On SVM Solution and The Kernel Trick

CS189/289A: Introduction to Machine Learning

Stella Yu

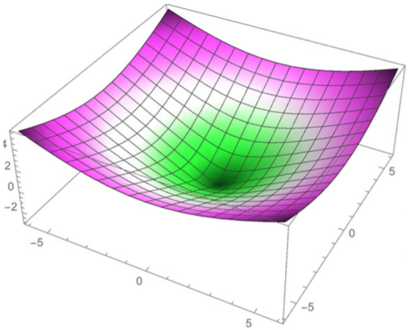
UC Berkeley

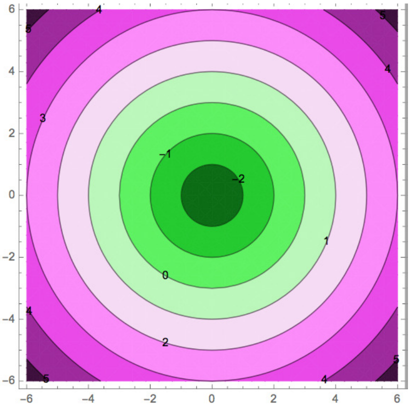
26 October 2017

Outline

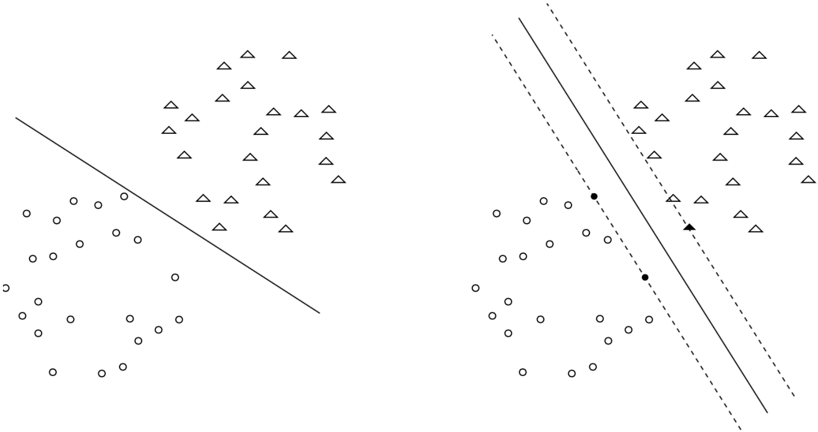
- SVM review
- SVM solution
- SVM and the kernel trick
- SVM for multiclass classification

Decision Function and Decision Boundary

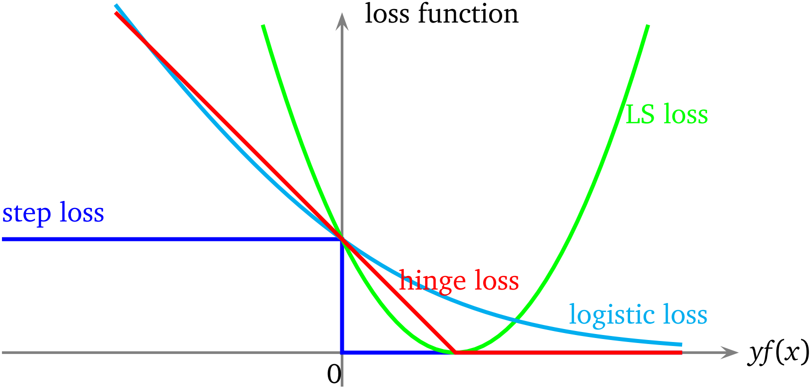




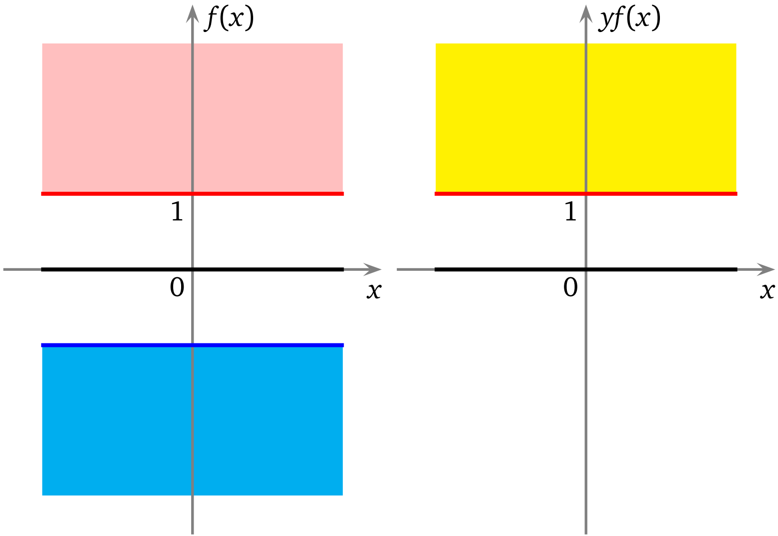
Maximum Margin Principle for the Primal SVM



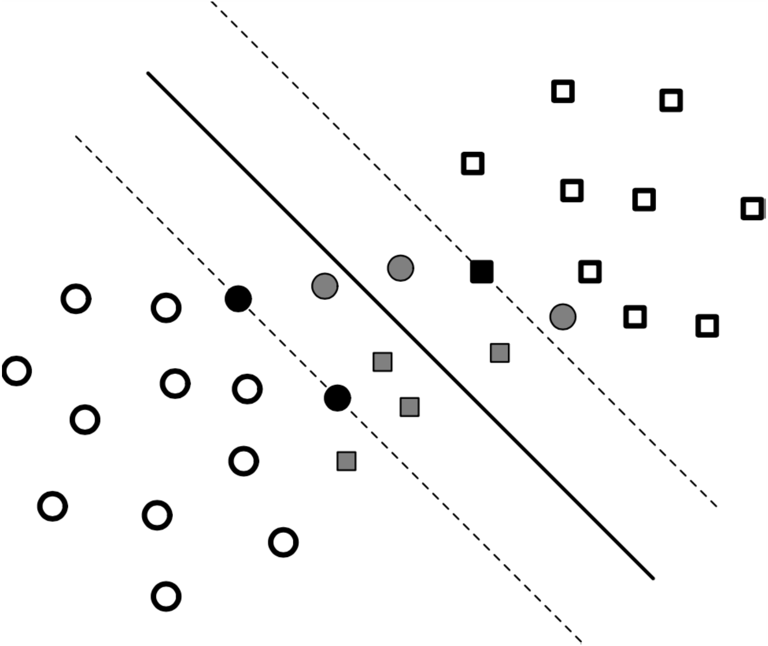
SVM = Tikhonov Regularization with Hinge Loss



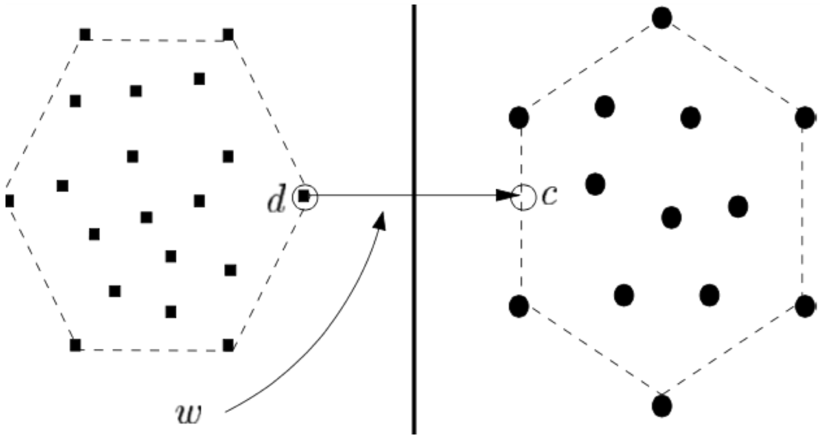
Classification Margin and Slack



Slack and Support Vectors for SVM



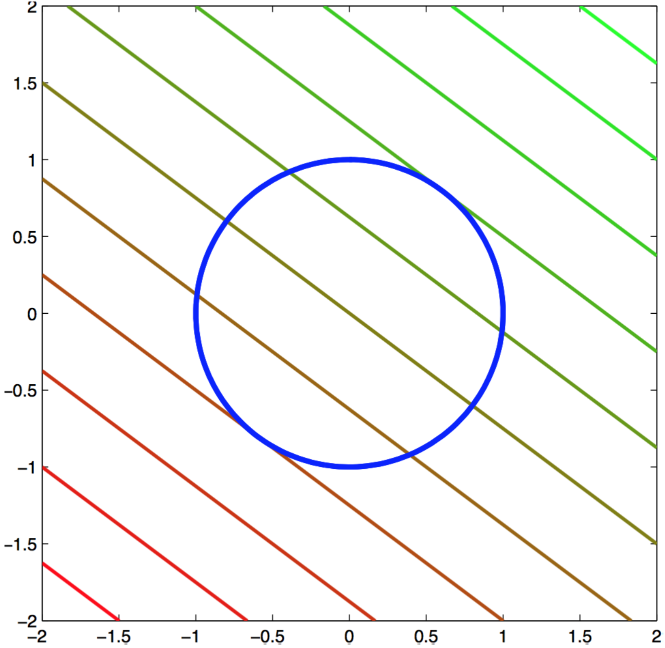
Convex Hull Interpretations from the Dual SVM



Solution: Solve The Dual of A Convex Program

- To every convex program corresponds a dual
- Solving the original (primal) is equivalent to solving the dual
- Dual often provides insight
- Can derive dual by using Lagrange multipliers to eliminate constraints

Lagrange Multiplier: Geometrical Intuition



Constrained Optimization Using the Lagrangian

- For convex functions $f(x), g(x)$,

$$\begin{aligned} \min \quad & f(x) && (1) \\ \text{s. t.} \quad & g(x) \geq 0 && (2) \end{aligned}$$

- Lagrangian:

$$\max_{\lambda \geq 0} \quad \min_{x} \quad L(x;\lambda) = f(x) - \lambda g(x) \tag{3}$$

- Karush-Kuhn Tucker (KKT) optimality conditions:

$$\begin{aligned} \text{stationarity:} \quad & \frac{\partial L}{\partial x} = \frac{\partial f(x)}{\partial x} - \lambda \frac{\partial g(x)}{\partial x} = 0 && (4) \\ \text{primal feasibility:} \quad & g(x) \geq 0 && (5) \\ \text{dual feasibility:} \quad & \lambda \geq 0 && (6) \\ \text{complementary slackness:} \quad & \lambda g(x) = 0 && (7) \end{aligned}$$

SVM Dual by Using Lagrange Multipliers

$$\min_{w,t,\xi} \quad \varepsilon(w,t,\xi) = \frac{1}{2}|w|^2 + C\sum_{i=1}^{n}\xi_i \tag{8}$$

$$\text{s. t.} \quad y_i(w\cdot x_i - t) \geq 1 - \xi_i, \qquad i = 1,2,\dots,n \tag{9}$$

$$\xi_i \geq 0, \qquad i = 1,2,\dots,n \tag{10}$$

$$L(w,t,\xi_1,\dots,\xi_n,\alpha_1,\dots,\alpha_n,\beta_1,\dots,\beta_n)$$

$$= \frac{1}{2}|w|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i(y_i(w\cdot x_i - t) - (1 - \xi_i)) - \sum_{i=1}^{n}\beta_i\xi_i \tag{11}$$

$$= \frac{1}{2}|w|^2 - \sum_{i=1}^{n}\alpha_i y_i(w\cdot x_i - t) + \sum_{i=1}^{n}\alpha_i + \sum_{i=1}^{n}(C - \alpha_i - \beta_i)\xi_i \tag{12}$$

SVM Solution: KKT Stationarity Conditions

- Lagrangian with one dual variable for each primal constraint:

$$L(w,t,\xi_1,\dots,\xi_n,\alpha_1,\dots,\alpha_n,\beta_1,\dots,\beta_n)$$
$$=\frac{1}{2}|w|^2-\sum_{i=1}^{n}\alpha_i y_i(w\cdot x_i-t)+\sum_{i=1}^{n}\alpha_i+\sum_{i=1}^{n}(C-\alpha_i-\beta_i)\xi_i \quad (13)$$

- KKT stationarity conditions:

$$\frac{\partial L}{\partial w}=w-\sum_{i=1}^{n}\alpha_i y_i x_i=0 \qquad \Rightarrow w=\sum_{i=1}^{n}\alpha_i y_i x_i \quad (14)$$
$$\frac{\partial L}{\partial t}=\sum_{i=1}^{n}\alpha_i y_i=0 \quad (15)$$
$$\frac{\partial L}{\partial \xi_i}=C-\alpha_i-\beta_i=0 \qquad \Rightarrow 0\le\alpha_i\le C \quad (16)$$

Simplification of the Lagrangian to Function Of α

$$
\begin{aligned}
& L(w, t, \xi_1, \ldots, \xi_n, \alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_n) \\
& = \frac{1}{2}|w|^2 - \sum_{i=1}^{n} \alpha_i y_i (x_i \cdot w - t) + \sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n} (C - \alpha_i - \beta_i)\xi_i && (17) \\
& = \frac{1}{2}|w|^2 - |w|^2 + \left(\sum_{i=1}^{n} \alpha_i y_i \right) t + \sum_{i=1}^{n} \alpha_i && (18) \\
& = -\frac{1}{2} \sum_{i=1}^{n} \alpha_i y_i x_i \cdot \left(\sum_{j=1}^{n} \alpha_j y_j x_j \right) + \sum_{i=1}^{n} \alpha_i && (19) \\
& = \alpha' 1 - \frac{1}{2} \alpha' G \alpha && (20) \\
& G_{ij} = y_i (x_i \cdot x_j) y_j \quad \Leftarrow \text{Gram Matrix} && (21)
\end{aligned}
$$

SVM Dual: Much Simpler Than The SVM Primal

$$\text{SVM Dual:} \tag{22}$$

$$\max_{\alpha} \quad L(\alpha) = \alpha' 1 - \frac{1}{2}\alpha' G \alpha \tag{23}$$

$$\text{where} \quad G_{ij} = y_i (x_i \cdot x_j) y_j \tag{24}$$

$$\text{s. t.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0, \quad 0 \le \alpha_i \le C, i = 1, \ldots, n. \tag{25}$$

$$\text{SVM Primal:} \tag{26}$$

$$\min_{w,t,\xi} \quad \varepsilon(w,t,\xi) = \frac{1}{2}|w|^2 + C\sum_{i=1}^{n} \xi_i \tag{27}$$

$$\text{s. t.} \quad y_i(w \cdot x_i - t) \ge 1 - \xi_i, \tag{28}$$

$$\xi_i \ge 0, \quad i = 1, 2, \ldots, n \tag{29}$$

KKT Feasibility and Complementary Conditions

- Primal feasibility:

$$y_i(w \cdot x_i - t) \geq 1 - \xi_i, \qquad i = 1, 2, \ldots, n \tag{30}$$

$$\xi_i \geq 0, \qquad i = 1, 2, \ldots, n \tag{31}$$

- Dual feasibility:

$$\alpha_i \geq 0, \qquad i = 1, 2, \ldots, n \tag{32}$$

$$\beta_i \geq 0, \qquad i = 1, 2, \ldots, n \tag{33}$$

- Complementary slackness relating each multiplier to its constraint:

$$\alpha_i \cdot (y_i(w \cdot x_i - t) - (1 - \xi_i)) = 0, \qquad i = 1, 2, \ldots, n \tag{34}$$

$$\beta_i \cdot \xi_i = 0, \qquad i = 1, 2, \ldots, n \tag{35}$$

Three Cases for The Training Instances

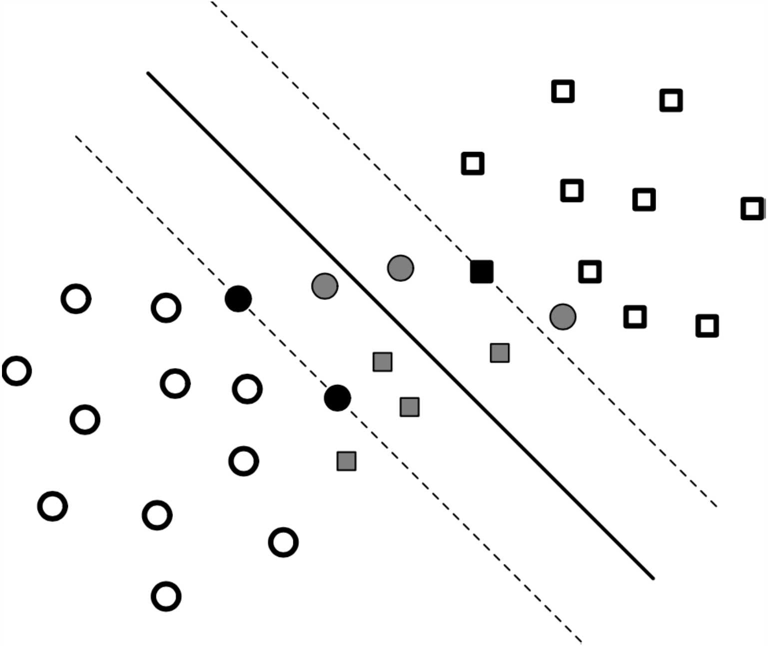
- Instances with $\alpha_i > 0$ are support vectors which participate in spanning the decision boundary.

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i = \sum_{\alpha_i > 0} \alpha_i y_i x_i \quad (36)$$

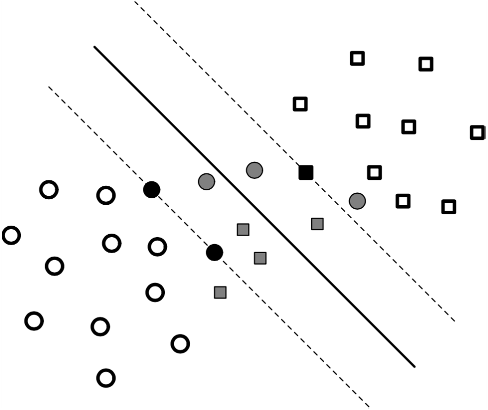
- Significance of the upperbound C on the α multipliers:
  1. $\alpha_i = 0$: These instances are outside or on the margin.
     Since $C - \alpha_i - \beta_i = 0$, $\alpha_i = 0$ implies $\beta_i = C$. Consequently, $\xi_i = 0$.
  2. $\alpha_i = C$: These are support vectors on or inside the margin.
     Since $C - \alpha_i - \beta_i = 0$, $\alpha_i = C$ implies $\beta_i = 0$ and $y_i f(x_i) = 1 - \xi_i$.
     Consequently, $\xi_i = 1 - y_i f(x_i) \geq 0$.
  3. $0 < \alpha_i < C$: These are the support vectors on the margin.
     Since $C - \alpha_i - \beta_i = 0$, $0 < \alpha_i < C$ implies $\beta_i = C - \alpha_i > 0$.
     Consequently, $\xi_i = 0$, and also:

$$y_i(w \cdot x_i - t) = 1 \quad \Rightarrow \quad t = w \cdot x_i - y_i \quad (37)$$

Slack and Support Vectors for SVM



Support Vectors and Their Dual Variables α



$$\alpha_i = 0 \Rightarrow \quad y_i f(x_i) \geq 1\text{: on or outside the margin} \tag{38}$$

$$0 < \alpha_i < C \Rightarrow \quad y_i f(x_i) = 1\text{: on the margin} \tag{39}$$

$$\alpha_i = C \Rightarrow \quad y_i f(x_i) \leq 1\text{: on or inside the margin} \tag{40}$$

$$\alpha_i = 0 \Leftarrow \quad y_i f(x_i) > 1\text{: outside the margin} \tag{41}$$

$$\alpha_i = C \Leftarrow \quad y_i f(x_i) < 1\text{: inside the margin} \tag{42}$$

SVM Solution: From the Dual to the Primal

- Solve α in the SVM dual:

$$\max_{\alpha} \quad L(\alpha) = \alpha' 1 - \frac{1}{2}\alpha' G \alpha \tag{43}$$

$$\text{where} \quad G_{ij} = y_i (x_i \cdot x_j) y_j \tag{44}$$

$$\text{s. t.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0, \quad 0 \le \alpha_i \le C, i = 1, \ldots, n. \tag{45}$$

- Solve w, t in the SVM Primal:

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i \tag{46}$$

$$t = \text{mean}(\{w \cdot x_i - y_i : \quad 0 < \alpha_i < C\}) \tag{47}$$

$$\xi_i = \begin{cases} 1 - y_i(w \cdot x_i - t), & \alpha_i = C \\ 0, & \text{otherwise} \end{cases} \tag{48}$$

The Kernel Trick

$$\text{SVM dual: } \max_{\alpha} \quad L(\alpha) = \alpha' 1 - \frac{1}{2}\alpha' G \alpha \tag{49}$$

$$\text{where} \quad G_{ij} = y_i \, (x_i \cdot x_j) \, y_j \tag{50}$$

$$\text{s. t.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, n. \tag{51}$$

$$\text{decision function:} \quad f(x) = w \cdot x - t = \left(\sum_{i=1}^{n} \alpha_i y_i x_i \right) \cdot x - t \tag{52}$$

$$= \sum_{i=1}^{n} \alpha_i y_i \, (x_i \cdot x) - t \tag{53}$$

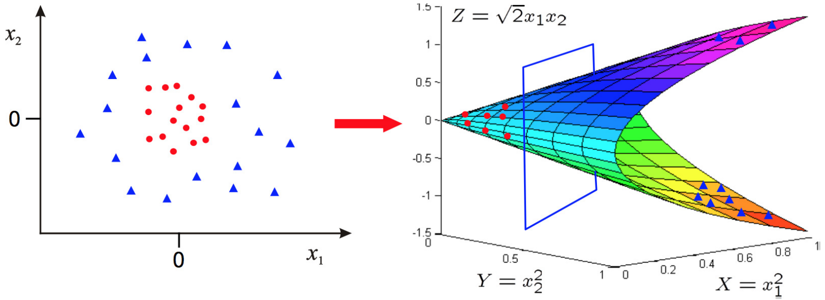
All we need is some measure of pairwise similarity (a scalar) between features: $x \cdot z$, not the features x and z themselves:

$$K(x, z) = x \cdot z \quad \Leftarrow \text{kernel function} \tag{54}$$

Linear SVM vs. Kernel SVM in General

	Linear SVM	**Kernel SVM**
training	$x_i, \quad i = 1, \ldots, n$	$K(x_i, x_j), \quad i, j = 1, \ldots, n$
testing	w	$K(x, x_i), \quad \alpha_i > 0$
memory	1 normal vector	#? support vectors
decision	feature space	sample space
boundary	linear	nonlinear

Not Linearly Separable: Feature Transformation



$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \rightarrow \phi(x) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \end{bmatrix} \tag{55}$$

$$\phi(x) \cdot \phi(z) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \end{bmatrix}' \begin{bmatrix} z_1^2 \\ z_2^2 \\ \sqrt{2}z_1z_2 \end{bmatrix} \tag{56}$$

$$= (x_1z_1)^2 + (x_2z_2)^2 + 2(x_1z_1x_2z_2) = (x'z)^2 \tag{57}$$

$$K(x, z) = (x \cdot z)^2 \tag{58}$$

Commonly Used Kernels

- Linear kernels

$$K(x,z) = x \cdot z \tag{59}$$

- polynomial kernels

$$K(x,z) = (1 + x \cdot z)^d, \quad d > 0 \tag{60}$$

- Sigmod kernels

$$K(x,z) = \tanh(a(x \cdot z) + b) \tag{61}$$

- Gaussian kernels or radial basis kernel

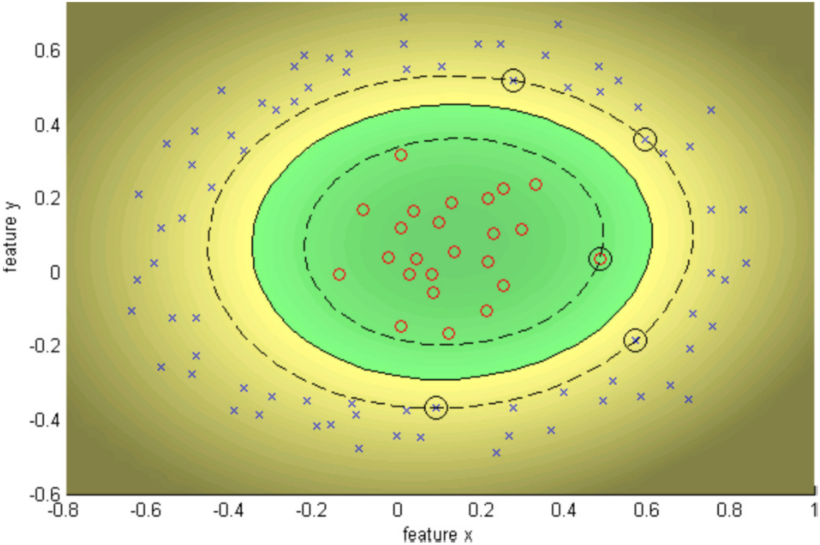
$$K(x,z) = \exp\left(-\frac{(x-z)^2}{2\sigma^2}\right) \tag{62}$$

Infinite dimensional feature space

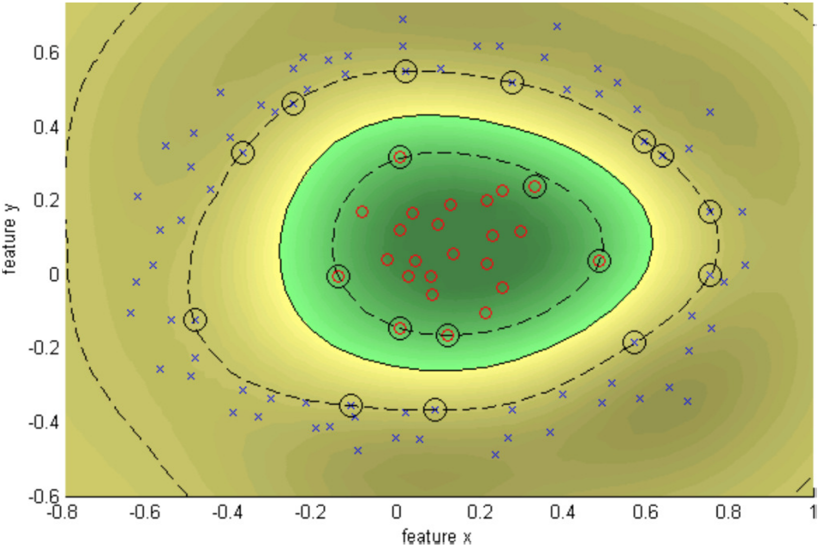
$$\exp\left(-\frac{1}{2}\|\mathbf{x}-\mathbf{x}'\|^2\right) = \sum_{j=0}^{\infty} \frac{(\mathbf{x}^\top \mathbf{x}')^j}{j!} \exp\left(-\frac{1}{2}\|\mathbf{x}\|^2\right) \exp\left(-\frac{1}{2}\|\mathbf{x}'\|^2\right)$$

$$= \sum_{j=0}^{\infty} \sum_{\sum n_i = j} \exp\left(-\frac{1}{2}\|\mathbf{x}\|^2\right) \frac{x_1^{n_1} \cdots x_k^{n_k}}{\sqrt{n_1! \cdots n_k!}} \exp\left(-\frac{1}{2}\|\mathbf{x}'\|^2\right) \frac{{x'}_1^{n_1} \cdots {x'}_k^{n_k}}{\sqrt{n_1! \cdots n_k!}}$$

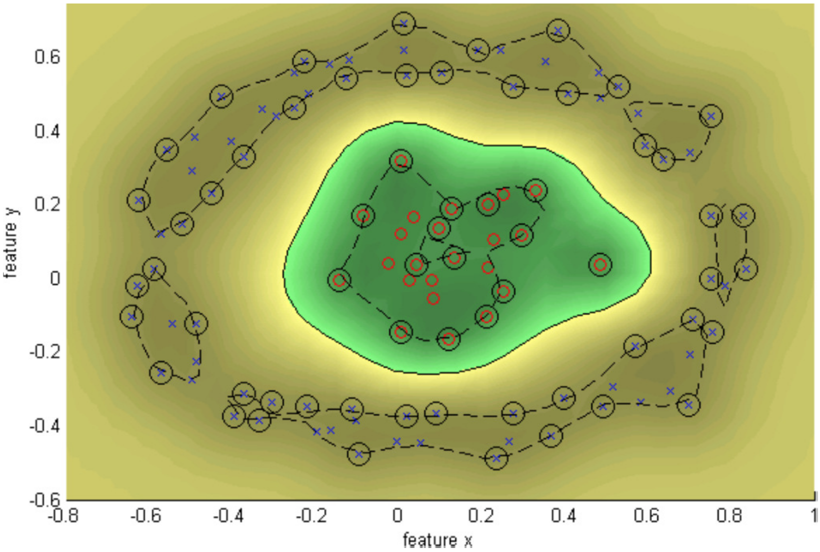
SVM Classifier with Gaussian Kernel: $\sigma = 1$



SVM Classifier with Gaussian Kernel: $\sigma = 0.25$



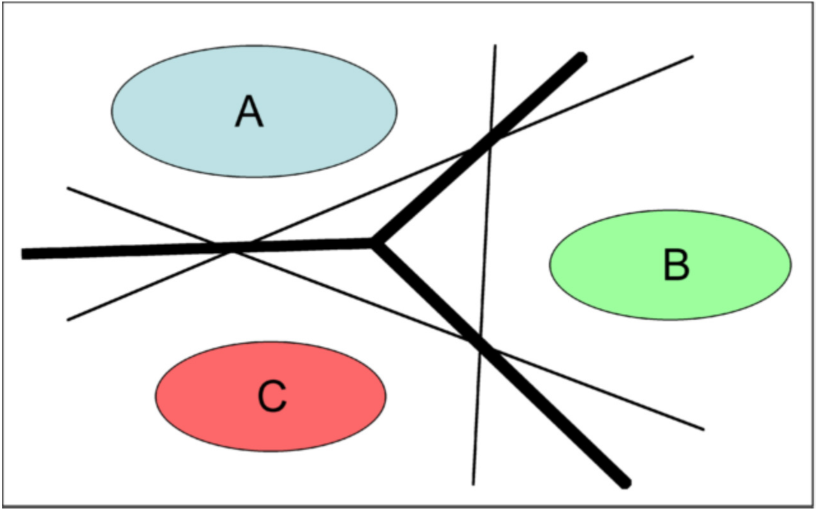
SVM Classifier with Gaussian Kernel: $\sigma = 0.10$



Kernel X

- The kernel trick isn't limited to SVMs.
- Works whenever we can express an algorithm using only sums, dot products of training examples.
- Kernel Fisher discriminant
- Kernel logistic regression
- kernel linear and ridge regression
- kernel SVD or PCA
- $\vdots$

Multiclass Classification: One vs. All



Multiclass Classification: Pairwise