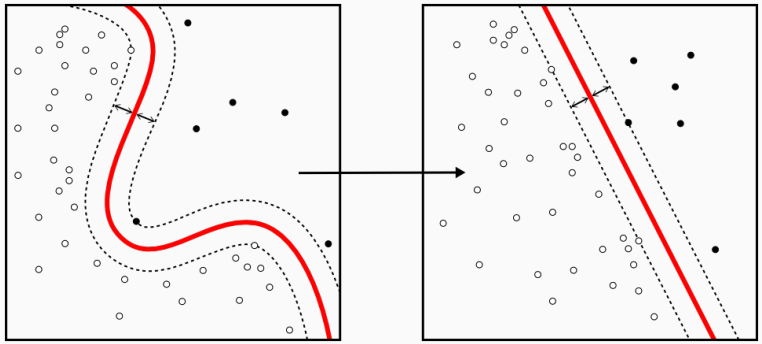
Support Vector Machine

Implementation

- solve the dual optimization problem
- expressed in terms of $\langle x_i, x_j \rangle$

Introducing Features

- use $f(X)$ instead of X
- linear separability
- "flexibility"



Kernel $\mathcal{K}(x, x') = \langle f(x), f(x') \rangle$

- resulting f: implicit and high-dimensional
- map data without extra computational cost
- choices of $\mathcal{K}$: RBF, polynomial

SVM for Given f

"Soft" Regime: large λ

$w^* = \lambda^{-1} \cdot \mathbb{E}[f(X)Y], \; b^* = 0$ — assume $\mathbb{E}[f] = 0, \mathbf{cov}(f) = I$

SVM for Given f

"Soft" Regime: large λ

$$w^* = \lambda^{-1} \cdot \mathbb{E}\left[f(X)Y\right], \; b^* = 0 \qquad - \text{assume } \mathbb{E}\left[f\right] = 0, \mathsf{cov}(f) = I$$

$$\hat{y}_{\mathsf{SVM}}(x; f) = \mathsf{sgn}\left(\langle w^*, f(x)\rangle + b^*\right) = \mathsf{sgn}\left(\left\langle \mathbb{E}\left[f(X)Y\right], f(x)\right\rangle\right)$$

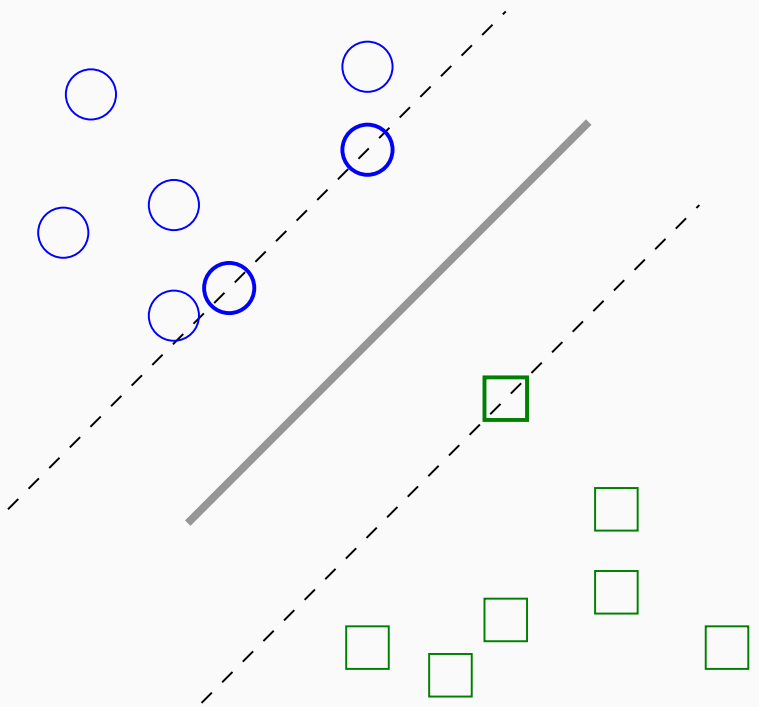
SVM for Given f

"Soft" Regime: large λ

$$w^* = \lambda^{-1} \cdot \mathbb{E}\left[f(X)Y\right], \; b^* = 0 \qquad - \text{assume } \mathbb{E}\left[f\right] = 0, \mathsf{cov}(f) = I$$

$$\hat{y}_{\mathsf{SVM}}(x; f) = \mathsf{sgn}\left(\langle w^*, f(x)\rangle + b^*\right) = \mathsf{sgn}\left(\left\langle \mathbb{E}\left[f(X)Y\right], f(x)\right\rangle\right)$$

$$\boxed{\mathbb{E}\left[f(X)Y\right] \propto \mathbb{E}\left[f(X)|Y=1\right] - \mathbb{E}\left[f(X)|Y=-1\right]}$$



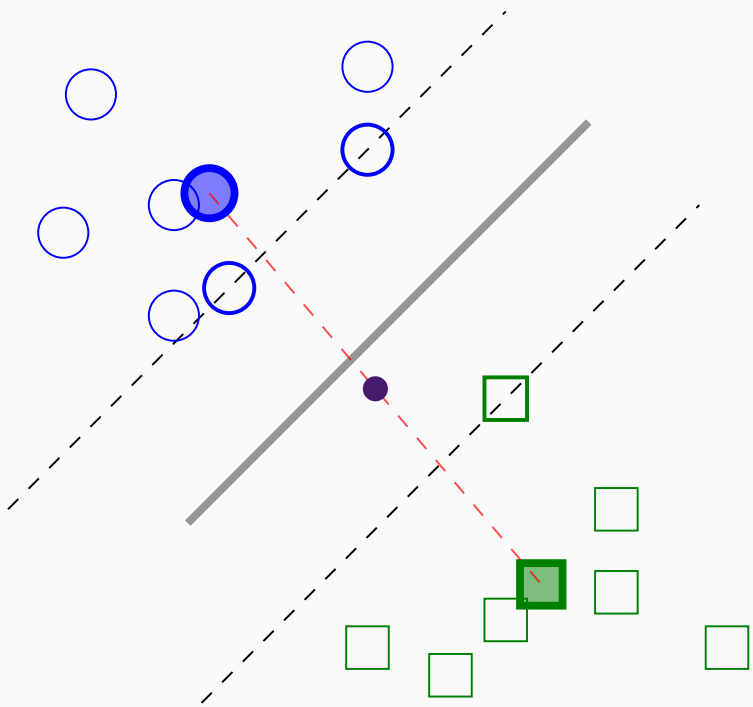
SVM for Given f

"Soft" Regime: large λ

$$w^* = \lambda^{-1} \cdot \mathbb{E}\left[f(X)Y\right], \; b^* = 0 \qquad - \text{assume } \mathbb{E}\left[f\right] = 0, \mathsf{cov}(f) = I$$

$$\hat{y}_{\mathsf{SVM}}(x; f) = \mathsf{sgn}\left(\langle w^*, f(x)\rangle + b^*\right) = \mathsf{sgn}\left(\left\langle \mathbb{E}\left[f(X)Y\right], f(x)\right\rangle\right)$$

$$\boxed{\mathbb{E}\left[f(X)Y\right] \propto \mathbb{E}\left[f(X)|Y=1\right] - \mathbb{E}\left[f(X)|Y=-1\right]}$$



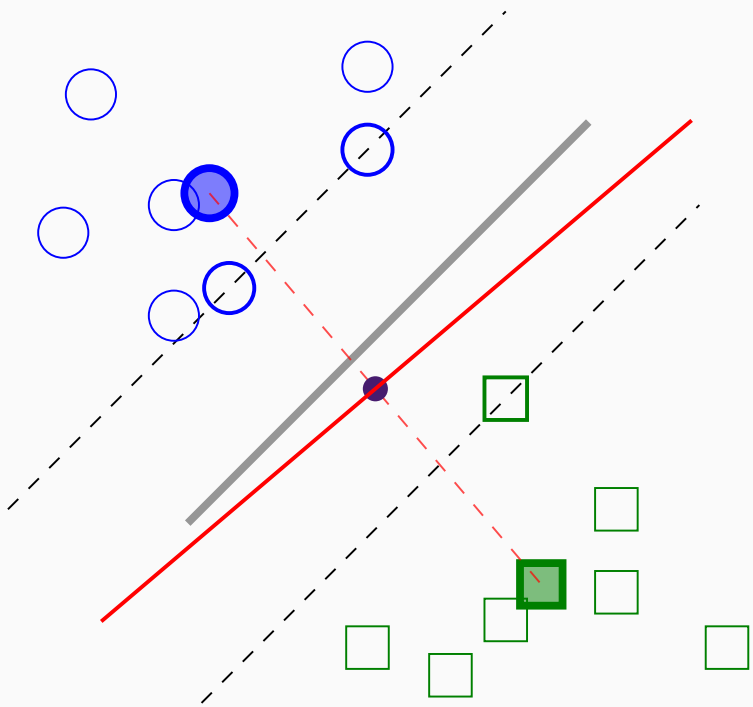
SVM for Given f

"Soft" Regime: large λ

$$w^* = \lambda^{-1} \cdot \mathbb{E}\left[f(X)Y\right], \; b^* = 0 \qquad - \text{assume } \mathbb{E}\left[f\right] = 0, \mathsf{cov}(f) = I$$

$$\hat{y}_{\mathsf{SVM}}(x; f) = \mathsf{sgn}\left(\langle w^*, f(x)\rangle + b^*\right) = \mathsf{sgn}\left(\left\langle \mathbb{E}\left[f(X)Y\right], f(x)\right\rangle\right)$$

$$\boxed{\mathbb{E}\left[f(X)Y\right] \propto \mathbb{E}\left[f(X)|Y=1\right] - \mathbb{E}\left[f(X)|Y=-1\right]}$$



SVM for Given f

"Soft" Regime: large λ

$$w^* = \lambda^{-1} \cdot \mathbb{E}\left[f(X)Y\right], \; b^* = 0 \qquad - \text{assume } \mathbb{E}\left[f\right] = 0, \mathsf{cov}(f) = I$$

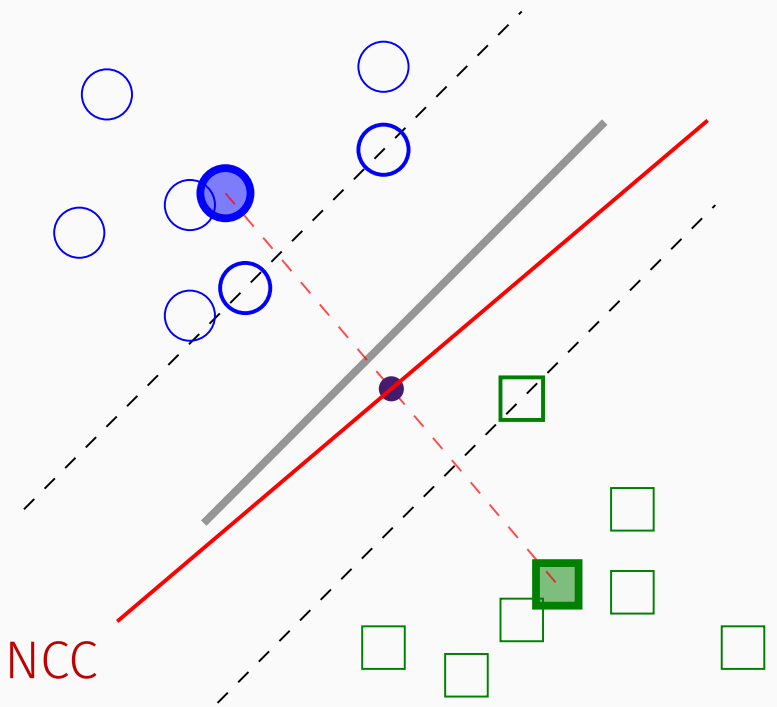
$$\hat{y}_{\mathsf{SVM}}(x; f) = \mathsf{sgn}\left(\langle w^*, f(x)\rangle + b^*\right) = \mathsf{sgn}\left(\left\langle \mathbb{E}\left[f(X)Y\right], f(x)\right\rangle\right)$$

$$\boxed{\mathbb{E}\left[f(X)Y\right] \propto \mathbb{E}\left[f(X)|Y=1\right] - \mathbb{E}\left[f(X)|Y=-1\right]}$$

Equivalent Form

$$\arg\min_y \left\| f(x) - \mathbb{E}\left[f(X)|Y=y\right] \right\|$$

- Nearest Centroid Classifier



SVM for Given f

"Soft" Regime: large λ

$$w^* = \lambda^{-1} \cdot \mathbb{E}\left[f(X)Y\right], \; b^* = 0 \qquad - \text{assume } \mathbb{E}\left[f\right] = 0, \mathsf{cov}(f) = I$$

$$\hat{y}_{\mathsf{SVM}}(x; f) = \mathsf{sgn}\left(\langle w^*, f(x)\rangle + b^*\right) = \mathsf{sgn}\left(\left\langle \mathbb{E}\left[f(X)Y\right], f(x)\right\rangle\right)$$

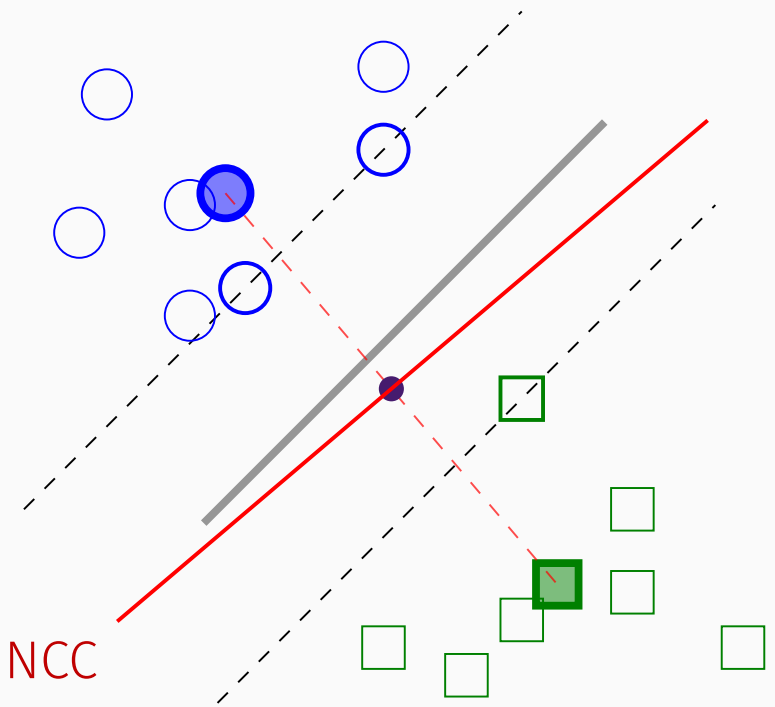
$$\boxed{\mathbb{E}\left[f(X)Y\right] \propto \mathbb{E}\left[f(X)|Y=1\right] - \mathbb{E}\left[f(X)|Y=-1\right]}$$

Equivalent Form

$$\arg\min_y \left\| f(x) - \mathbb{E}\left[f(X)|Y=y\right] \right\|$$

- Nearest Centroid Classifier

How to pick f?



SVM Decision

$$\hat{y}_{\mathrm{SVM}}(x; f) = \operatorname{sgn}\left(\left\langle \mathbb{E}\left[f(X)Y\right], f(x)\right\rangle\right)$$

SVM Decision

$$\hat{y}_{\mathrm{SVM}}(x; f) = \operatorname{sgn}\left(\left\langle \mathbb{E}\left[f(X)Y\right], f(x)\right\rangle\right)$$

- Based on a linear combination of $f_1, \ldots, f_d$

SVM Decision From Feature Geometry

SVM Decision

$$\hat{y}_{\mathrm{SVM}}(x; f) = \mathrm{sgn}\left(\left\langle \mathbb{E}\left[f(X)Y\right], f(x)\right\rangle\right)$$

- Based on a linear combination of $f_1, \ldots, f_d$

Optimal Decision

$$\hat{y}_{\mathrm{MAP}}(x) = \mathrm{sgn}(f^*(x)) \qquad f^*\text{: max. corr. function}$$

SVM Decision From Feature Geometry

SVM Decision

$$\hat{y}_{\mathrm{SVM}}(x; f) = \mathbf{sgn}\big(\big\langle \mathbb{E}\big[f(X)Y\big], f(x)\big\rangle\big)$$

- Based on a linear combination of $f_1, \ldots, f_d$

Optimal Decision

$$\hat{y}_{\mathrm{MAP}}(x) = \mathbf{sgn}(f^*(x)) \qquad f^*\text{: max. corr. function}$$

- SVM is optimal if $f = f^*$ $(d = 1)$

SVM Decision From Feature Geometry

SVM Decision

$$\hat{y}_{\mathrm{SVM}}(x; f) = \mathrm{sgn}\left(\left\langle \mathbb{E}\left[f(X)Y\right], f(x)\right\rangle\right)$$

- Based on a linear combination of $f_1, \ldots, f_d$

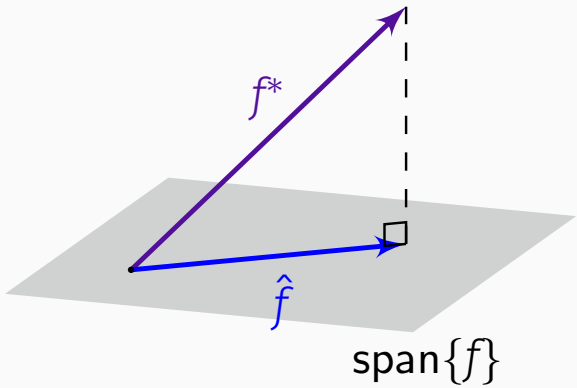
Optimal Decision

$$\hat{y}_{\mathrm{MAP}}(x) = \mathrm{sgn}(f^*(x)) \qquad f^*\text{: max. corr. function}$$

- SVM is optimal if $f = f^*$ $(d = 1)$

Optimal Linear Combination

$\hat{f}$: project f^* onto $\mathrm{span}\{f_1, \ldots, f_d\}$



SVM Decision From Feature Geometry

SVM Decision

$$\hat{y}_{\mathrm{SVM}}(x; f) = \mathsf{sgn}\left(\left\langle \mathbb{E}\left[f(X)Y\right], f(x)\right\rangle\right)$$

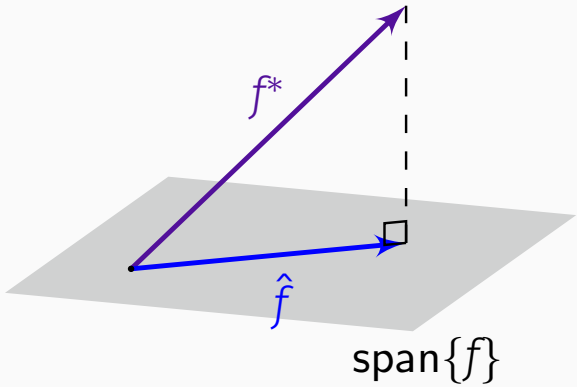
- Based on a linear combination of $f_1, \ldots, f_d$

Optimal Decision

$$\hat{y}_{\mathrm{MAP}}(x) = \mathsf{sgn}(f^*(x)) \qquad f^*\text{: max. corr. function}$$

- SVM is optimal if $f = f^*$ $(d = 1)$

Optimal Linear Combination



- $\hat{f}$: project f^* onto $\mathsf{span}\{f_1, \ldots, f_d\}$
- Feature space $\mathcal{F} \triangleq \{\phi\colon \mathcal{X} \to \mathbb{R}\}$
- Inner product $\langle \phi_1, \phi_2 \rangle_{\mathcal{F}} \triangleq \mathbb{E}\left[\phi_1(X)\phi_2(X)\right]$ (data dependent)

SVM Decision From Feature Geometry

SVM Decision

$$\hat{y}_{\mathrm{SVM}}(x; f) = \mathbf{sgn}\left(\left\langle \mathbb{E}\left[f(X)Y\right], f(x)\right\rangle\right) = \mathbf{sgn}\left(\hat{f}(x)\right)$$

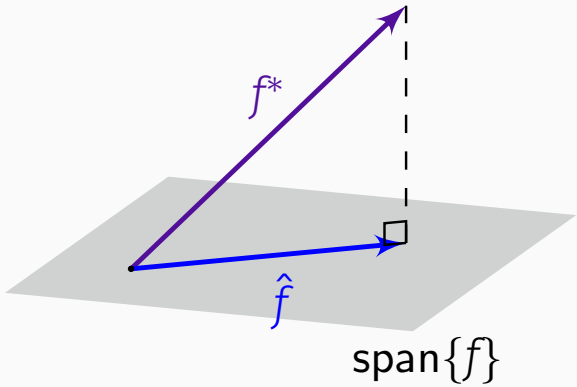
- Based on a linear combination of $f_1, \ldots, f_d$

Optimal Decision

$$\hat{y}_{\mathrm{MAP}}(x) = \mathbf{sgn}(f^*(x)) \qquad f^*\text{: max. corr. function}$$

- SVM is optimal if $f = f^*$ $(d = 1)$

Optimal Linear Combination



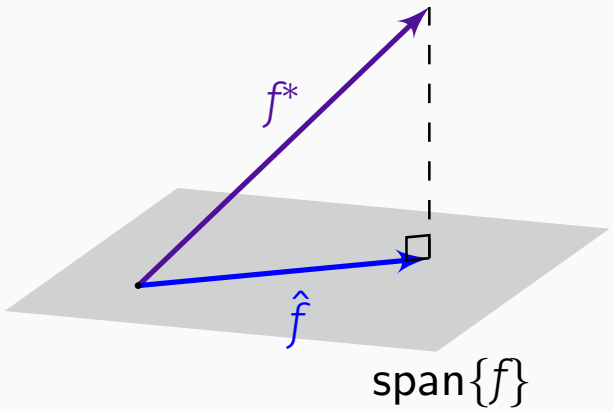
- $\hat{f}$: project f^* onto $\mathbf{span}\{f_1, \ldots, f_d\}$
- Feature space $\mathcal{F} \triangleq \{\phi\colon \mathcal{X} \to \mathbb{R}\}$
- Inner product $\langle \phi_1, \phi_2 \rangle_{\mathcal{F}} \triangleq \mathbb{E}\left[\phi_1(X)\phi_2(X)\right]$ (data dependent)

Feature Geometry and SVM Loss

SVM Decision

$$\hat{y}_{\text{SVM}}(x; f) = \text{sgn}\left(\hat{f}(x)\right)$$

- $\hat{f} = \Pi\left(f^*; \text{span}\{f\}\right)$

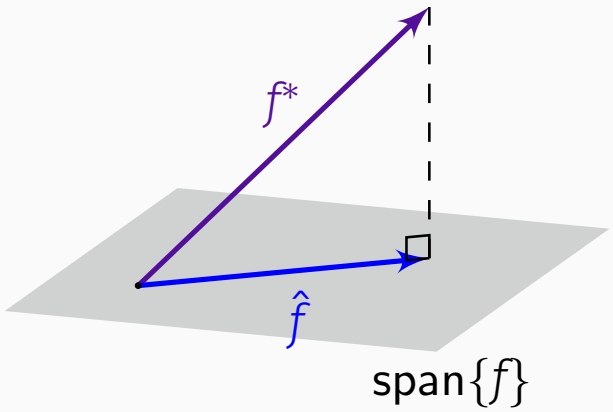


Feature Geometry and SVM Loss

SVM Decision

$$\hat{y}_{\text{SVM}}(x; f) = \text{sgn}\left(\hat{f}(x)\right)$$

- $\hat{f} = \Pi\left(f^*; \text{span}\{f\}\right)$



Optimal Decision

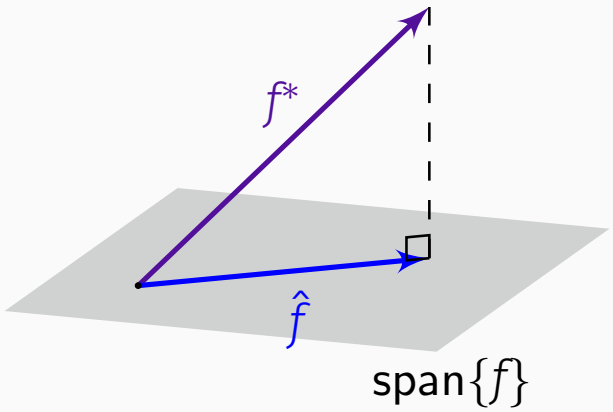
$$\hat{f} = f^* \implies \hat{y}_{\text{SVM}} = \hat{y}_{\text{MAP}}$$

Feature Geometry and SVM Loss

SVM Decision

$$\hat{y}_{\text{SVM}}(x; f) = \text{sgn}\left(\hat{f}(x)\right)$$

- $\hat{f} = \Pi\left(f^*; \text{span}\{f\}\right)$



Optimal Decision

$$\hat{f} = f^* \implies \hat{y}_{\text{SVM}} = \hat{y}_{\text{MAP}}$$

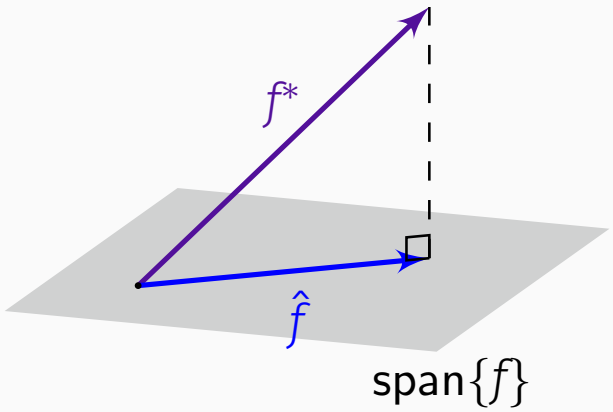
- require $f^* \in \text{span}\{f\}$

Feature Geometry and SVM Loss

SVM Decision

$$\hat{y}_{\text{SVM}}(x; f) = \operatorname{sgn}\left(\hat{f}(x)\right)$$

- $\hat{f} = \Pi\left(f^*; \operatorname{span}\{f\}\right)$



Optimal Decision

$$\hat{f} = f^* \implies \hat{y}_{\text{SVM}} = \hat{y}_{\text{MAP}}$$

- require $f^* \in \operatorname{span}\{f\}$

SVM Loss

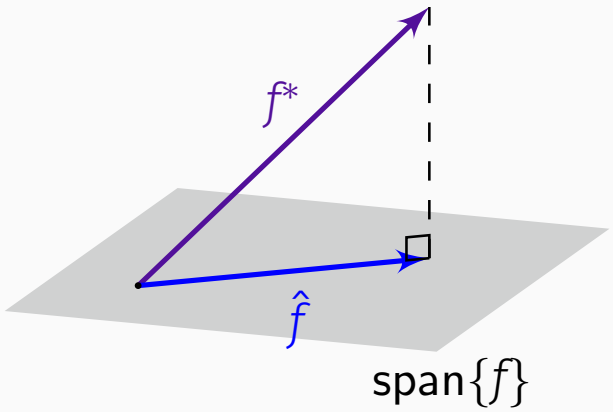
$$\text{Put } w^*, b^* \text{ in } \implies L^*_{\text{SVM}}(f) = 1 - \frac{C}{\lambda} \cdot \|\hat{f}\|^2_{\mathcal{F}}$$

Feature Geometry and SVM Loss

SVM Decision

$$\hat{y}_{\text{SVM}}(x; f) = \operatorname{sgn}\left(\hat{f}(x)\right)$$

- $\hat{f} = \Pi\left(f^*; \operatorname{span}\{f\}\right)$



Optimal Decision

$$\hat{f} = f^* \implies \hat{y}_{\text{SVM}} = \hat{y}_{\text{MAP}}$$

- require $f^* \in \operatorname{span}\{f\}$

SVM Loss

$$\text{Put } w^*, b^* \text{ in } \implies L^*_{\text{SVM}}(f) = 1 - \frac{C}{\lambda} \cdot \|\hat{f}\|^2_{\mathcal{F}}$$

- "oracle" feature extractor: minimize $L^*_{\text{SVM}}(f)$ over f

Kernel SVM Decision

Decision in terms of kernel $k(x, x') = \langle f(x), f(x') \rangle$

Kernel SVM Decision

Decision in terms of kernel $k(x, x') = \langle f(x), f(x') \rangle$

General Form

Rewrite $\hat{y}_{\text{SVM}}(x; f) = \textbf{sgn}\left(\left\langle \mathbb{E}\left[f(X)Y\right], f(x)\right\rangle\right)$

Kernel SVM Decision

Decision in terms of kernel $k(x, x') = \langle f(x), f(x') \rangle$

General Form

Rewrite $\hat{y}_{\text{SVM}}(x; f) = \textbf{sgn}\left(\left\langle \mathbb{E}\left[f(X)Y\right], f(x)\right\rangle\right)$

Kernel SVM Decision

Decision in terms of kernel $k(x, x') = \langle f(x), f(x') \rangle$

General Form

Rewrite $\hat{y}_{\text{SVM}}(x; f) = \text{sgn}\left(\left\langle \mathbb{E}\left[f(X)Y\right], f(x)\right\rangle\right)$

$= \text{sgn}(\mathbb{E}\left[k(X, x) \cdot Y\right])$

Kernel SVM Decision

Decision in terms of kernel $k(x, x') = \langle f(x), f(x') \rangle$

General Form

Rewrite $\hat{y}_{\text{SVM}}(x; f) = \text{sgn}\left(\left\langle \mathbb{E}\left[f(X)Y\right], f(x)\right\rangle\right)$

$$
\begin{aligned}
&= \text{sgn}(\mathbb{E}\left[k(X, x) \cdot Y\right]) \\
&= \text{sgn}(\mathbb{E}\left[k(X, x) f^*(X)\right]) && \text{— Replace } P_{XY} \text{ by } P_X P_Y \text{ and } f^*
\end{aligned}
$$

Kernel SVM Decision

Decision in terms of kernel $k(x, x') = \langle f(x), f(x') \rangle$

General Form

Rewrite $\hat{y}_{\text{SVM}}(x; f) = \text{sgn}\left(\left\langle \mathbb{E}\left[f(X)Y\right], f(x)\right\rangle\right)$

$$
\begin{aligned}
&= \text{sgn}(\mathbb{E}\left[k(X, x) \cdot Y\right]) \\
&= \text{sgn}(\mathbb{E}\left[k(X, x) f^*(X)\right]) \qquad \text{— Replace } P_{XY} \text{ by } P_X P_Y \text{ and } f^*
\end{aligned}
$$

Key Quantity

$\mathbb{E}\left[k(X, x) f^*(X)\right]$

Kernel SVM Decision

Decision in terms of kernel $k(x, x') = \langle f(x), f(x') \rangle$

General Form

Rewrite $\hat{y}_{\text{SVM}}(x; f) = \text{sgn}\left(\left\langle \mathbb{E}\left[f(X)Y\right], f(x)\right\rangle\right)$

$$
\begin{aligned}
&= \text{sgn}(\mathbb{E}\left[k(X, x) \cdot Y\right]) \\
&= \text{sgn}(\mathbb{E}\left[k(X, x) f^*(X)\right]) \qquad \text{— Replace } P_{XY} \text{ by } P_X P_Y \text{ and } f^*
\end{aligned}
$$

Key Quantity

$$\mathbb{E}\left[k(X, x) f^*(X)\right] = \int k(x, x') \cdot f^*(x')\, p_X(x') dx'$$

Kernel SVM Decision

Decision in terms of kernel $k(x, x') = \langle f(x), f(x') \rangle$

General Form

Rewrite $\hat{y}_{\text{SVM}}(x; f) = \text{sgn}\left(\left\langle \mathbb{E}\left[f(X)Y\right], f(x)\right\rangle\right)$

$$= \text{sgn}(\mathbb{E}\left[k(X, x) \cdot Y\right])$$

$$= \text{sgn}(\mathbb{E}\left[k(X, x) f^*(X)\right]) \qquad \text{— Replace } P_{XY} \text{ by } P_X P_Y \text{ and } f^*$$

Key Quantity

$$\mathbb{E}\left[k(X, x) f^*(X)\right] = \int k(x, x') \cdot f^*(x')\, p_X(x') dx' = [k \circ f^*](x)$$

— "kernel" of integral operator

Kernel SVM Decision

Decision in terms of kernel $k(x, x') = \langle f(x), f(x') \rangle$

General Form

Rewrite $\hat{y}_{\text{SVM}}(x; f) = \text{sgn}\left(\left\langle \mathbb{E}\left[f(X)Y\right], f(x)\right\rangle\right)$

$$= \text{sgn}(\mathbb{E}\left[k(X, x) \cdot Y\right])$$

$$= \text{sgn}(\mathbb{E}\left[k(X, x) f^*(X)\right]) \qquad \text{— Replace } P_{XY} \text{ by } P_X P_Y \text{ and } f^*$$

Key Quantity

$$\mathbb{E}\left[k(X, x) f^*(X)\right] = \int k(x, x') \cdot f^*(x')\, p_X(x') dx' = [k \circ f^*](x)$$

— "kernel" of integral operator

How to pick k?

Kernel and Feature Subspace

Kernel for Feature Subspace $\mathcal{G}$

$k_{\mathcal{G}}(x, x') \triangleq \langle \phi(x), \phi(x') \rangle \qquad \phi = (\phi_1, \ldots, \phi_d)^{\mathrm{T}}$: orthonormal basis

Kernel and Feature Subspace

Kernel for Feature Subspace $\mathcal{G}$

$k_{\mathcal{G}}(x, x') \triangleq \langle \phi(x), \phi(x') \rangle$ $\quad$ $\phi = (\phi_1, \ldots, \phi_d)^{\mathrm{T}}$: orthonormal basis

- $k_{\mathcal{G}} \circ f = \Pi(f; \mathcal{G})$ — Projection onto $\mathcal{G}$

Kernel and Feature Subspace

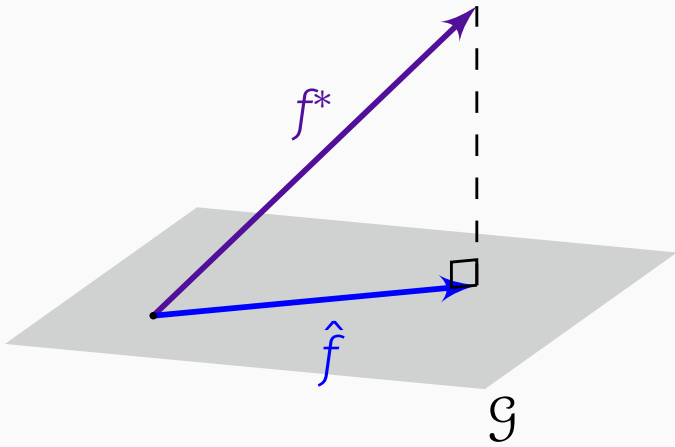
Kernel for Feature Subspace $\mathcal{G}$

$k_{\mathcal{G}}(x, x') \triangleq \langle \phi(x), \phi(x') \rangle \qquad \phi = (\phi_1, \ldots, \phi_d)^{\mathrm{T}}$: orthonormal basis

- $k_{\mathcal{G}} \circ f = \Pi(f; \mathcal{G})$ — Projection onto $\mathcal{G}$

SVM With $k_{\mathcal{G}}$

- Decision $\hat{y}_{\mathrm{SVM}}(x; k_{\mathcal{G}}) = \mathsf{sgn}\left(\hat{f}(x)\right)$
- Loss $L^*_{\mathrm{SVM}}(k_{\mathcal{G}}) = 1 - \dfrac{C}{\lambda} \cdot \|\hat{f}\|^2_{\mathcal{F}}$
- $\hat{f} \triangleq \Pi(f^*; \mathcal{G})$



Kernel and Feature Subspace

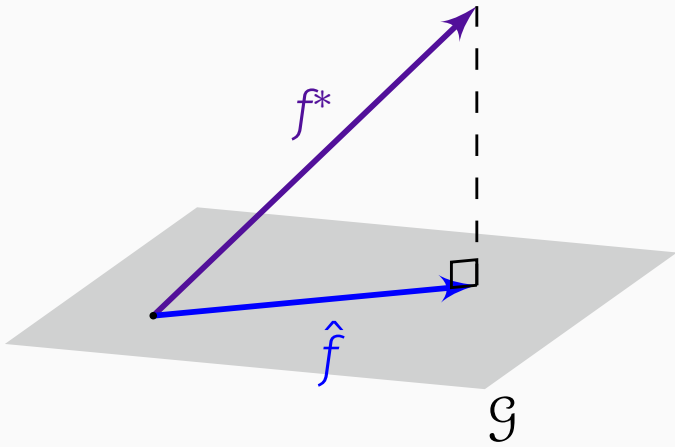
Kernel for Feature Subspace $\mathcal{G}$

$\kappa_{\mathcal{G}}(x, x') \triangleq \langle \phi(x), \phi(x') \rangle \quad \phi = (\phi_1, \ldots, \phi_d)^{\mathrm{T}}$: orthonormal basis

- $\kappa_{\mathcal{G}} \circ f = \Pi(f; \mathcal{G})$ — Projection onto $\mathcal{G}$

SVM With $\kappa_{\mathcal{G}}$

- Decision $\hat{y}_{\mathrm{SVM}}(x; \kappa_{\mathcal{G}}) = \mathrm{sgn}\left(\hat{f}(x)\right)$
- Loss $L^*_{\mathrm{SVM}}(\kappa_{\mathcal{G}}) = 1 - \frac{C}{\lambda} \cdot \|\hat{f}\|^2_{\mathcal{F}}$
- $\hat{f} \triangleq \Pi(f^*; \mathcal{G})$



Quality of $\kappa_{\mathcal{G}}$

- Determined by $\|\hat{f}\|_{\mathcal{F}}$, or the angle between f^* and $\mathcal{G}$

Kernel and Feature Subspace

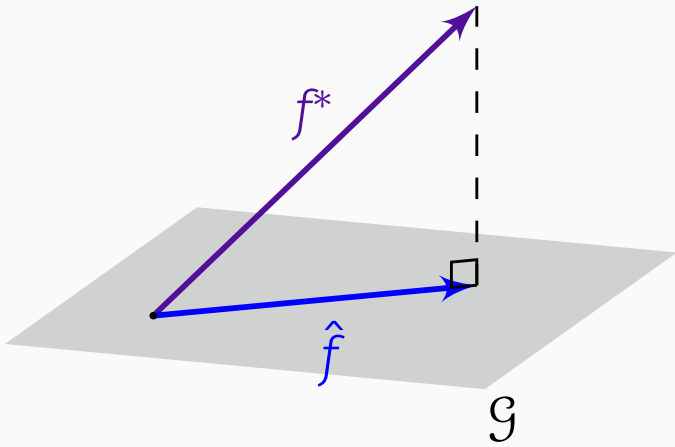
Kernel for Feature Subspace $\mathcal{G}$

$\kappa_{\mathcal{G}}(x, x') \triangleq \langle \phi(x), \phi(x') \rangle \quad \phi = (\phi_1, \ldots, \phi_d)^{\mathrm{T}}$: orthonormal basis

- $\kappa_{\mathcal{G}} \circ f = \Pi(f; \mathcal{G})$ — Projection onto $\mathcal{G}$

SVM With $\kappa_{\mathcal{G}}$

- Decision $\hat{y}_{\mathrm{SVM}}(x; \kappa_{\mathcal{G}}) = \mathrm{sgn}\left(\hat{f}(x)\right)$
- Loss $L^*_{\mathrm{SVM}}(\kappa_{\mathcal{G}}) = 1 - \frac{C}{\lambda} \cdot \|\hat{f}\|^2_{\mathcal{F}}$
- $\hat{f} \triangleq \Pi(f^*; \mathcal{G})$



Quality of $\kappa_{\mathcal{G}}$

- Determined by $\|\hat{f}\|_{\mathcal{F}}$, or the angle between f^* and $\mathcal{G}$
- Optimal choice: $\mathcal{G} \ni f^*$

Kernel on Parameterized Family

Two classes from the same family $\pi(x;\theta)$

- $Y = 1 \leftrightarrow \theta_0$
- $Y = -1 \leftrightarrow -\theta_0$

$$P_{X|Y=1}(x) = \pi(x;\theta_0)$$
$$P_{X|Y=-1}(x) = \pi(x;-\theta_0)$$

Kernel on Parameterized Family

Two classes from the same family $\pi(x;\theta)$

- $Y = 1 \leftrightarrow \theta_0$
- $Y = -1 \leftrightarrow -\theta_0$

$$P_{X|Y=1}(x) = \pi(x;\theta_0)$$
$$P_{X|Y=-1}(x) = \pi(x;-\theta_0)$$

Optimal Kernel

$$f^*(x) \propto \frac{P_{X|Y=1}(x) - P_{X|Y=-1}(x)}{P_X(x)}$$

Kernel on Parameterized Family

Two classes from the same family $\pi(x;\theta)$

- $Y = 1 \leftrightarrow \theta_0$
- $Y = -1 \leftrightarrow -\theta_0$

$$P_{X|Y=1}(x) = \pi(x;\theta_0)$$
$$P_{X|Y=-1}(x) = \pi(x;-\theta_0)$$

Optimal Kernel

$$f^*(x) \propto \frac{P_{X|Y=1}(x) - P_{X|Y=-1}(x)}{P_X(x)} \propto \frac{\nabla_\theta \, \pi(x;\theta)|_{\theta=0}}{\pi(x;0)}$$

Kernel on Parameterized Family

Two classes from the same family $\pi(x;\theta)$

- $Y = 1 \leftrightarrow \theta_0$
- $Y = -1 \leftrightarrow -\theta_0$

$$P_{X|Y=1}(x) = \pi(x;\theta_0)$$
$$P_{X|Y=-1}(x) = \pi(x;-\theta_0)$$

Optimal Kernel

$$f^*(x) \propto \frac{P_{X|Y=1}(x) - P_{X|Y=-1}(x)}{P_X(x)} \propto \frac{\nabla_\theta \, \pi(x;\theta)|_{\theta=0}}{\pi(x;0)}$$
$$= \nabla_\theta \log \pi(x;\theta)|_{\theta=0}$$

Kernel on Parameterized Family

Two classes from the same family $\pi(x;\theta)$

- $Y = 1 \leftrightarrow \theta_0$
- $Y = -1 \leftrightarrow -\theta_0$

$$P_{X|Y=1}(x) = \pi(x;\theta_0)$$
$$P_{X|Y=-1}(x) = \pi(x;-\theta_0)$$

Optimal Kernel

$$f^*(x) \propto \frac{P_{X|Y=1}(x) - P_{X|Y=-1}(x)}{P_X(x)} \propto \frac{\nabla_\theta \, \pi(x;\theta)|_{\theta=0}}{\pi(x;0)}$$
$$= \nabla_\theta \log \pi(x;\theta)|_{\theta=0}$$

- f^* is score function $s(x)$

Kernel on Parameterized Family

Two classes from the same family $\pi(x;\theta)$

- $Y = 1 \leftrightarrow \theta_0$
- $Y = -1 \leftrightarrow -\theta_0$

$$P_{X|Y=1}(x) = \pi(x;\theta_0)$$
$$P_{X|Y=-1}(x) = \pi(x;-\theta_0)$$

Optimal Kernel

$$f^*(x) \propto \frac{P_{X|Y=1}(x) - P_{X|Y=-1}(x)}{P_X(x)} \propto \frac{\nabla_\theta\, \pi(x;\theta)|_{\theta=0}}{\pi(x;0)}$$
$$= \nabla_\theta \log \pi(x;\theta)|_{\theta=0}$$

- f^* is score function $s(x)$
- optimal $k(x,x') = \langle s(x), s(x') \rangle$

Kernel on Parameterized Family

Two classes from the same family $\pi(x;\theta)$

- $Y = 1 \leftrightarrow \theta_0$ $\qquad P_{X|Y=1}(x) = \pi(x;\theta_0)$
- $Y = -1 \leftrightarrow -\theta_0$ $\qquad P_{X|Y=-1}(x) = \pi(x;-\theta_0)$

Optimal Kernel

$$f^*(x) \propto \frac{P_{X|Y=1}(x) - P_{X|Y=-1}(x)}{P_X(x)} \propto \frac{\nabla_\theta \, \pi(x;\theta)|_{\theta=0}}{\pi(x;0)}$$
$$= \nabla_\theta \log \pi(x;\theta)|_{\theta=0}$$

- f^* is score function $s(x)$
- optimal $k(x,x') = \langle s(x), s(x') \rangle$ — Fisher kernel!

Kernel on Parameterized Family

Two classes from the same family $\pi(x;\theta)$

- $Y = 1 \leftrightarrow \theta_0$ $\qquad P_{X|Y=1}(x) = \pi(x;\theta_0)$
- $Y = -1 \leftrightarrow -\theta_0$ $\qquad P_{X|Y=-1}(x) = \pi(x;-\theta_0)$

Optimal Kernel

$$f^*(x) \propto \frac{P_{X|Y=1}(x) - P_{X|Y=-1}(x)}{P_X(x)} \propto \frac{\nabla_\theta\, \pi(x;\theta)|_{\theta=0}}{\pi(x;0)}$$
$$= \nabla_\theta \log \pi(x;\theta)|_{\theta=0}$$

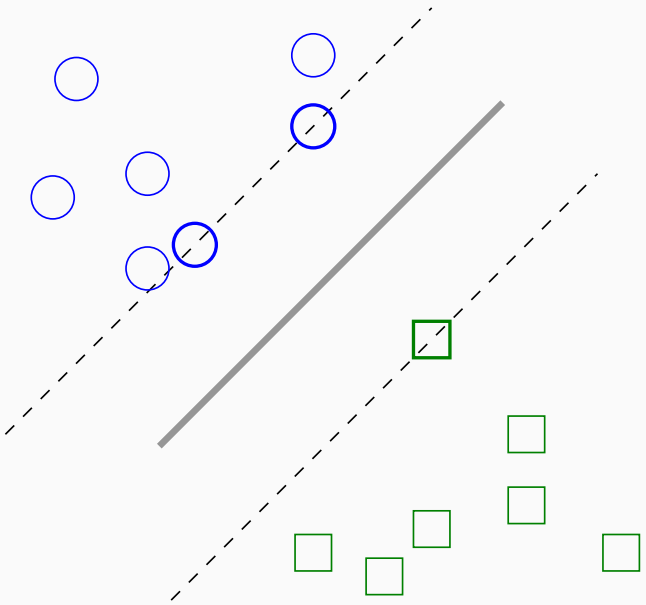
- f^* is score function $s(x)$
- optimal $k(x,x') = \langle s(x), s(x') \rangle$ — Fisher kernel!

Quality of Kernel $k_{\mathcal{G}}$

- $\| \Pi(s;\mathcal{G}) \|_{\mathcal{F}}$

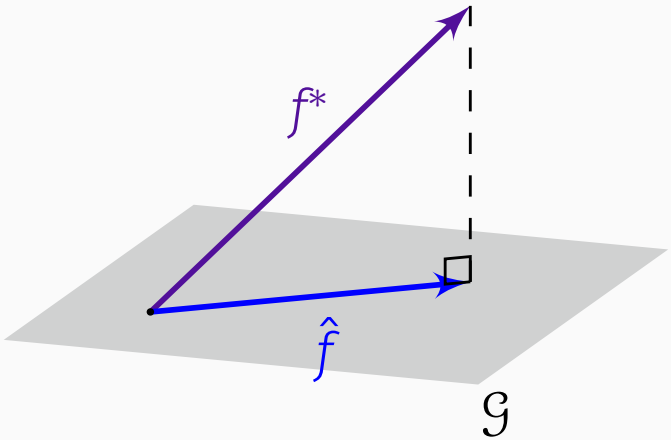
Kernel Subspace and Feature Extraction

Euclidean Geometry



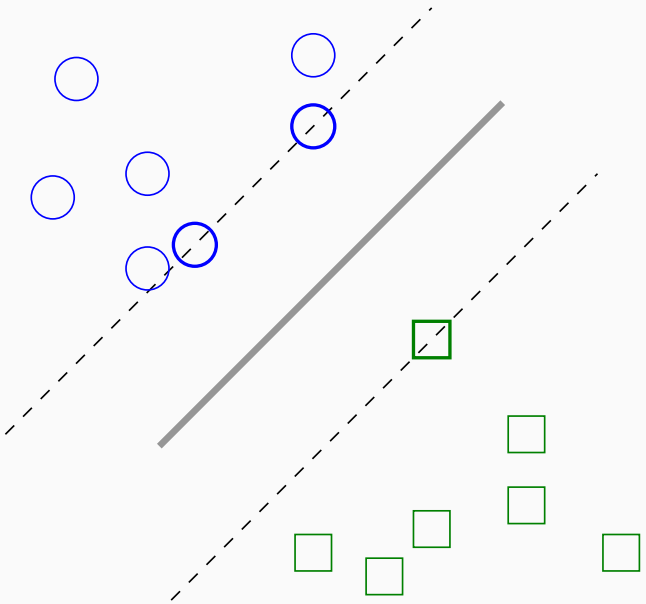
$\rightarrow$

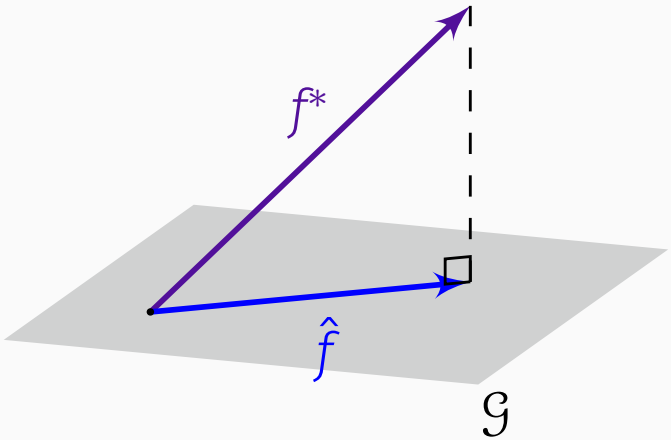
Feature Geometry



Kernel Subspace and Feature Extraction

Euclidean Geometry $\rightarrow$ Feature Geometry





- Feature subspace $\mathcal{G} \leftrightarrow$ Kernel $k_{\mathcal{G}}$
- Performance decided by $\angle(f^*, \mathcal{G})$
- Quantitative metric of kernel quality