Data Mining Techniques

CS 6220 - Section 3 - Fall 2016

Lecture 6: Classification 3

Jan-Willem van de Meent
(*credit*: Yijun Zhao, Arthur Gretton)

Class Schedule Updates

SCHEDULE

Note: This schedule is subject to change and will be adjusted as needed throughout the semester.

Wk	Day	Lectures	Homework	Project
1	07 Sep	Introduction 1: Course Overview		
	09 Sep	Introduction 2: Linear regression, Overfitting, Cross validation		
2	14 Sep	Introduction 3: Probability, Bayes Rule, Conjugacy	#1 out	Vote on type
	16 Sep	Classification 1: k-NN, Logistic Regression, Linear Discriminant Analysis		
3	21 Sep	Classification 2: Naive Bayes, Support Vector Machines		
	23 Sep	Classification 3: Non-linear SVMs, Kernels		
4	28 Sep	Classification 4: Ensemble Methods, Boosting, Random Forests	#2 out	Teams due
	30 Sep	Clustering 1: K-means, K-medioids	#3 due	
5	05 Oct	Clustering 2: DBSCAN, Mixture Models		
	07 Oct	Clustering 3: Expectation Maximimization		
6	12 Oct	Topic Models: pLSA, Latent Dirichlet Allocation	#3 out	
	14 Oct	Dimensionality Reduction 1: PCA, SVD, ICA	#2 due	
7	19 Oct	Dimensionality Reduction 2: Random Projections		
	21 Oct	Recommender Systems		
8	26 Oct	Midterm exam		

Class Schedule Updates

SCHEDULE

Note: This schedule is subject to change and will be adjusted as needed throughout the semester.

Wk	Day	Lectures	Homework	Project
1	07 Sep	Introduction 1: Course Overview		
	09 Sep	Introduction 2: Linear regression, Overfitting, Cross validation		
2	14 Sep	Introduction 3: Probability, Bayes Rule, Conjugacy	#1 out	Vote on type
	16 Sep	Classification 1: k-NN, Logistic Regression, Linear Discriminant Analysis		
3	21 Sep	Classification 2: Naive Bayes, Support Vector Machines		
	23 Sep	Classification 3: Non-linear SVMs, Kernels		
4	28 Sep	Classification 4: Ensemble Methods, Boosting, Random Forests	#2 out	Teams due
	30 Sep	Clustering 1: K-means, K-medioids	#3 due	
5	05 Oct	Clustering 2: DBSCAN, Mixture Models		
	07 Oct	Clustering 3: Expectation Maximimization		
6	12 Oct	Topic Models: pLSA, Latent Dirichlet Allocation	#3 out	
	14 Oct	Dimensionality Reduction 1: PCA, SVD, ICA	#2 due	
7	19 Oct	Dimensionality Reduction 2: Random Projections		
	21 Oct	Recommender Systems		
8	26 Oct	Midterm exam		

1 extra

1 instead of 2

Class Schedule Updates

SCHEDULE

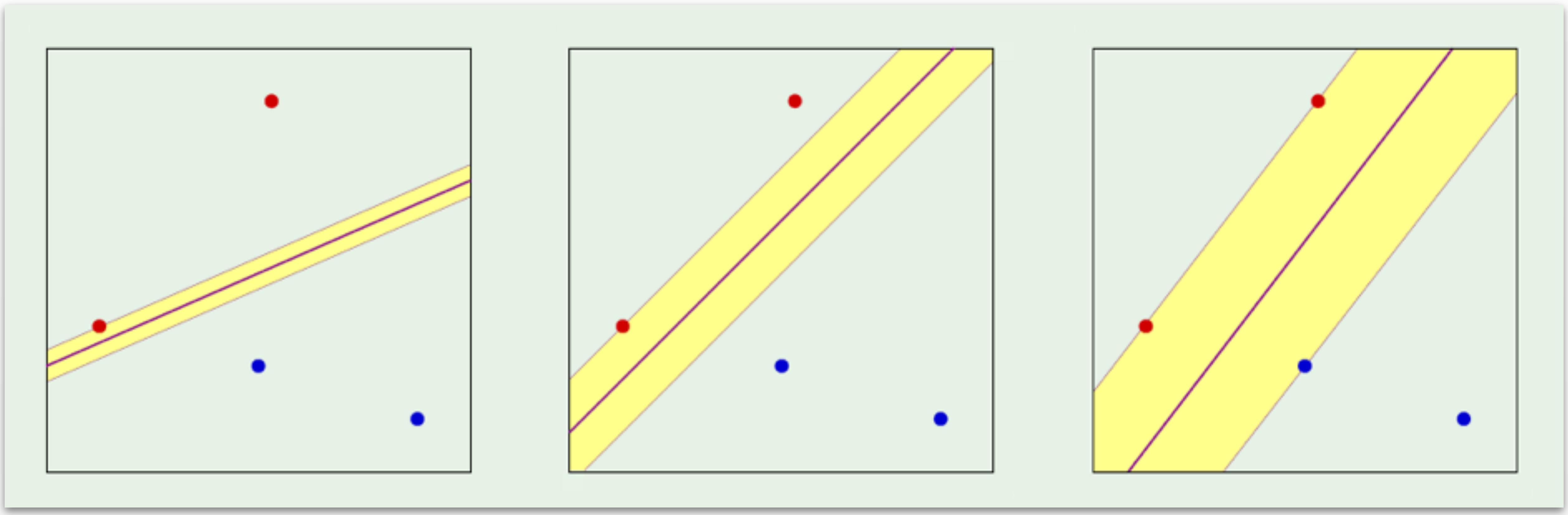
Note: This schedule is subject to change and will be adjusted as needed throughout the semester.

Wk	Day	Lectures	Homework	Project
1	07 Sep	Introduction 1: Course Overview		
	09 Sep	Introduction 2: Linear regression, Overfitting, Cross validation		
2	14 Sep	Introduction 3: Probability, Bayes Rule, Conjugacy	#1 out	Vote on type
	16 Sep	Classification 1: k-NN, Logistic Regression, Linear Discriminant Analysis		
3	21 Sep	Classification 2: Naive Bayes, Support Vector Machines		
	23 Sep	Classification 3: Non-linear SVMs, Kernels		
4	28 Sep	Classification 4: Ensemble Methods, Boosting, Random Forests	#2 out	Teams due
	30 Sep	Clustering 1: K-means, K-medioids	#3 due	
5	05 Oct	Clustering 2: DBSCAN, Mixture Models		
	07 Oct	Clustering 3: Expectation Maximimization		
6	12 Oct	Topic Models: pLSA, Latent Dirichlet Allocation	#3 out	
	14 Oct	Dimensionality Reduction 1: PCA, SVD, ICA	#2 due	
7	19 Oct	Dimensionality Reduction 2: Random Projections		
	21 Oct	Recommender Systems		
8	26 Oct	Midterm exam		

Project teams due next week

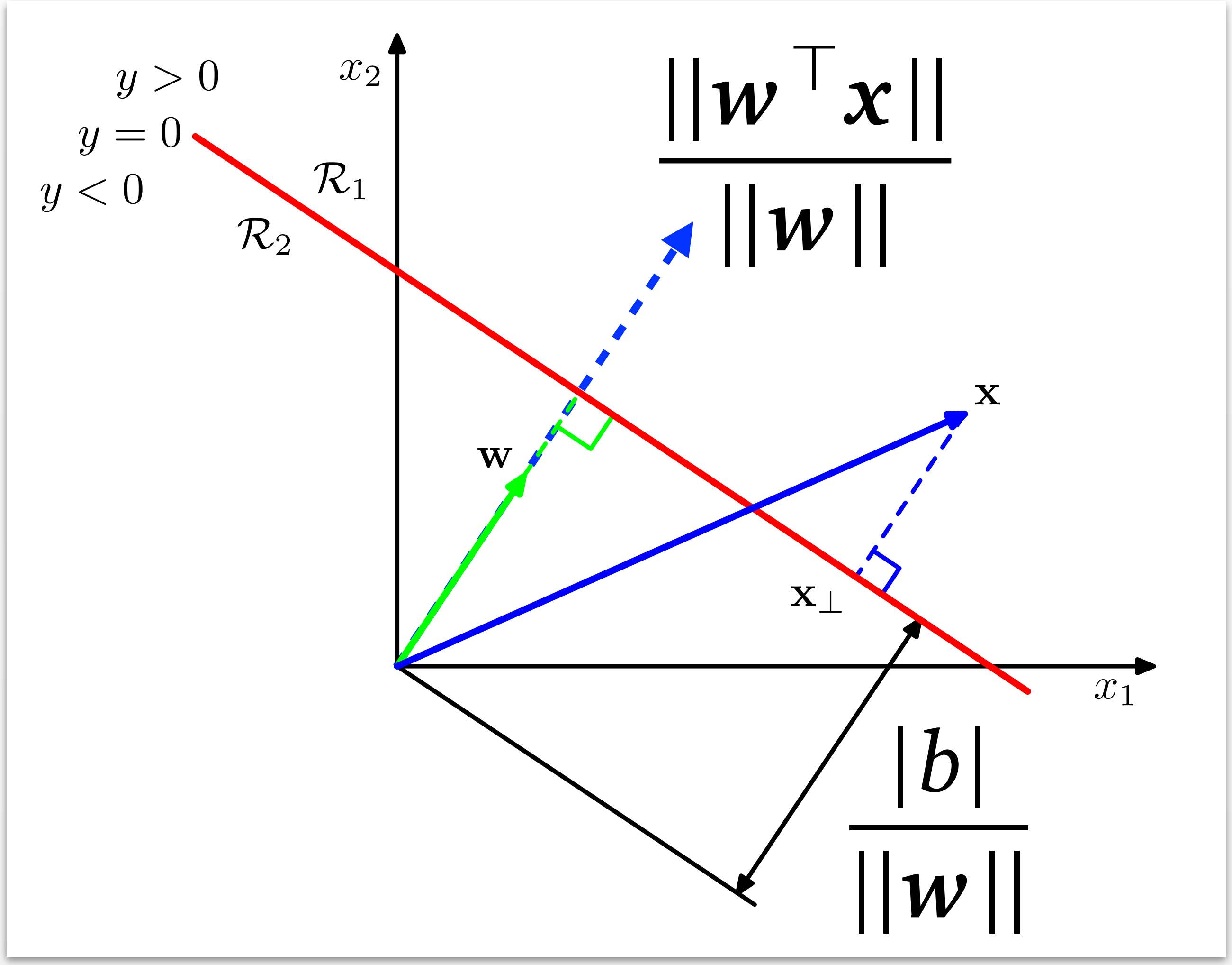
Support Vector Machines
(recap)

Max Margin Classifiers



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

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

SVMs as Convex Optimization

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

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

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

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

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

SVMs as Convex Optimization

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

Generalized Lagrangian

$$\mathcal{L}(\boldsymbol{w}, b, \alpha) = \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} - \sum_{i=1}^{m} \alpha_i \big(y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - 1\big)$$

Dual problem

$$W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j \boldsymbol{x}_i^\top \boldsymbol{x}_j$$

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

$$\mathcal{L}(\boldsymbol{w}, b, \alpha) = \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} - \sum_{i=1}^{m} \alpha_i \big(y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - 1\big)$$

Dual problem

$$W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$$

SVMs as Convex Optimization

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

Dual problem

$$W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$$

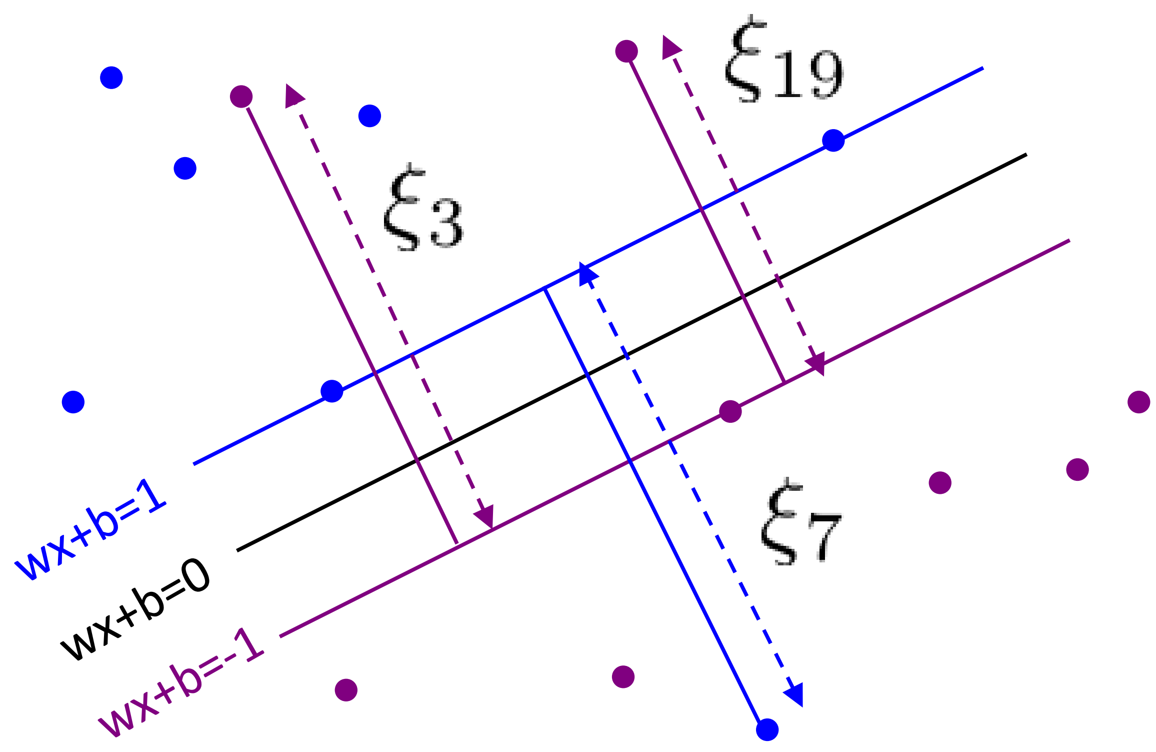
Sum over support vectors during prediction

$$\boldsymbol{w}^\top \boldsymbol{x} + b = \sum_{i=1}^{m} \alpha_i y_i \langle \boldsymbol{x}_i, \boldsymbol{x} \rangle + b$$

Soft-margin SVMs

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

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



Loss Function

$$
\begin{aligned}
\arg \min_{\mathbf{w}, b, \xi \geq 0} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{m}\xi_i \\
\text{s.t.} \quad & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m
\end{aligned}
$$

$$
E^{\mathrm{SVM}}(\boldsymbol{w}) = \sum_{i=1}^{m} \xi_i + \frac{1}{2C}\boldsymbol{w}^\top \boldsymbol{w}
$$

Loss Function

$$
\begin{aligned}
\arg \min_{\mathbf{w}, b, \xi \geq 0} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{m} \xi_i \\
\text{s.t.} \quad & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m
\end{aligned}
$$

$$
\begin{aligned}
E^{\mathrm{SVM}}(\boldsymbol{w}) &= \sum_{i=1}^{m} \xi_i + \frac{1}{2C}\boldsymbol{w}^\top\boldsymbol{w} \\
&= \sum_{i=1}^{m} \big(1 - y_i(\boldsymbol{w}^\top\boldsymbol{x}_i + b)\big)_+ + \frac{1}{2C}\boldsymbol{w}^\top\boldsymbol{w}
\end{aligned}
$$

Loss Function

$$\begin{aligned} \arg\min_{\mathbf{w},b,\xi\geq 0} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m \end{aligned}$$

$$\begin{aligned} E^{\mathrm{SVM}}(\boldsymbol{w}) &= \sum_{i=1}^{m}\xi_i + \frac{1}{2C}\boldsymbol{w}^\top\boldsymbol{w} \\ &= \boxed{\sum_{i=1}^{m}\big(1 - y_i(\boldsymbol{w}^\top\boldsymbol{x}_i + b)\big)_+} + \frac{1}{2C}\boldsymbol{w}^\top\boldsymbol{w} \end{aligned}$$

Hinge Loss

Loss Function

$$
\begin{aligned}
\arg \min_{\mathbf{w}, b, \xi \geq 0} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{m} \xi_i \\
\text{s.t.} \quad & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m
\end{aligned}
$$

$$
\begin{aligned}
E^{\text{SVM}}(\boldsymbol{w}) &= \sum_{i=1}^{m} \xi_i + \frac{1}{2C}\boldsymbol{w}^\top \boldsymbol{w} \\
&= \sum_{i=1}^{m} \big(1 - y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)\big)_+ + \boxed{\frac{1}{2C}\boldsymbol{w}^\top \boldsymbol{w}}
\end{aligned}
$$

Regularization

Loss Function

$$\begin{aligned} \arg\min_{\mathbf{w},b,\xi\geq 0} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m \end{aligned}$$

$$\begin{aligned} E^{\text{SVM}}(\boldsymbol{w}) &= \sum_{i=1}^{m}\xi_i + \frac{1}{2C}\boldsymbol{w}^\top\boldsymbol{w} \\ &= \sum_{i=1}^{m}\left(1 - y_i(\boldsymbol{w}^\top\boldsymbol{x}_i + b)\right)_+ + \boxed{\lambda\boldsymbol{w}^\top\boldsymbol{w}} \end{aligned}$$

Regularization

Relationship to Logistic Regression

$$E^{\mathrm{SVM}}(\boldsymbol{w}) = \sum_{i=1}^{m} \left(1 - y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)\right)_+ + \lambda \boldsymbol{w}^\top \boldsymbol{w}$$

Relationship to Logistic Regression

$$E^{\mathrm{SVM}}(\boldsymbol{w}) = \sum_{i=1}^{m} \big(1 - y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)\big)_+ + \lambda \boldsymbol{w}^\top \boldsymbol{w}$$

$$\begin{aligned} E^{\mathrm{LR}}(\boldsymbol{w}) &= -\log p(y \mid \boldsymbol{x}, \boldsymbol{w}, b) \\ &= -\sum_{i=1}^{N} \log \frac{1}{(1 + e^{-y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)})} \\ &= \sum_{i=1}^{N} \log(1 + e^{-y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)}) \end{aligned}$$

Relationship to Logistic Regression

$$E^{\mathrm{SVM}}(\boldsymbol{w}) = \sum_{i=1}^{m} \left(1 - y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)\right)_+ + \lambda \boldsymbol{w}^\top \boldsymbol{w}$$

$$E^{\mathrm{LR}}(\boldsymbol{w}) = -\log p(y \,|\, \boldsymbol{x}, \boldsymbol{w}, b)$$

Relationship to Logistic Regression

$$E^{\mathrm{SVM}}(\boldsymbol{w}) = \sum_{i=1}^{m} \left(1 - y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)\right)_+ + \lambda \boldsymbol{w}^\top \boldsymbol{w}$$

$$\begin{aligned} E^{\mathrm{LR}}(\boldsymbol{w}) &= -\log p(y \mid \boldsymbol{x}, \boldsymbol{w}, b) \\ &= -\sum_{i=1}^{N} \log \frac{1}{(1 + e^{-y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)})} \end{aligned}$$

Relationship to Logistic Regression

$$E^{\mathrm{SVM}}(\boldsymbol{w}) = \sum_{i=1}^{m} \left(1 - y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)\right)_+ + \lambda \boldsymbol{w}^\top \boldsymbol{w}$$

$$\begin{aligned} E^{\mathrm{LR}}(\boldsymbol{w}) &= -\log p(y \,|\, \boldsymbol{x}, \boldsymbol{w}, b) \\ &= -\sum_{i=1}^{N} \log \frac{1}{(1 + e^{-y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)})} \\ &= \sum_{i=1}^{N} \log(1 + e^{-y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)}) \end{aligned}$$

Relationship to Logistic Regression

$$E^{\mathrm{SVM}}(\boldsymbol{w}) = \sum_{i=1}^{m} \left(1 - y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)\right)_+ + \lambda \boldsymbol{w}^\top \boldsymbol{w}$$

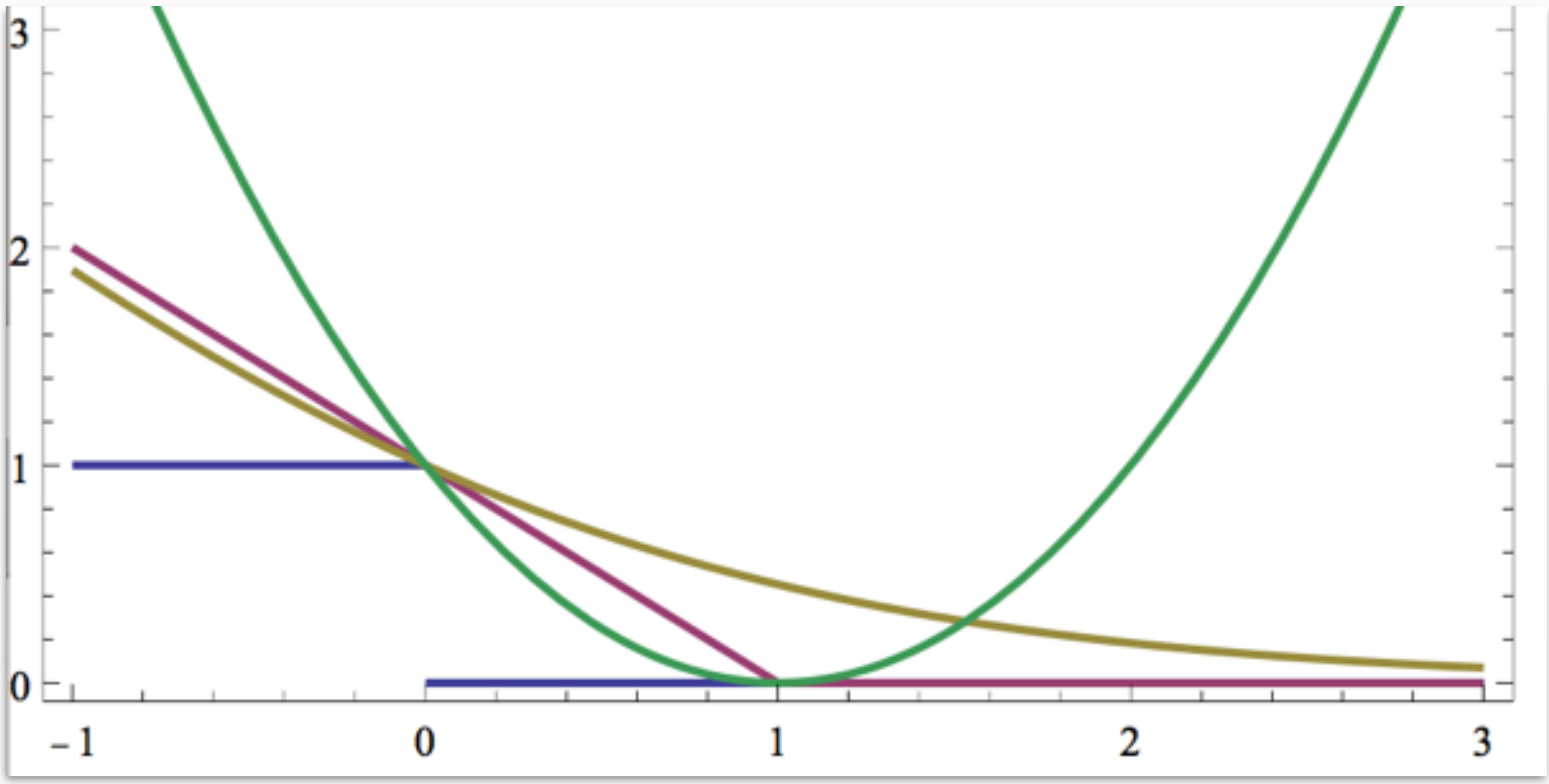
$$E^{\mathrm{LR}}(\boldsymbol{w}) = \sum_{i=1}^{N} \log(1 + e^{-y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)})$$

Relationship to Logistic Regression

$$E^{\mathrm{SVM}}(\boldsymbol{w}) = \sum_{i=1}^{m} \left(1 - y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)\right)_+ + \lambda \boldsymbol{w}^\top \boldsymbol{w}$$

$$E^{\mathrm{LR}}(\boldsymbol{w}) = \sum_{i=1}^{N} \log(1 + e^{-y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)}) + \lambda \boldsymbol{w}^\top \boldsymbol{w}$$

Loss Functions



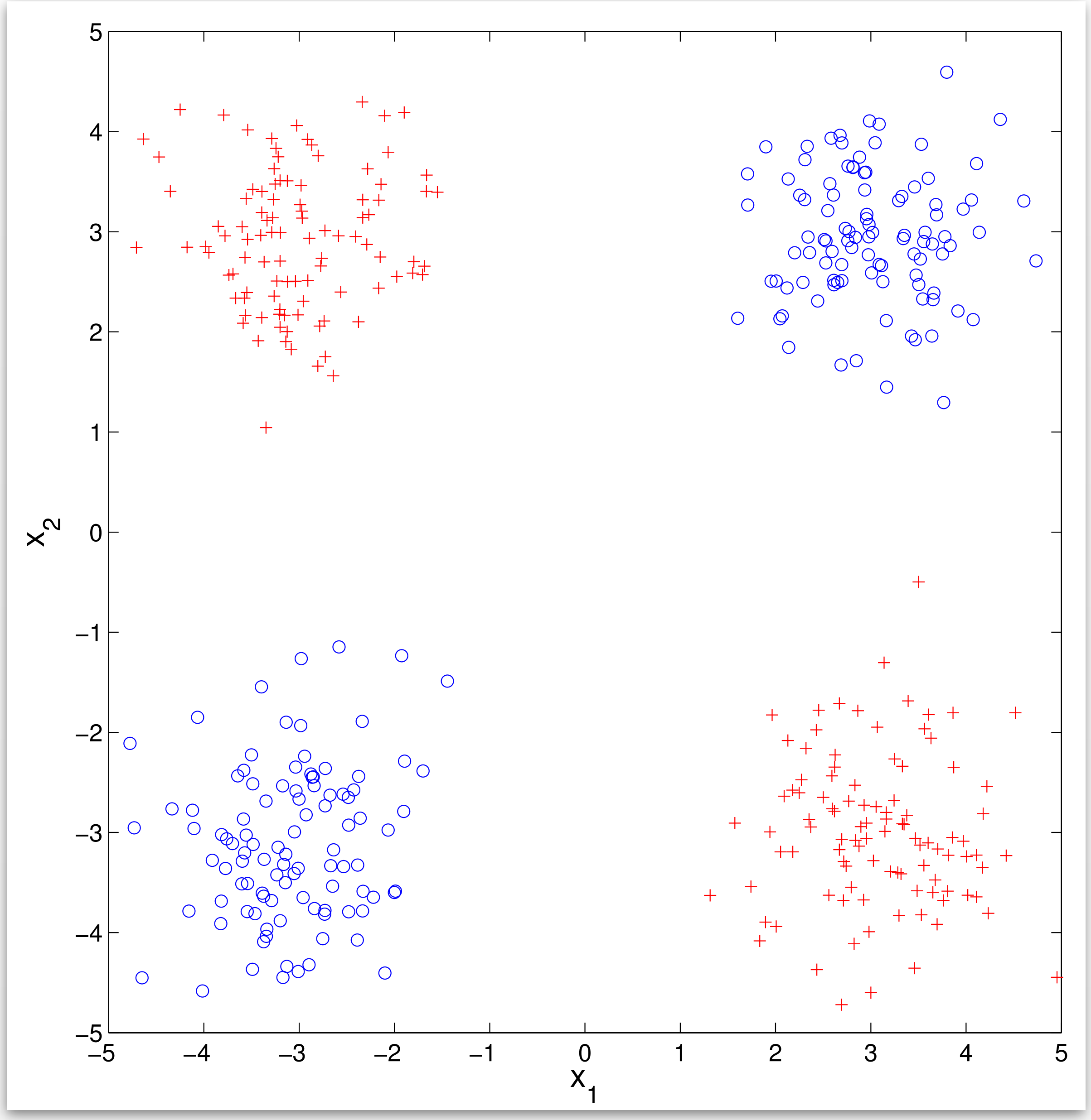
squared loss: $\frac{1}{2}(\boldsymbol{w}^\top \boldsymbol{x} - y)^2$

logistic loss: $\log\left(1 + \exp(-y\boldsymbol{w}^\top \boldsymbol{x})\right)$

hinge loss: $\max\{0, 1 - y\boldsymbol{w}^\top \boldsymbol{x}\}$

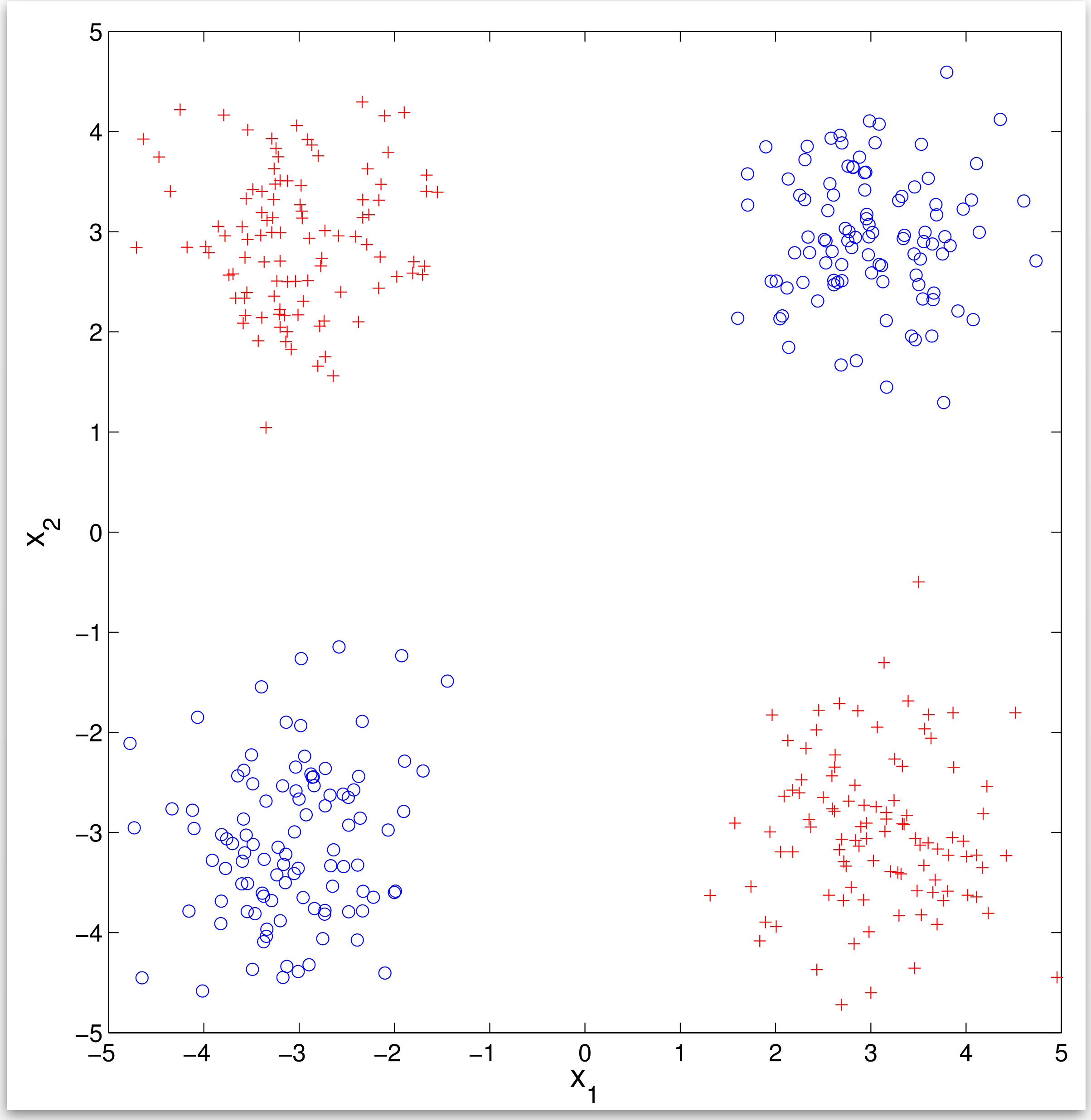
Nonlinear SVMs

Inseparable Problems



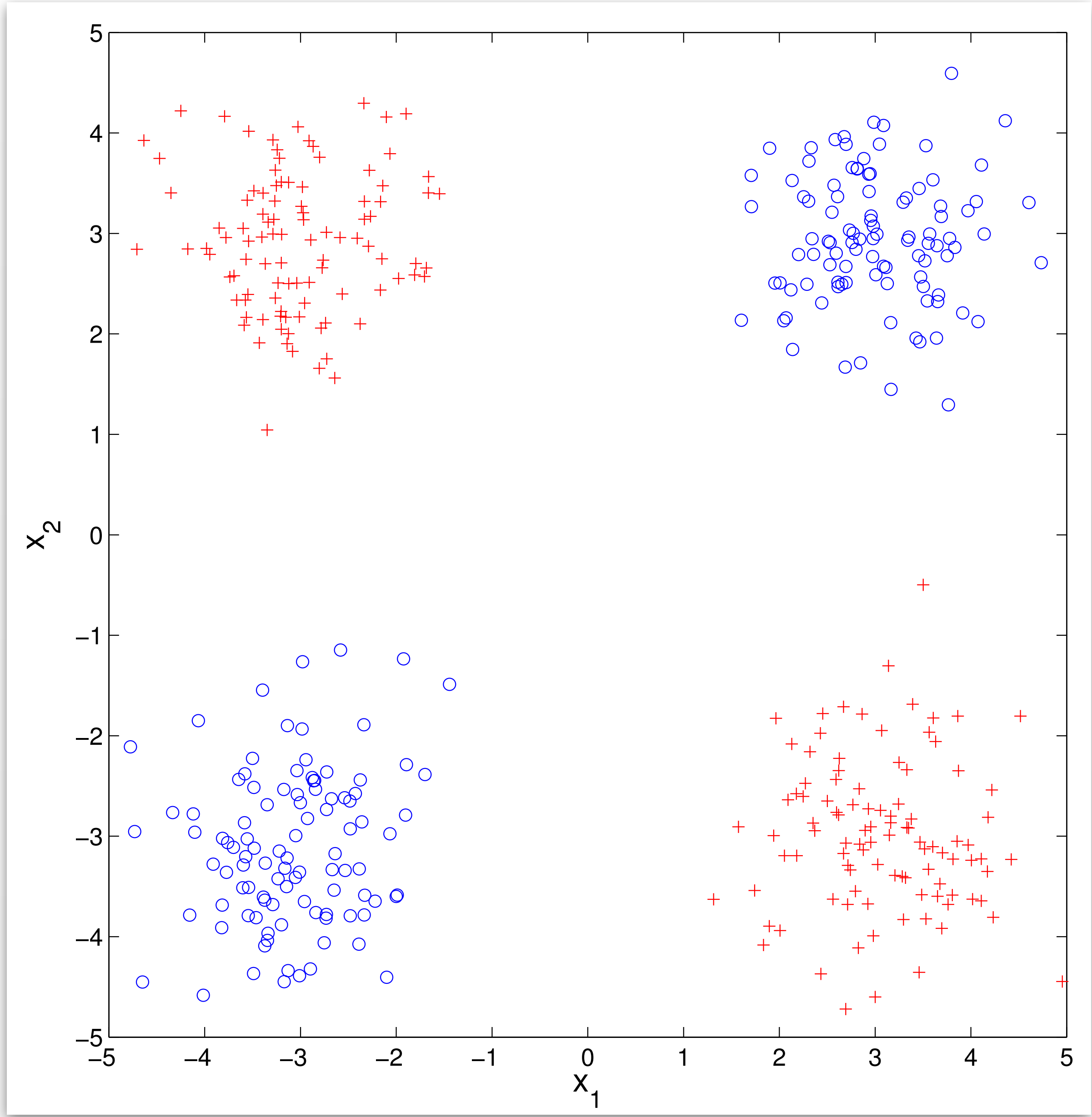
No linear classifier

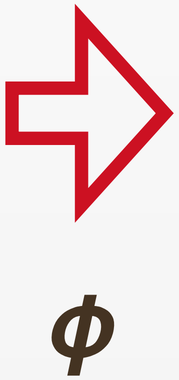
Inseparable Problems

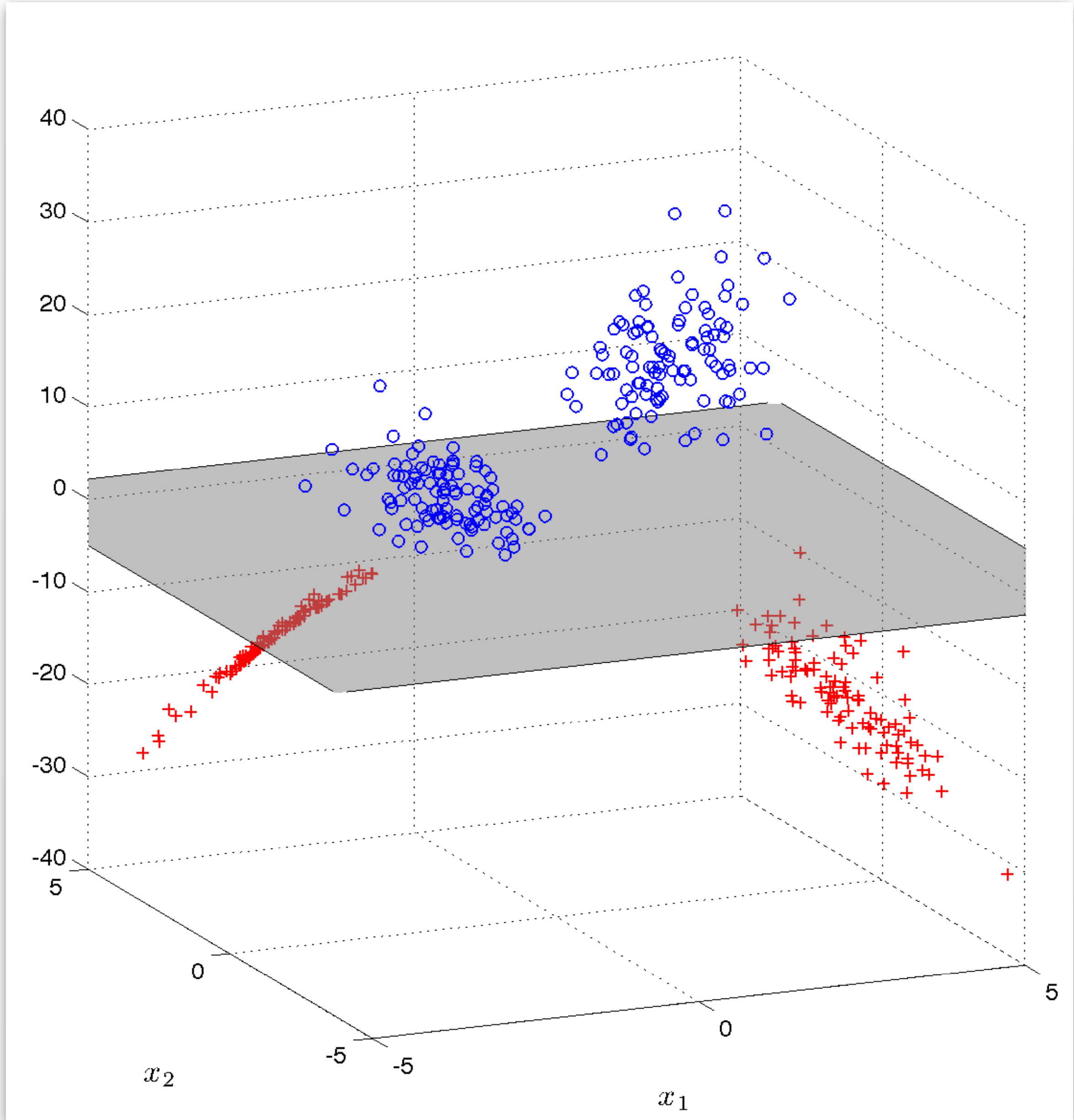


Idea: Map features onto higher dimensional space

Feature Map







$$\phi(x) = \begin{bmatrix} x_1 & x_2 & x_1 x_2 \end{bmatrix} \in \mathbb{R}^3$$

SVMs with Feature Maps

Dual problem

$$W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j \boldsymbol{x}_i^\top \boldsymbol{x}_j$$

Dual problem with feature map

$$W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j \phi(\boldsymbol{x}_i)^\top \phi(\boldsymbol{x}_j)$$

Computational Cost

Example: Mapping with linear and quadratic terms

$$\boldsymbol{x} = (x_1, \ldots, x_d)$$

$$\phi(\boldsymbol{x}) = (1, x_1, \ldots, x_d, x_1 x_1, x_1 x_2, \ldots x_d x_d)$$

Computational Cost

Example: Mapping with linear and quadratic terms

$$\boldsymbol{x} = (x_1, \ldots, x_d)$$

$$\phi(\boldsymbol{x}) = \boxed{(1, x_1, \ldots, x_d, x_1x_1, x_1x_2, \ldots x_dx_d)}$$

1+d+d²/2 terms

Computational Cost

Example: Mapping with linear and quadratic terms

$$\boldsymbol{x} = (x_1, \ldots, x_d)$$

$$\phi(\boldsymbol{x}) = (1, x_1, \ldots, x_d, x_1 x_1, x_1 x_2, \ldots x_d x_d)$$

Polynomial	ϕ (**x**)	Cost	100 features
Quadratic	> $d^2/2$ terms up to degree 2	$d^2 N^2 /4$	2,500 N^2
Cubic	> $d^3/6$ terms up to degree 3	$d^3 N^2 /12$	83,000 N^2
Quartic	> $d^4/24$ terms up to degree 4	$d^4 N^2 /48$	1,960,000 N^2

Kernel Trick

Define a kernel function such that

$$k(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x})^\top \phi(\boldsymbol{x}')$$

k can be cheaper to evaluate than ϕ!

Kernel Trick

Define a kernel function such that

$$k(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x})^\top \phi(\boldsymbol{x}')$$

k can be cheaper to evaluate than ϕ!

$$\begin{aligned}
\boldsymbol{x} &= (x_1, x_2) \\
k(\boldsymbol{x}, \boldsymbol{x}') &= (1 + \boldsymbol{x}^\top \boldsymbol{x}')^2 \\
&= 1 + x_1^2 {x'}_1^2 + x_2^2 {x'}_2^2 + 2 x_1 {x'}_1 + 2 x_2 {x'}_2 + 2 x_1 {x'}_1 x_2 {x'}_2 \\
\phi(\boldsymbol{x}) &= (1, x_1^2, x_2^2, \sqrt{2} x_1, \sqrt{2} x_2, \sqrt{2} x_1 x_2)
\end{aligned}$$

Computational Cost

Kernel for polynomials up to degree q

$$\boldsymbol{x} = (x_1, \ldots, x_d)$$

$$k(\boldsymbol{x}, \boldsymbol{x}') = (1 + \boldsymbol{x}^\top \boldsymbol{x}')^q$$

Computational Cost

Kernel for polynomials up to degree q

$$\boldsymbol{x} = (x_1, \ldots, x_d)$$

$$k(\boldsymbol{x}, \boldsymbol{x}') = (1 + \boldsymbol{x}^\top \boldsymbol{x}')^q$$

Polynomial	$\phi(\mathbf{x})$	Cost	100 features
Quadratic	> $d^2/2$ terms up to degree 2	$d^2 N^2/4$	2,500 N^2
Cubic	> $d^3/6$ terms up to degree 3	$d^3 N^2/12$	83,000 N^2
Quartic	> $d^4/24$ terms up to degree 4	$d^4 N^2/48$	1,960,000 N^2

Computational Cost

Kernel for polynomials up to degree q

$$\boldsymbol{x} = (x_1, \ldots, x_d)$$

$$k(\boldsymbol{x}, \boldsymbol{x}') = (1 + \boldsymbol{x}^\top \boldsymbol{x}')^q$$

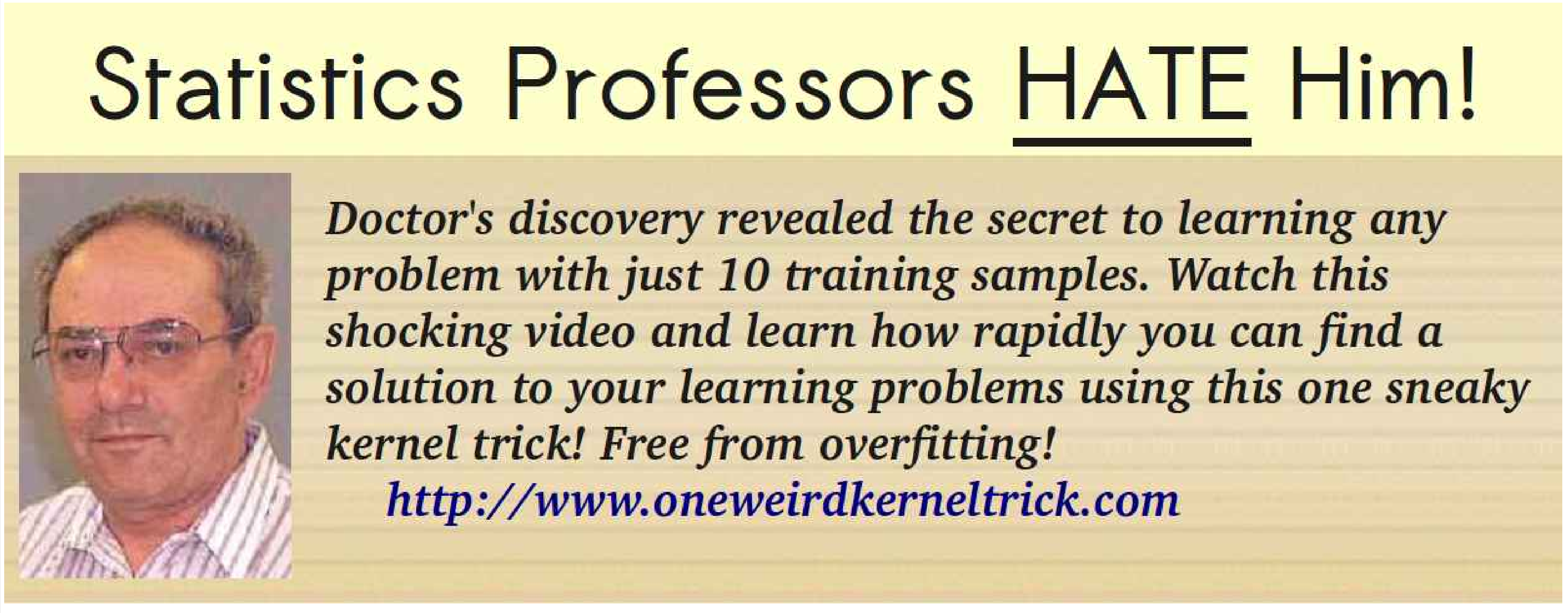
Polynomial	ϕ (**x**)	Cost	100 features
Quadratic	> $d^2/2$ terms up to degree 2	$d^{\cancel{2}} N^2 / \cancel{4}$	100 ~~2,500~~ N^2
Cubic	> $d^3/6$ terms up to degree 3	$d^{\cancel{3}} N^2 / \cancel{12}$	100 ~~83,000~~ N^2
Quartic	> $d^4/24$ terms up to degree 4	$d^{\cancel{4}} N^2 / \cancel{48}$	100 ~~1,960,000~~ N^2

Computational Cost

Kernel for polynomials up to degree q

$$\boldsymbol{x} = (x_1, \ldots, x_d)$$

$$k(\boldsymbol{x}, \boldsymbol{x}') = (1 + \boldsymbol{x}^\top \boldsymbol{x}')^q$$



Kernels

Borrowing from:
Arthur Gretton
(Gatsby, UCL)

Hilbert Spaces

Definition (Inner product)

Let $\mathcal{H}$ be a vector space over $\mathbb{R}$. A function $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is an inner product on $\mathcal{H}$ if

1. Linear: $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{H}}$
2. Symmetric: $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$
3. $\langle f, f \rangle_{\mathcal{H}} \geq 0$ and $\langle f, f \rangle_{\mathcal{H}} = 0$ if and only if $f = 0$.

Norm induced by the inner product: $\|f\|_{\mathcal{H}} := \sqrt{\langle f, f \rangle_{\mathcal{H}}}$

Definition (Hilbert space)

Inner product space containing Cauchy sequence limits.

Example: Fourier Bases

$$\langle f, f' \rangle := \int_{-\infty}^{\infty} dx\, f(x)^* f'(x)$$

Example: Fourier Bases

$$\langle f, f' \rangle := \int_{-\infty}^{\infty} dx\, f(x)^* f'(x)$$

$$f := e^{i\omega x}$$

$$f' := e^{i\omega' x}$$

Example: Fourier Bases

$$\langle f, f' \rangle := \int_{-\infty}^{\infty} dx\, f(x)^* f'(x)$$

$$f := e^{i\omega x}$$

$$f' := e^{i\omega' x}$$

$$\langle f, f' \rangle := \int_{-\infty}^{\infty} dx\ \exp^{i(\omega' - \omega)x}$$

$$= \delta(\omega, \omega')$$

Example: Fourier Bases

$$\langle f, f' \rangle := \int_{-\infty}^{\infty} dx\, f(x)^* f'(x)$$

$$f := e^{i\omega x}$$

$$f' := e^{i\omega' x}$$

$$\langle f, f' \rangle := \int_{-\infty}^{\infty} dx \; \exp^{i(\omega' - \omega)x}$$

$$= \delta(\omega, \omega')$$

Fourier modes define a vector space

Kernels

Definition

Let $\mathcal{X}$ be a non-empty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **kernel** if there exists an $\mathbb{R}$-Hilbert space and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, x' \in \mathcal{X}$,

$$k(x, x') := \langle \phi(x), \phi(x') \rangle_{\mathcal{H}} .$$

- Almost no conditions on $\mathcal{X}$ (eg, $\mathcal{X}$ itself doesn't need an inner product, eg. documents).
- A single kernel can correspond to several possible features. A trivial example for $\mathcal{X} := \mathbb{R}$:

$$\phi_1(x) = x \qquad \text{and} \qquad \phi_2(x) = \begin{bmatrix} x/\sqrt{2} \\ x/\sqrt{2} \end{bmatrix}$$

Sums, Transformations, Products

Theorem (Sums of kernels are kernels)

Given $\alpha > 0$ and k, k_1 and k_2 all kernels on $\mathcal{X}$, then αk and $k_1 + k_2$ are kernels on $\mathcal{X}$.

(Proof via positive definiteness: later!) A difference of kernels may not be a kernel (**why?**)

Theorem (Mappings between spaces)

Let $\mathcal{X}$ and $\widetilde{\mathcal{X}}$ be sets, and define a map $A : \mathcal{X} \to \widetilde{\mathcal{X}}$. Define the kernel k on $\widetilde{\mathcal{X}}$. Then the kernel $k(A(x), A(x'))$ is a kernel on $\mathcal{X}$.

Example: $k(x, x') = x^2 \, (x')^2$.

Theorem (Products of kernels are kernels)

Given k_1 on $\mathcal{X}_1$ and k_2 on $\mathcal{X}_2$, then $k_1 \times k_2$ is a kernel on $\mathcal{X}_1 \times \mathcal{X}_2$. If $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}$, then $k := k_1 \times k_2$ is a kernel on $\mathcal{X}$.

Polynomial Kernels

Theorem (Polynomial kernels)

Let $x, x' \in \mathbb{R}^d$ for $d \geq 1$, and let $m \geq 1$ be an integer and $c \geq 0$ be a positive real. Then

$$k(x, x') := \left(\langle x, x' \rangle + c\right)^m$$

is a valid kernel.

To prove: expand into a sum (with non-negative scalars) of kernels $\langle x, x' \rangle$ raised to integer powers. These individual terms are valid kernels by the product rule.

Infinite Sequences

Definition

The space ℓ_2 (**square** summable sequences) comprises all sequences $a := (a_i)_{i \geq 1}$ for which

$$\|a\|_{\ell_2}^2 = \sum_{i=1}^{\infty} a_i^2 < \infty.$$

Definition

Given sequence of functions $(\phi_i(x))_{i \geq 1}$ in ℓ_2 where $\phi_i : \mathcal{X} \to \mathbb{R}$ is the ith coordinate of $\phi(x)$. Then

$$k(x, x') := \sum_{i=1}^{\infty} \phi_i(x) \phi_i(x') \tag{1}$$

Infinite Sequences

Why square summable? By Cauchy-Schwarz,

$$\left|\sum_{i=1}^{\infty} \phi_i(x)\phi_i(x')\right| \leq \|\phi(x)\|_{\ell_2} \left\|\phi(x')\right\|_{\ell_2},$$

so the sequence defining the inner product converges for all $x, x' \in \mathcal{X}$

Taylor Series Kernels

Definition (Taylor series kernel)

For $r \in (0, \infty]$, with $a_n \geq 0$ for all $n \geq 0$

$$f(z) = \sum_{n=0}^{\infty} a_n z^n \qquad |z| < r, \ z \in \mathbb{R},$$

Define $\mathcal{X}$ to be the $\sqrt{r}$-ball in $\mathbb{R}^d$, so$\|x\| < \sqrt{r}$,

$$k(x, x') = f\left(\langle x, x' \rangle\right) = \sum_{n=0}^{\infty} a_n \langle x, x' \rangle^n .$$

Example (Exponential kernel)

$$k(x, x') := \exp\left(\langle x, x' \rangle\right) .$$

Gaussian Kernel

(also known as Radial Basis Function (RBF) kernel)

Example (Gaussian kernel)

The Gaussian kernel on $\mathbb{R}^d$ is defined as

$$k(x, x') := \exp\left(-\gamma^{-2} \left\|x - x'\right\|^2\right).$$

Proof: an exercise! Use product rule, mapping rule, exponential kernel.

Gaussian Kernel

(also known as Radial Basis Function (RBF) kernel)

Example (Gaussian kernel)

The Gaussian kernel on $\mathbb{R}^d$ is defined as

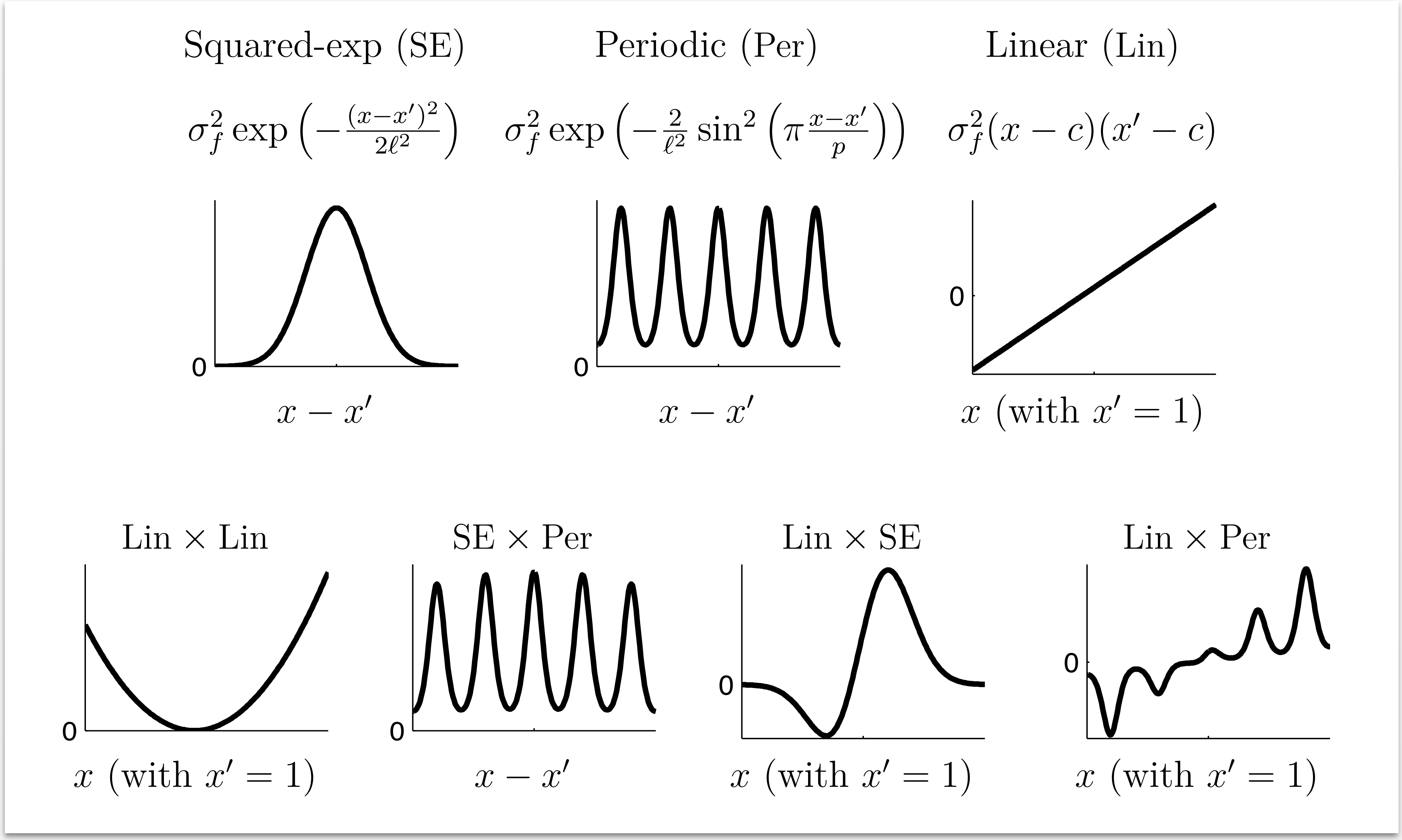
$$k(x, x') := \exp\left(-\gamma^{-2} \|x - x'\|^2\right).$$

Proof: an exercise! Use product rule, mapping rule, exponential kernel.

Squared Exponential (SE) $\quad k(\boldsymbol{x}, \boldsymbol{x}') = \exp^{-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x}'}$

Automatic Relevance Determination (ARD) $\quad k(\boldsymbol{x}, \boldsymbol{x}') = \exp^{-\frac{1}{2}\sum_{i=1}^{d} \frac{(x_i - x_i')^2}{\sigma_i^2}}$

Products of Kernels



source: David Duvenaud (PhD Thesis)

Positive Definiteness

Definition (Positive definite functions)

A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite if $\forall n \geq 1$, $\forall (a_1, \dots a_n) \in \mathbb{R}^n$, $\forall (x_1, \dots, x_n) \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0.$$

The function $k(\cdot, \cdot)$ is strictly positive definite if for mutually distinct x_i, the equality holds only when all the a_i are zero.

Mercer's Theorem

Theorem

Let $\mathcal{H}$ be a Hilbert space, $\mathcal{X}$ a non-empty set and $\phi : \mathcal{X} \to \mathcal{H}$. Then $\langle \phi(x), \phi(y) \rangle_{\mathcal{H}} =: k(x, y)$ is positive definite.

Proof.

$$\begin{aligned}
\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) &= \sum_{i=1}^{n} \sum_{j=1}^{n} \langle a_i \phi(x_i), a_j \phi(x_j) \rangle_{\mathcal{H}} \\
&= \left\| \sum_{i=1}^{n} a_i \phi(x_i) \right\|_{\mathcal{H}}^{2} \geq 0.
\end{aligned}$$

Reverse also holds: positive definite $k(x, x')$ is inner product in a unique $\mathcal{H}$ (Moore-Aronsajn: coming later!). $\square$

Kernelized SVMs

Dual problem

$$W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j \boldsymbol{x}_i^\top \boldsymbol{x}_j$$

Dual problem with ***feature map***

$$W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j \phi(\boldsymbol{x}_i)^\top \phi(\boldsymbol{x}_j)$$

Kernelized SVMs

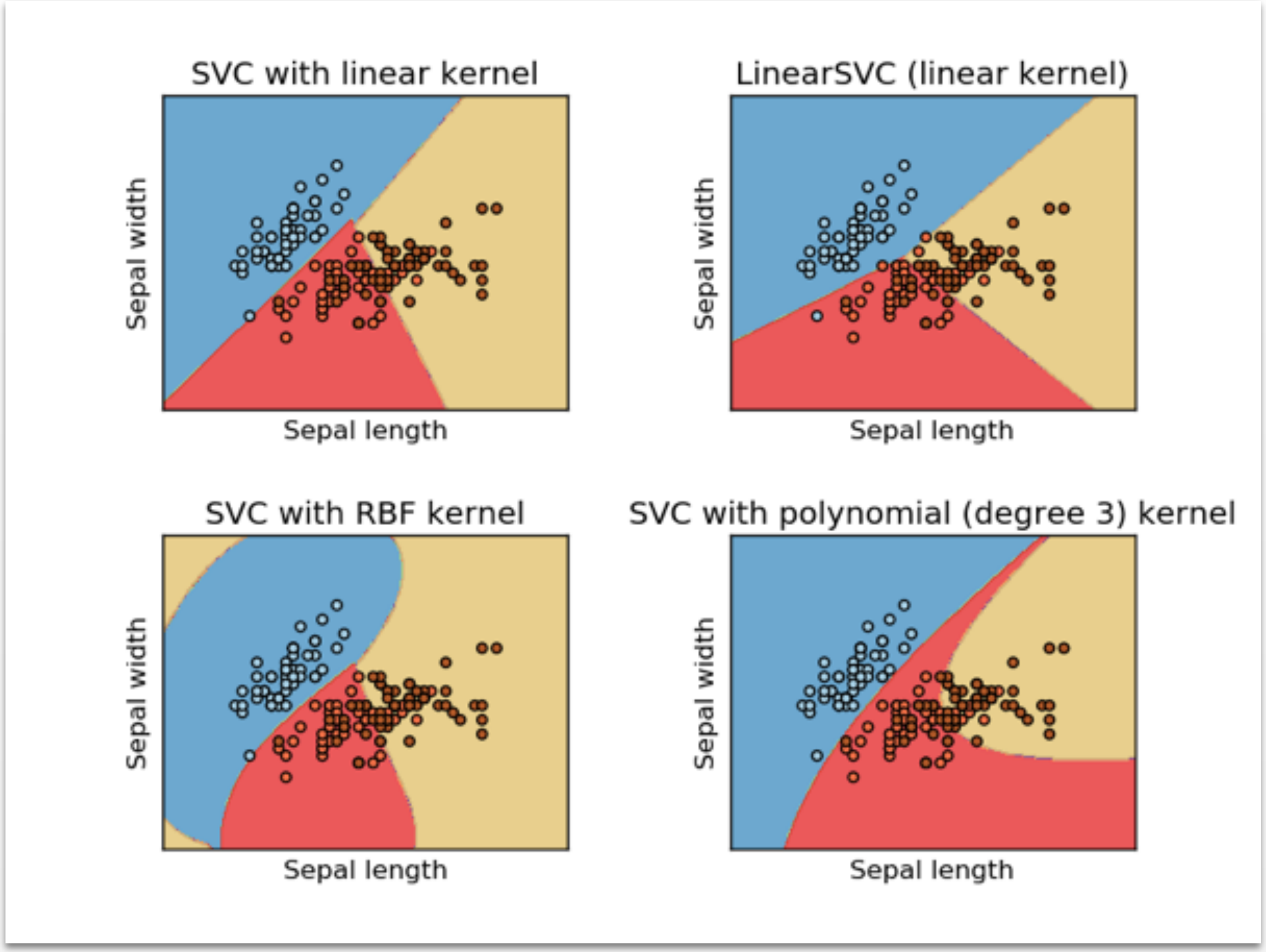
Dual problem

$$W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j \boldsymbol{x}_i^\top \boldsymbol{x}_j$$

Dual problem with ***kernel***

$$W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j k(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

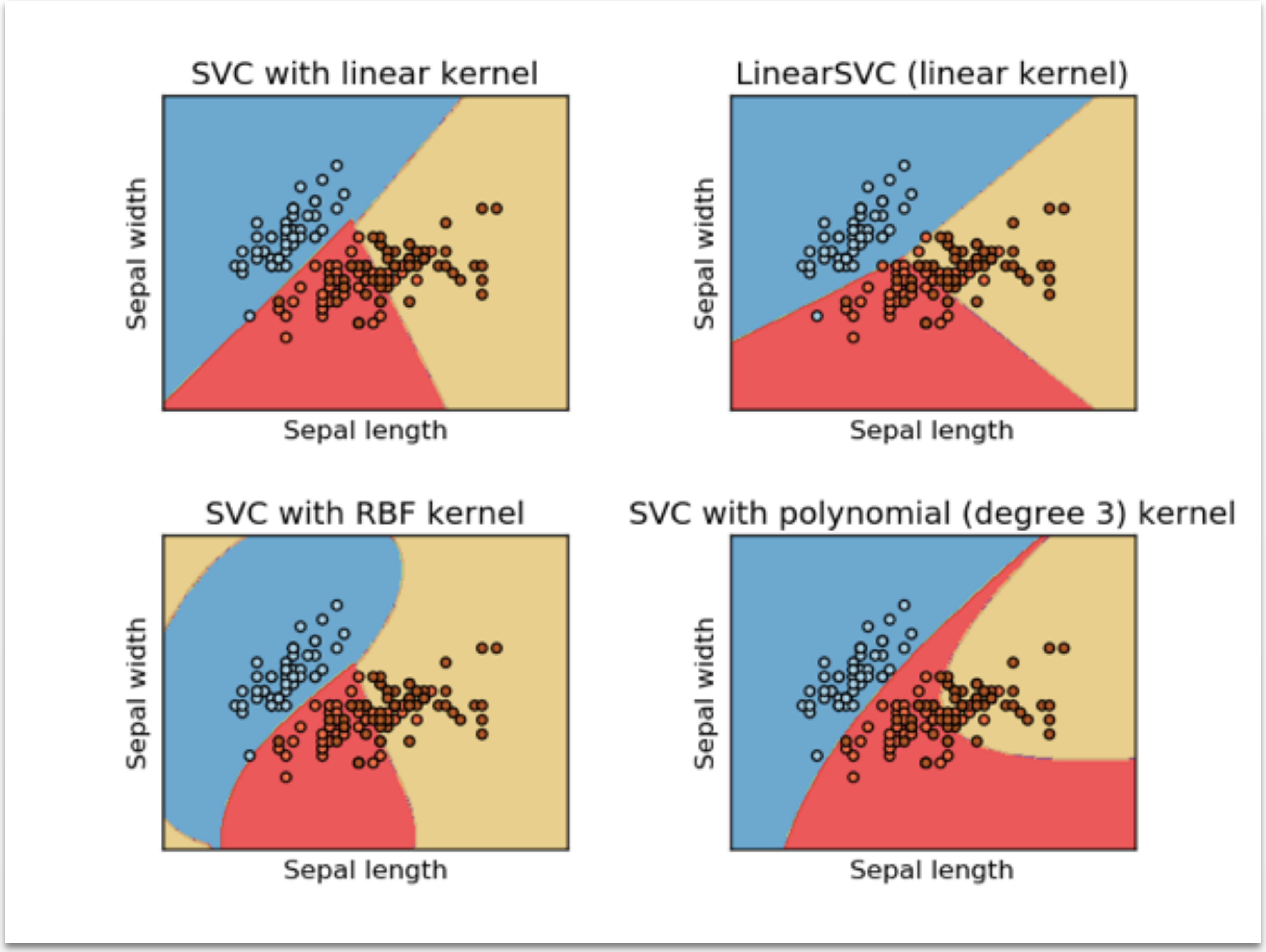
Kernelized SVMs



Generalization to multiple classes:

Train multiple **one-vs-all** or **one-vs-one** classifiers

Kernelized SVMs

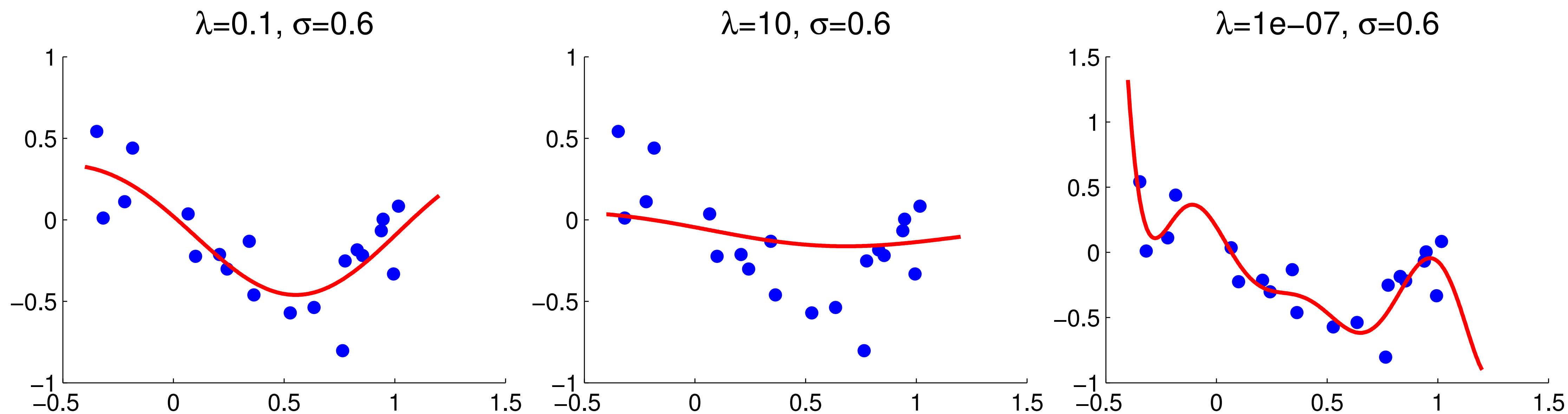


Generalization to multiple classes:

Train multiple **one-vs-all** or **one-vs-one** classifiers

Kernel Ridge Regression

$$f^* = \arg\min_{f \in \mathcal{H}} \left(\sum_{i=1}^{n} (y_i - \langle f, \phi(x_i) \rangle_{\mathcal{H}})^2 + \lambda \|f\|_{\mathcal{H}}^2 \right).$$



Optimization Problem

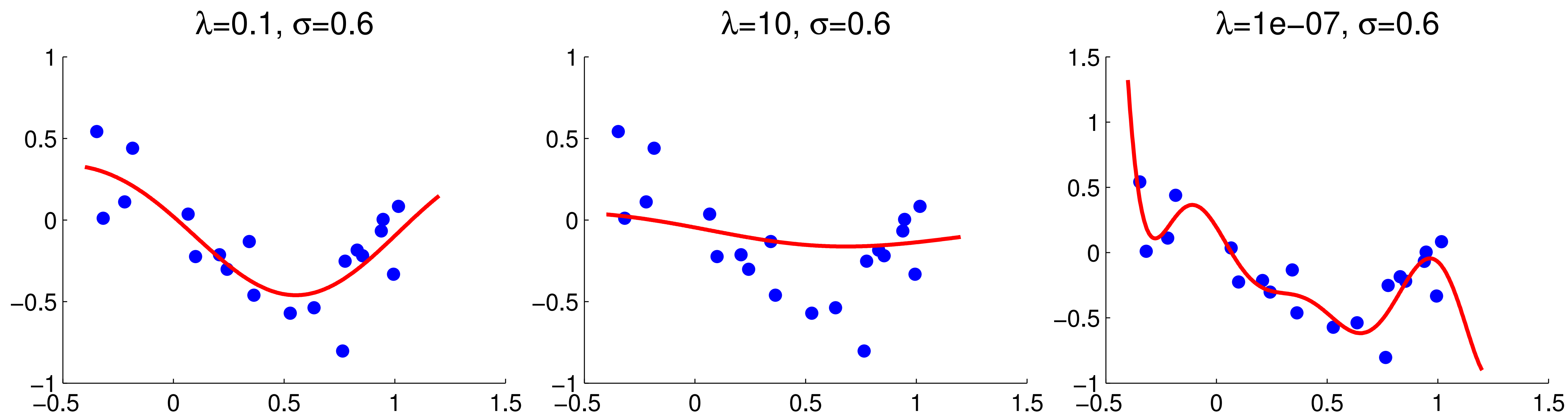
$$\min \lambda \|w\|^2 + \sum \xi_i^2$$
$$\text{s.t. } \xi_i = y_i - \langle w, x_i \rangle$$

Solve for Dual Problem

$$w = \frac{1}{2\lambda} \sum \alpha_i x_i$$
$$\xi = \frac{\alpha_i}{2}$$

Kernel Ridge Regression

$$f^* = \arg\min_{f \in \mathcal{H}} \left(\sum_{i=1}^{n} (y_i - \langle f, \phi(x_i) \rangle_{\mathcal{H}})^2 + \lambda \|f\|_{\mathcal{H}}^2 \right).$$



Closed form Solution

$$\boldsymbol{\alpha} = 2\lambda(\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}\boldsymbol{y}$$

$$f(\boldsymbol{x}) = \boldsymbol{y}^\top(\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}\boldsymbol{k}$$

$$\boldsymbol{y} := (y_1, \ldots, y_n)$$

$$\boldsymbol{K}_{ij} := k(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

$$\boldsymbol{k}_i(\boldsymbol{x}) := k(\boldsymbol{x}_i, \boldsymbol{x})$$