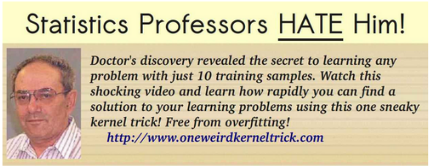
Introduction to Machine Learning

Machine Learning: Jordan Boyd-Graber
University of Maryland

KERNEL SVMS

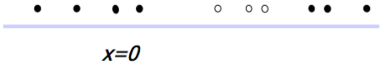


Statistics Professors HATE Him!

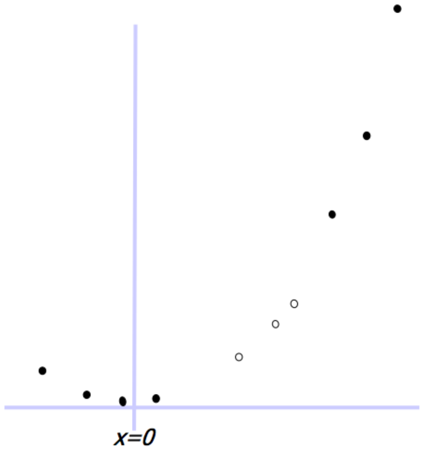
Doctor's discovery revealed the secret to learning any problem with just 10 training samples. Watch this shocking video and learn how rapidly you can find a solution to your learning problems using this one sneaky kernel trick! Free from overfitting!
http://www.oneweirdkerneltrick.com

Slides adapted from Jerry Zhu

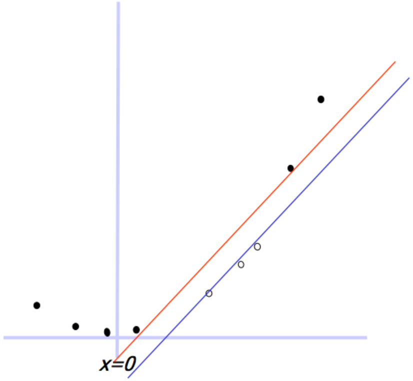
Can you solve this with linear separator?



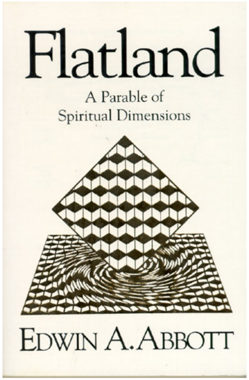
Can you solve this with linear separator?



Can you solve this with linear separator?



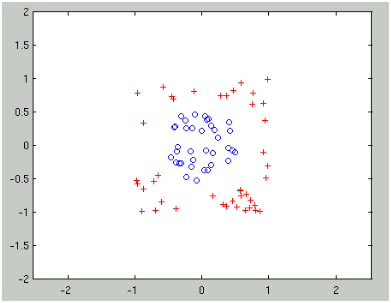
Adding another dimension

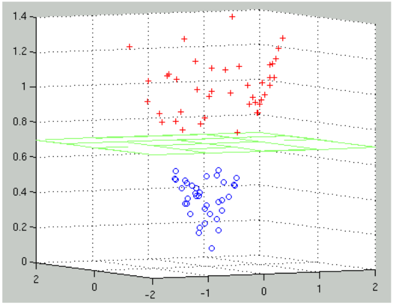


Behold yon miserable creature. That Point is a Being like ourselves, but confined to the non-dimensional Gulf. He is himself his own World, his own Universe; of any other than himself he can form no conception; he knows not Length, nor Breadth, nor Height, for he has had no experience of them; he has no cognizance even of the number Two; nor has he a thought of Plurality, for he is himself his One and All, being really Nothing. Yet mark his perfect self-contentment, and hence learn this lesson, that to be self-contented is to be vile and ignorant, and that to aspire is better than to be blindly and impotently happy.

Problems get easier in higher dimensions

$$(x_1, x_2) \Rightarrow (x_1, x_2, \sqrt{x_1^2 + x_2^2})$$





What's special about SVMs?

$$\max_{\vec{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{i=1}^{m} \alpha_i \alpha_j y_i y_j (\vec{x}_i \cdot \vec{x}_j) \tag{1}$$

What's special about SVMs?

$$\max_{\vec{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{i=1}^{m} \alpha_i \alpha_j y_i y_j (\vec{x}_i \cdot \vec{x}_j) \tag{1}$$

- This dot product is basically just how much x_i looks like x_j. Can we generalize that?

What's special about SVMs?

$$\max_{\vec{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{i=1}^{m} \alpha_i \alpha_j y_i y_j (\vec{x}_i \cdot \vec{x}_j) \tag{1}$$

- This dot product is basically just how much x_i looks like x_j. Can we generalize that?
- Kernels!

What's a kernel?

- A function $K : \mathscr{X} \times \mathscr{X} \mapsto \mathrm{R}$ is a kernel over $\mathscr{X}$.
- This is equivalent to taking the dot product $\langle \phi(x_1), \phi(x_2) \rangle$ for some mapping
- **Mercer's Theorem**: So long as the function is continuous and symmetric, then K admits an expansion of the form

$$K(x, x') = \sum_{n=0}^{\infty} a_n \phi_n(x) \phi_n(x') \tag{2}$$

What's a kernel?

- A function $K : \mathscr{X} \times \mathscr{X} \mapsto \mathrm{R}$ is a kernel over $\mathscr{X}$.
- This is equivalent to taking the dot product $\langle \phi(x_1), \phi(x_2) \rangle$ for some mapping
- **Mercer's Theorem**: So long as the function is continuous and symmetric, then K admits an expansion of the form

$$K(x, x') = \sum_{n=0}^{\infty} a_n \phi_n(x) \phi_n(x') \tag{2}$$

- The computational cost is just in computing the kernel

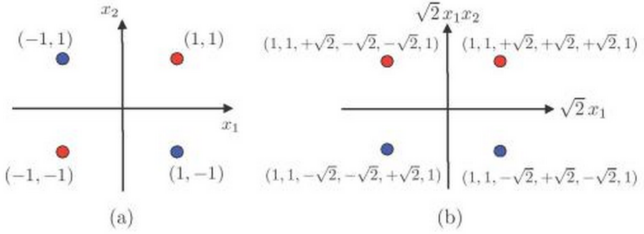
Polynomial Kernel

$$K(x, x') = (x \cdot x' + c)^d \tag{3}$$

Polynomial Kernel

$$K(x, x') = (x \cdot x' + c)^d \tag{3}$$

When $d = 2$:


x_2
$(-1, 1)$
$(1, 1)$
x_1
$(-1, -1)$
$(1, -1)$
(a)
$\sqrt{2}x_1x_2$
$(1, 1, +\sqrt{2}, -\sqrt{2}, -\sqrt{2}, 1)$
$(1, 1, +\sqrt{2}, +\sqrt{2}, +\sqrt{2}, 1)$
$\sqrt{2}x_1$
$(1, 1, -\sqrt{2}, -\sqrt{2}, +\sqrt{2}, 1)$
$(1, 1, -\sqrt{2}, +\sqrt{2}, -\sqrt{2}, 1)$
(b)

Gaussian Kernel

$$K(x, x') = \exp - \frac{\left\| x' - x \right\|^2}{2\sigma^2} \tag{4}$$

Gaussian Kernel

$$K(x, x') = \exp - \frac{\left\| x' - x \right\|^2}{2\sigma^2} \tag{4}$$

which can be rewritten as

$$K(x, x') = \sum_n \frac{(x \cdot x')^n}{\sigma^n n!} \tag{5}$$

(All polynomials!)

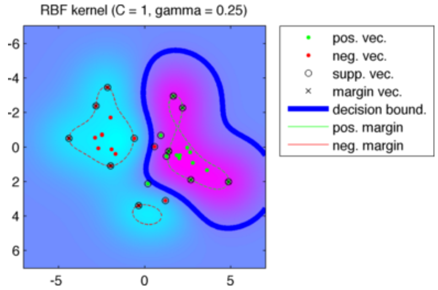
Gaussian Kernel

$$K(x, x') = \exp - \frac{\left\| x' - x \right\|^2}{2\sigma^2} \tag{4}$$

which can be rewritten as

$$K(x, x') = \sum_n \frac{(x \cdot x')^n}{\sigma^n n!} \tag{5}$$

(All polynomials!)

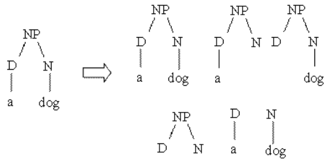


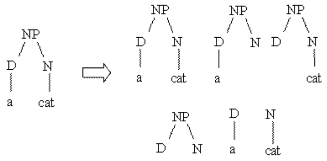
Tree Kernels

- Sometimes we have example *x* that are hard to express as vectors
- For example sentences “a dog” and “a cat”: internal syntax structure

Tree Kernels

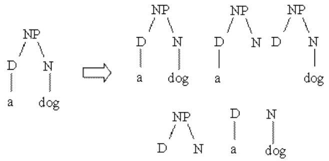
- Sometimes we have example *x* that are hard to express as vectors
- For example sentences “a dog” and “a cat”: internal syntax structure

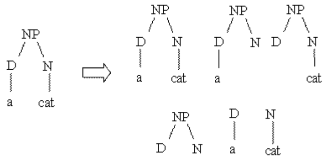




Tree Kernels

- Sometimes we have example *x* that are hard to express as vectors
- For example sentences “a dog” and “a cat”: internal syntax structure





3/5 structures match, so tree kernel returns .6

What does this do to learnability?

- Kernelized hypothesis spaces are obviously more complicated
- What does this do to complexity?

What does this do to learnability?

- Kernelized hypothesis spaces are obviously more complicated
- What does this do to complexity?
- Rademacher complexity for a kernel with radius Λ and data with radius r: $S \subset \{x : K(x,x) \leq r^2\}$, $H = \{x \mapsto w \cdot \phi(x) : ||w|| \leq \Lambda\}$

$$\hat{\mathscr{R}}_S(H) \leq \sqrt{\frac{r^2\Lambda^2}{m}} \tag{6}$$

What does this do to learnability?

- Kernelized hypothesis spaces are obviously more complicated
- What does this do to complexity?
- Rademacher complexity for a kernel with radius Λ and data with radius r: $S \subset \{x : K(x,x) \leq r^2\}$, $H = \{x \mapsto w \cdot \phi(x) : \|w\| \leq \Lambda\}$

$$\hat{\mathscr{R}}_S(H) \leq \sqrt{\frac{r^2\Lambda^2}{m}} \tag{6}$$

- Proof requires real analysis

Margin learnability

- With probability $1-\delta$:

$$R(h) \leq \hat{R}_\rho(h) + 2\sqrt{\frac{r^2\Lambda^2/\rho^2}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \tag{7}$$

Margin learnability

- With probability $1-\delta$:

$$R(h) \leq \hat{R}_\rho(h) + 2\sqrt{\frac{r^2\Lambda^2/\rho^2}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \tag{7}$$

- So if you can find a simple kernel representation that induces a margin, use it!

Margin learnability

- With probability $1-\delta$:

$$R(h) \leq \hat{R}_\rho(h) + 2\sqrt{\frac{r^2\Lambda^2/\rho^2}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \tag{7}$$

- So if you can find a simple kernel representation that induces a margin, use it!
- . . . so long as you can handle the computational complexity

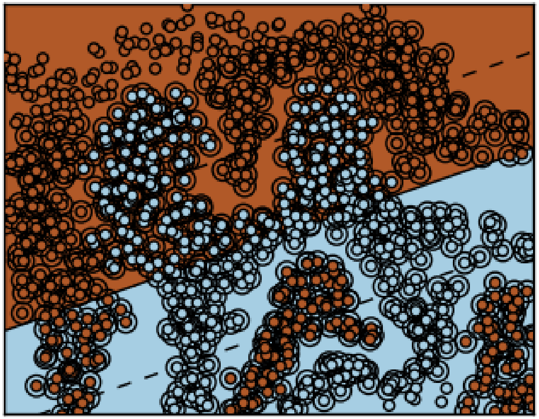
How does it effect optimization

- Replace all dot product with kernel evaluations $K(x_1, x_2)$
- Makes computation more expensive, overall structure is the same
- Try linear first!

Kernelized SVM

```
X, Y = read_data("ex8a.txt")
clf = svm.SVC(kernel=kk, degree=dd, gamma=gg)
clf.fit(X, Y)
```

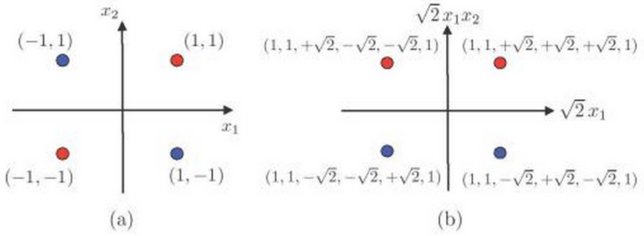
Linear Kernel Doesn't Work



Polynomial Kernel

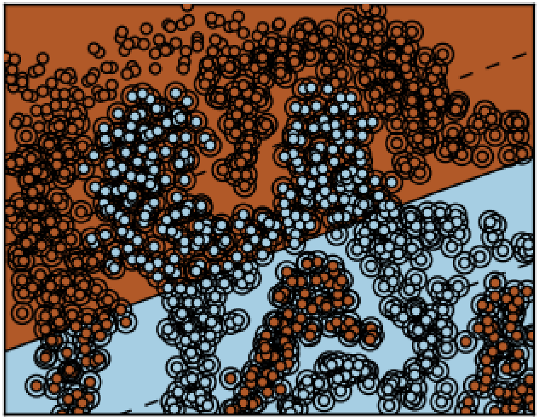
$$K(x, x') = (x \cdot x' + c)^d \tag{8}$$

When $d = 2$:

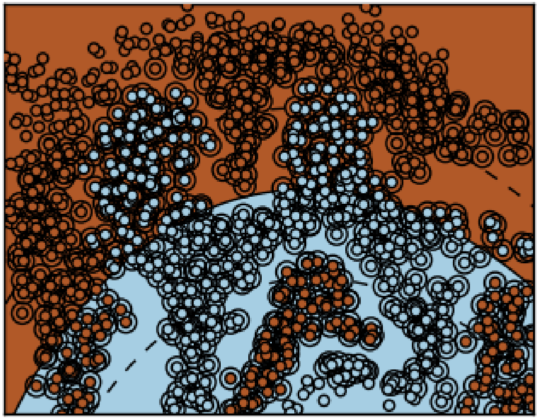


x_2
x_1
$(-1, 1)$
$(1, 1)$
$(-1, -1)$
$(1, -1)$
(a)
$\sqrt{2}x_1x_2$
$\sqrt{2}x_1$
$(1, 1, +\sqrt{2}, -\sqrt{2}, -\sqrt{2}, 1)$
$(1, 1, +\sqrt{2}, +\sqrt{2}, +\sqrt{2}, 1)$
$(1, 1, -\sqrt{2}, -\sqrt{2}, +\sqrt{2}, 1)$
$(1, 1, -\sqrt{2}, +\sqrt{2}, -\sqrt{2}, 1)$
(b)

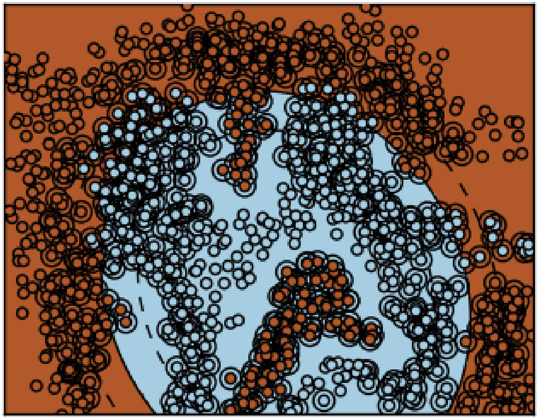
Polynomial Kernel $d = 1, c = 5$



Polynomial Kernel $d = 2, c = 5$

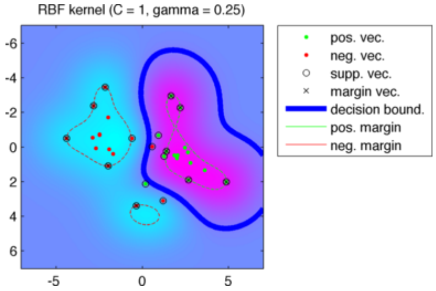


Polynomial Kernel $d = 3, c = 5$

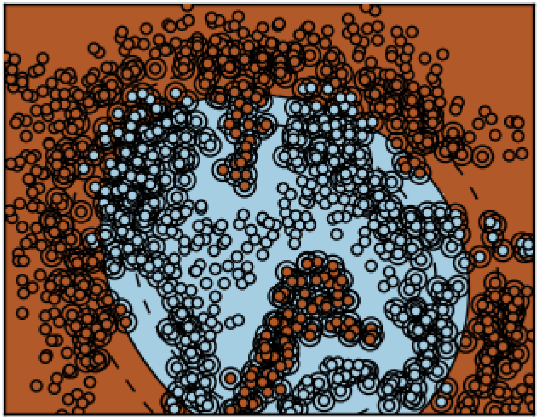


Gaussian Kernel

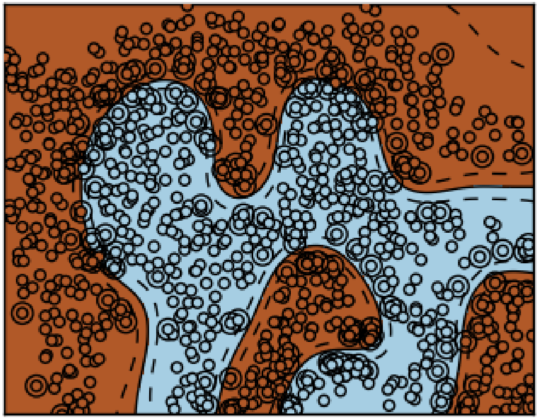
$$K(x,x') = \exp\left(\gamma \left\|x' - x\right\|^2\right) \tag{9}$$



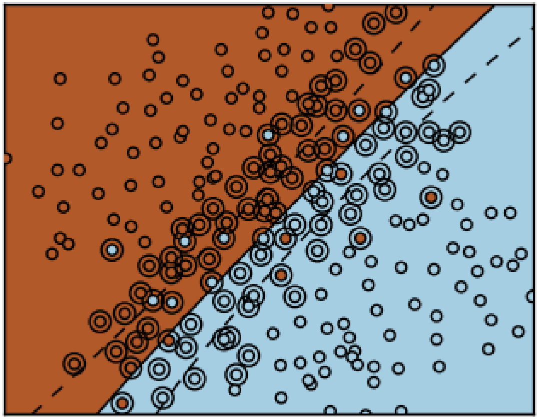
RBF Kernel $\gamma = 2$



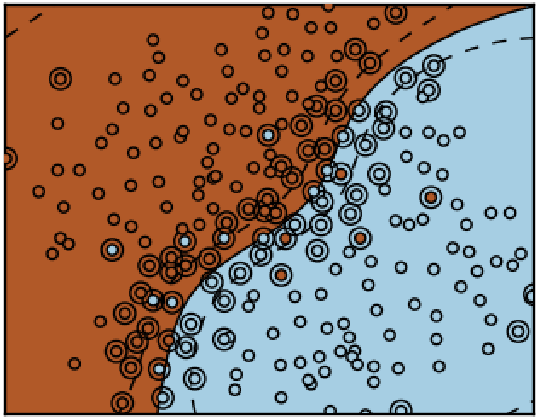
RBF Kernel $\gamma = 100$



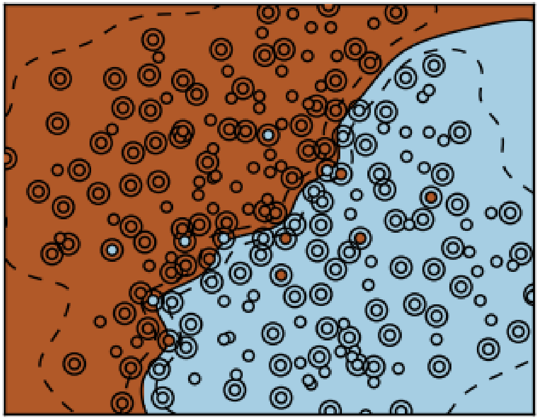
RBF Kernel $\gamma = 1$



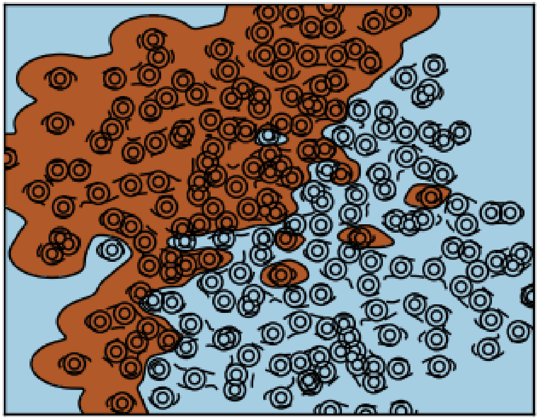
RBF Kernel $\gamma = 10$



RBF Kernel $\gamma = 100$



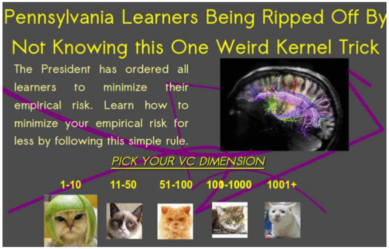
RBF Kernel $\gamma = 1000$



Be careful!

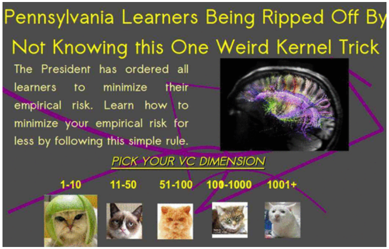
- Which has the lowest training error?
- Which one would generalize best?

Recap


Pennsylvania Learners Being Ripped Off By Not Knowing this One Weird Kernel Trick
The President has ordered all learners to minimize their empirical risk. Learn how to minimize your empirical risk for less by following this simple rule.
PICK YOUR VC DIMENSION
1-10
11-50
51-100
100-1000
1001+

- This completes our discussion of SVMs
- Workhorse method of machine learning
- Flexible, fast, effective

Recap


Pennsylvania Learners Being Ripped Off By
Not Knowing this One Weird Kernel Trick
The President has ordered all learners to minimize their empirical risk. Learn how to minimize your empirical risk for less by following this simple rule.
PICK YOUR VC DIMENSION
1-10
11-50
51-100
100-1000
1001+

- This completes our discussion of SVMs
- Workhorse method of machine learning
- Flexible, fast, effective
- Kernels: applicable to wide range of data, inner product trick keeps method simple