Kernels

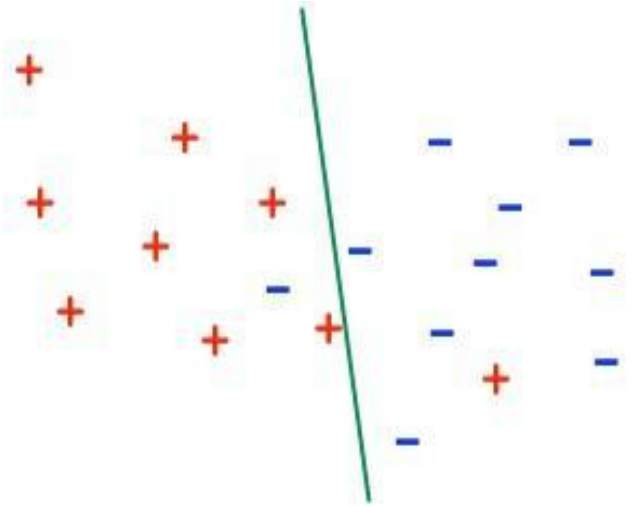
MGTF 495

Class Outline

- Parametric Methods
  - Discriminative Methods
    - Logistic Regression
      - Gradient Descent
      - Newton-Raphson
      - Hands-On
    - SVM
      - Hands-On
    - Perceptron
      - Hands-On
  - Kernels
  - Richer Output Spaces

Deviations from linear separability

Noise

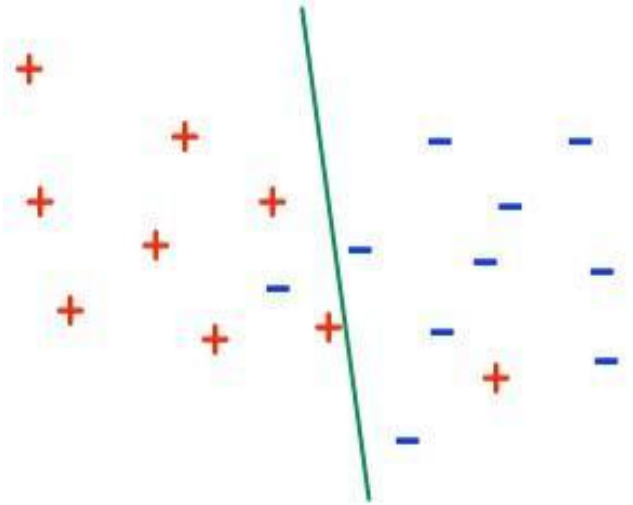


Find a separator that minimizes a convex loss function related to the number of mistakes.

e.g. SVM, logistic regression.

Deviations from linear separability

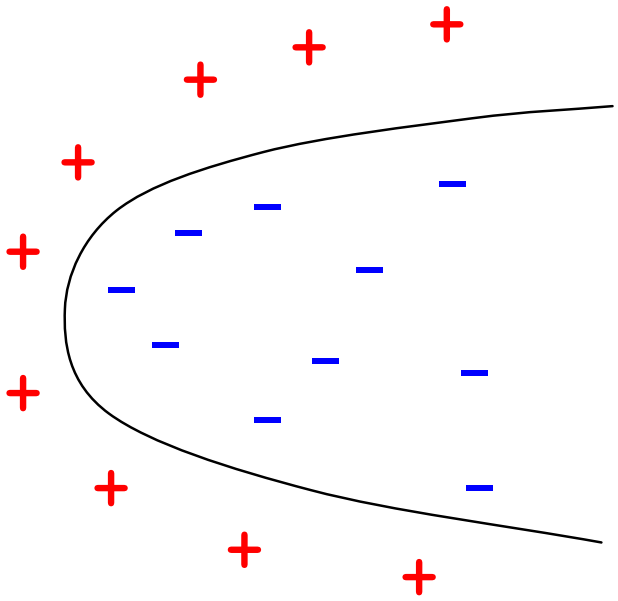
Noise



Find a separator that minimizes a convex loss function related to the number of mistakes.

e.g. SVM, logistic regression.

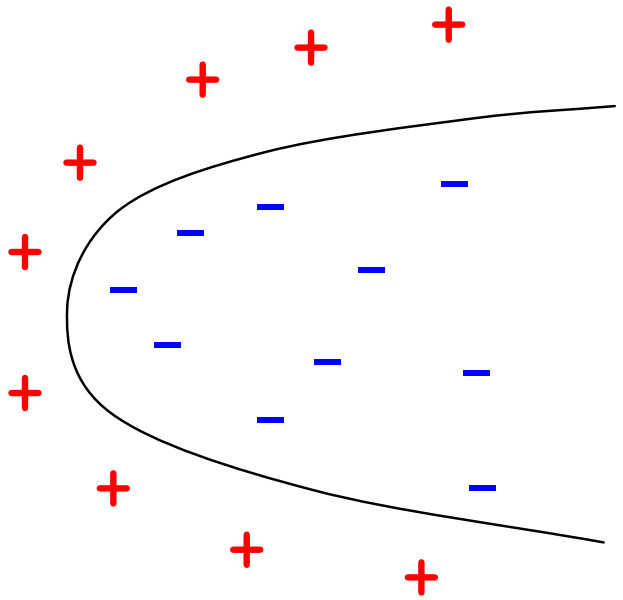
Systematic deviation



What to do with this?

Systematic inseparability

In this case, the actual boundary looks quadratic.



Quick fix: in addition to the regular features $x = (x_1, x_2, \ldots, x_p)$, add in extra features:

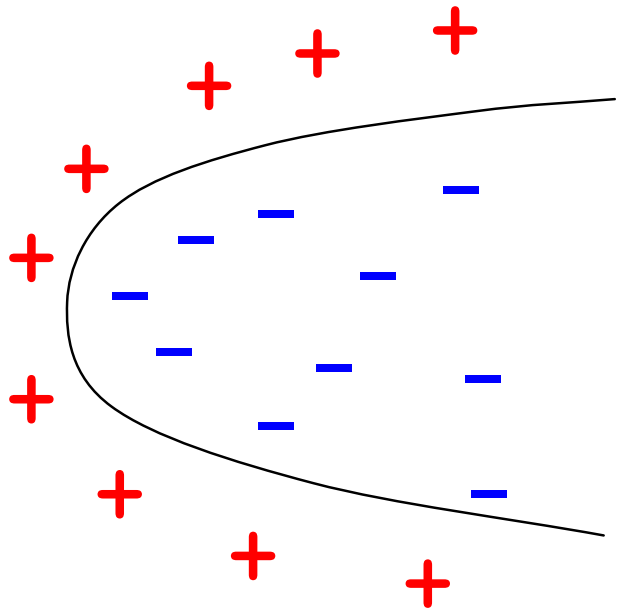
$$x_1^2,\ x_2^2,\ \ldots,\ x_p^2$$

$$x_1x_2,\ x_1x_3,\ \ldots,\ x_{p-1}x_p$$

The new, enhanced data vectors are of the form:

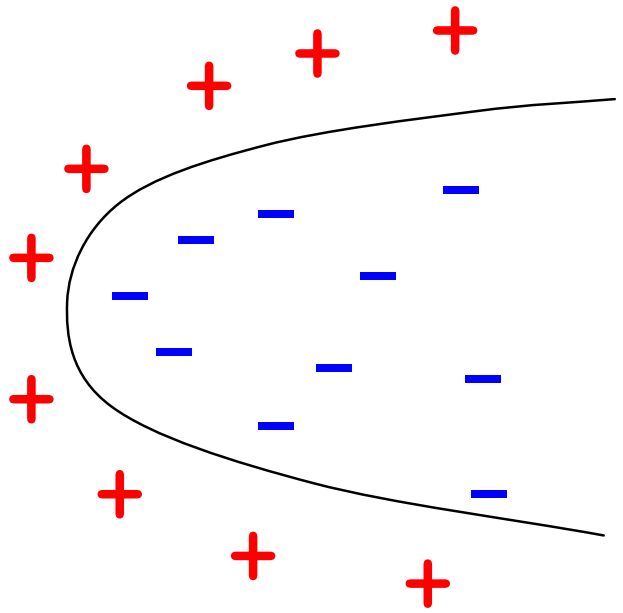
$$\Phi(x) = (1, \sqrt{2}x_1, \ldots, \sqrt{2}x_p, x_1^2, \ldots, x_p^2, \sqrt{2}x_1x_2, \ldots, \sqrt{2}x_{p-1}x_p).$$

Adding new features



Actual boundary is something like $x_1 = x_2^2 + 5$.

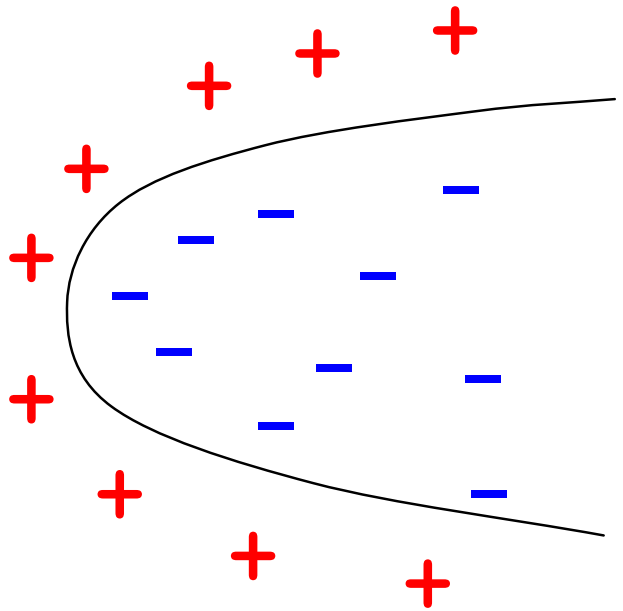
Adding new features



Actual boundary is something like $x_1 = x_2^2 + 5$.

- This is quadratic in $x = (1, x_1, x_2)$
- But it is linear in $\Phi(x) = (1, x_1, x_2, x_1^2, x_2^2, x_1 x_2)$

Adding new features



Actual boundary is something like $x_1 = x_2^2 + 5$.

- This is quadratic in $x = (1, x_1, x_2)$
- But it is linear in $\Phi(x) = (1, x_1, x_2, x_1^2, x_2^2, x_1 x_2)$

By embedding the data in a higher-dimensional feature space, we can keep using a linear classifier!

Quick quiz

1. Suppose $x = (1, x_1, x_2, x_3)$. What is the dimension of $\Phi(x)$?

Quick quiz

1. Suppose $x = (1, x_1, x_2, x_3)$. What is the dimension of $\Phi(x)$?

Ten

Features: $1, x_1, x_2, x_3, x_1^2, x_2^2, x_3^2, x_1x_2, x_2x_3, x_1x_3$

Quick quiz

1. Suppose $x = (1, x_1, x_2, x_3)$. What is the dimension of $\Phi(x)$?

Ten

Features: $1, x_1, x_2, x_3, x_1^2, x_2^2, x_3^2, x_1x_2, x_2x_3, x_1x_3$

2. What if $x = (1, x_1, \ldots, x_p)$?

Quick quiz

1. Suppose $x = (1, x_1, x_2, x_3)$. What is the dimension of $\Phi(x)$?

Ten

Features: $1, x_1, x_2, x_3, x_1^2, x_2^2, x_3^2, x_1x_2, x_2x_3, x_1x_3$

2. What if $x = (1, x_1, \ldots, x_p)$?

Use the formula – $(1 + 2p + \binom{p}{2})$

Perceptron revisited

Learning in the higher-dimensional feature space:

- $w = 0$
- while some y $(w \cdot \Phi(x)) < 0$:
  - $w = w + y\,\Phi(x)$

Final w is a weighted linear sum of various $\Phi(x)$.

Perceptron revisited

Learning in the higher-dimensional feature space:

- $w = 0$
- while some y $(w \cdot \Phi(x)) < 0$:
  - $w = w + y\,\Phi(x)$

Final w is a weighted linear sum of various $\Phi(x)$.

Problem: number of features has now increased dramatically.

For MNIST, from 784 to 307720!

Perceptron revisited

Learning in the higher-dimensional feature space:

- $w = 0$
- while some y $(w \cdot \Phi(x)) < 0$:
  - $w = w + y\,\Phi(x)$

Final w is a weighted linear sum of various $\Phi(x)$.

Problem: number of features has now increased dramatically.

For MNIST, from 784 to 307720!

The **kernel trick** (Aizenman, Braverman, Rozonoer (1964)):

- The only time we ever access $\Phi(x)$ is to compute $w \cdot \Phi(x)$. If, say,

$$w = a_1\Phi(x^{(1)}) + a_2\Phi(x^{(2)}) + a_3\Phi(x^{(3)}),$$

then $w \cdot \Phi(x)$ is a weighted sum of dot products $\Phi(x) \cdot \Phi(x^{(i)})$.

Perceptron revisited

Learning in the higher-dimensional feature space:

- $w = 0$
- while some y $(w \cdot \Phi(x)) < 0$:
  - $w = w + y\,\Phi(x)$

Final w is a weighted linear sum of various $\Phi(x)$.

Problem: number of features has now increased dramatically.

For MNIST, from 784 to 307720!

The **kernel trick** (Aizenman, Braverman, Rozonoer (1964)):

- The only time we ever access $\Phi(x)$ is to compute $w \cdot \Phi(x)$. If, say,

$$w = a_1\Phi(x^{(1)}) + a_2\Phi(x^{(2)}) + a_3\Phi(x^{(3)}),$$

  then $w \cdot \Phi(x)$ is a weighted sum of dot products $\Phi(x) \cdot \Phi(x^{(i)})$.

- Can we compute such dot products without writing out the $\Phi(x)$'s?

Computing dot products

In 2-d:

$$
\begin{aligned}
&\Phi(x) \cdot \Phi(z) \\
&= (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2) \cdot (1, \sqrt{2}z_1, \sqrt{2}z_2, z_1^2, z_2^2, \sqrt{2}z_1z_2) \\
&= 1 + 2x_1z_1 + 2x_2z_2 + x_1^2z_1^2 + x_2^2z_2^2 + 2x_1x_2z_1z_2 \\
&= (1 + x_1z_1 + x_2z_2)^2 \quad = \quad (1 + x \cdot z)^2
\end{aligned}
$$

Computing dot products

In 2-d:

$$\begin{aligned}
&\Phi(x) \cdot \Phi(z) \\
&= (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2) \cdot (1, \sqrt{2}z_1, \sqrt{2}z_2, z_1^2, z_2^2, \sqrt{2}z_1z_2) \\
&= 1 + 2x_1z_1 + 2x_2z_2 + x_1^2z_1^2 + x_2^2z_2^2 + 2x_1x_2z_1z_2 \\
&= (1 + x_1z_1 + x_2z_2)^2 \quad = \quad (1 + x \cdot z)^2
\end{aligned}$$

In p dimensions:

$$\begin{aligned}
&\Phi(x) \cdot \Phi(z) \\
&= (1, \sqrt{2}x_1, \ldots, \sqrt{2}x_p, x_1^2, \ldots, x_p^2, \sqrt{2}x_1x_2, \ldots, \sqrt{2}x_{p-1}x_p) \cdot \\
&\quad (1, \sqrt{2}z_1, \ldots, \sqrt{2}z_p, z_1^2, \ldots, z_p^2, \sqrt{2}z_1z_2, \ldots, \sqrt{2}z_{p-1}z_p) \\
&= 1 + 2\sum_i x_iz_i + \sum_i x_i^2z_i^2 + 2\sum_{i \neq j} x_ix_jz_iz_j \\
&= (1 + x_1z_1 + \cdots + x_pz_p)^2 \quad = \quad (1 + x \cdot z)^2
\end{aligned}$$

Computing dot products

In 2-d:

$$\begin{aligned}
&\Phi(x) \cdot \Phi(z) \\
&= (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2) \cdot (1, \sqrt{2}z_1, \sqrt{2}z_2, z_1^2, z_2^2, \sqrt{2}z_1z_2) \\
&= 1 + 2x_1z_1 + 2x_2z_2 + x_1^2z_1^2 + x_2^2z_2^2 + 2x_1x_2z_1z_2 \\
&= (1 + x_1z_1 + x_2z_2)^2 \quad = \quad (1 + x \cdot z)^2
\end{aligned}$$

In p dimensions:

$$\begin{aligned}
&\Phi(x) \cdot \Phi(z) \\
&= (1, \sqrt{2}x_1, \ldots, \sqrt{2}x_p, x_1^2, \ldots, x_p^2, \sqrt{2}x_1x_2, \ldots, \sqrt{2}x_{p-1}x_p) \cdot \\
&\quad (1, \sqrt{2}z_1, \ldots, \sqrt{2}z_p, z_1^2, \ldots, z_p^2, \sqrt{2}z_1z_2, \ldots, \sqrt{2}z_{p-1}z_p) \\
&= 1 + 2\sum_i x_iz_i + \sum_i x_i^2z_i^2 + 2\sum_{i \neq j} x_ix_jz_iz_j \\
&= (1 + x_1z_1 + \cdots + x_pz_p)^2 \quad = \quad (1 + x \cdot z)^2
\end{aligned}$$

For MNIST: computing dot products in the 307720-dimensional quadratic feature space takes time proportional to just 784, the original dimension!

The kernel trick

Why does it work?

The kernel trick

Why does it work?

1. The only time we ever look at data – during training or subsequent classification – is to compute dot products $w \cdot \Phi(x)$.

The kernel trick

Why does it work?

1. The only time we ever look at data – during training or subsequent classification – is to compute dot products $w \cdot \Phi(x)$.

2. And w itself is a linear combination of $\Phi(x)$'s. If, say,

$$w = \alpha_1 \Phi(x^{(1)}) + \alpha_{22} \Phi(x^{(22)}) + \alpha_{37} \Phi(x^{(37)}),$$

we can store w as $[(1, \alpha_1), (22, \alpha_{22}), (37, \alpha_{37})]$.

3. Dot products $\Phi(x) \cdot \Phi(x')$ can be computed efficiently, without ever writing out the high-dimensional embedding $\Phi(\cdot)$.

Kernel Perceptron

Learning from data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathcal{X} \times \{-1, 1\}$

Primal form:

- $w = 0$
- while there is some i with $y^{(i)}(w \cdot \Phi(x^{(i)})) < 0$:
  - $w = w + y^{(i)}\, \Phi(x^{(i)})$

Kernel Perceptron

Learning from data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathcal{X} \times \{-1, 1\}$

Primal form:

- $w = 0$
- while there is some i with $y^{(i)}(w \cdot \Phi(x^{(i)})) < 0$:
  - $w = w + y^{(i)}\, \Phi(x^{(i)})$

Dual form: $w = \sum_j \alpha_j y^{(j)} \Phi(x^{(j)})$, where $\alpha \in \mathbb{R}^n$.

- $\alpha = 0$
- while there is some i with $y^{(i)} \left(\sum_j \alpha_j y^{(j)} \underbrace{\Phi(x^{(j)}) \cdot \Phi(x^{(i)})}_{\text{efficient}} \right) < 0$:
  - $\alpha_i = \alpha_i + 1$

Kernel Perceptron

Learning from data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathcal{X} \times \{-1, 1\}$

Primal form:

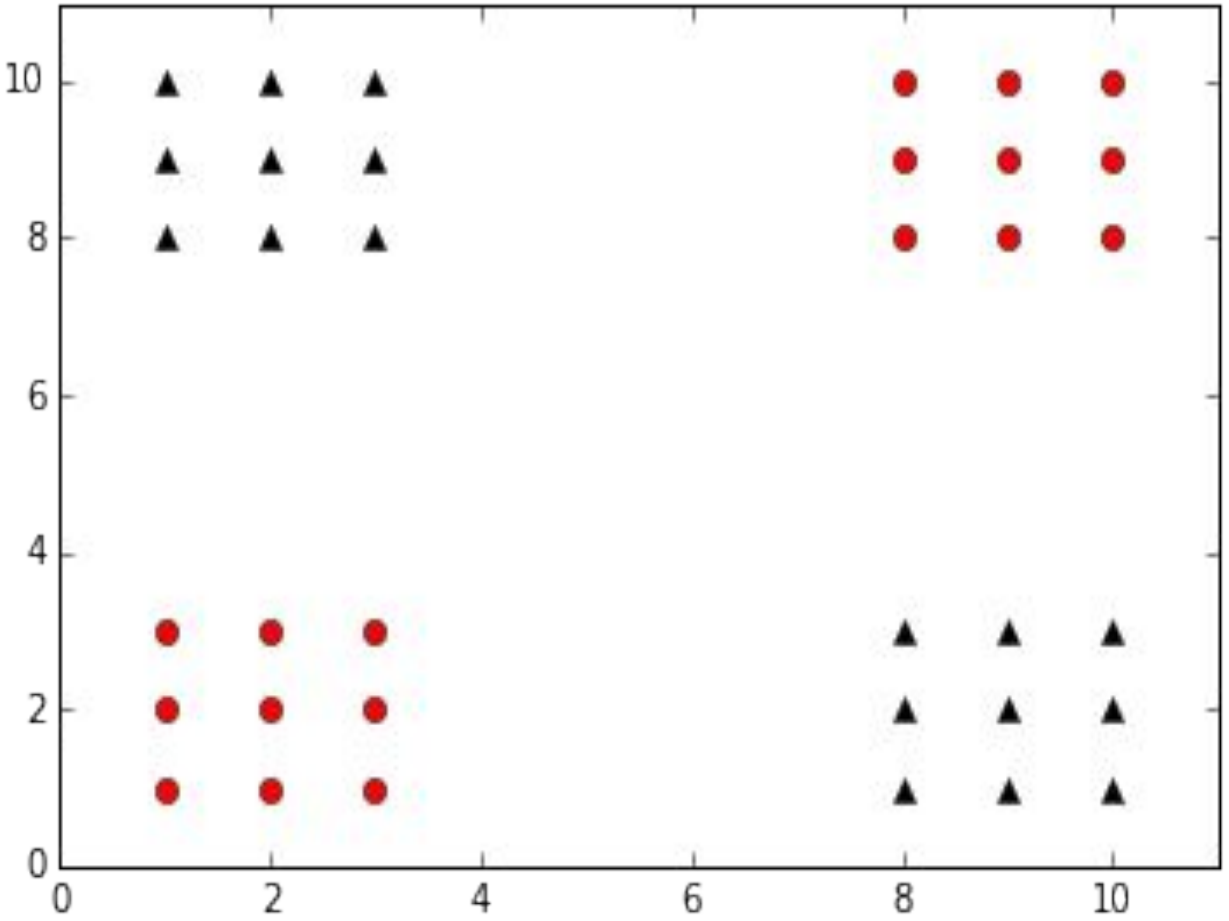
- $w = 0$
- while there is some i with $y^{(i)}(w \cdot \Phi(x^{(i)})) < 0$:
  - $w = w + y^{(i)}\,\Phi(x^{(i)})$

Dual form: $w = \sum_j \alpha_j y^{(j)} \Phi(x^{(j)})$, where $\alpha \in \mathbb{R}^n$.

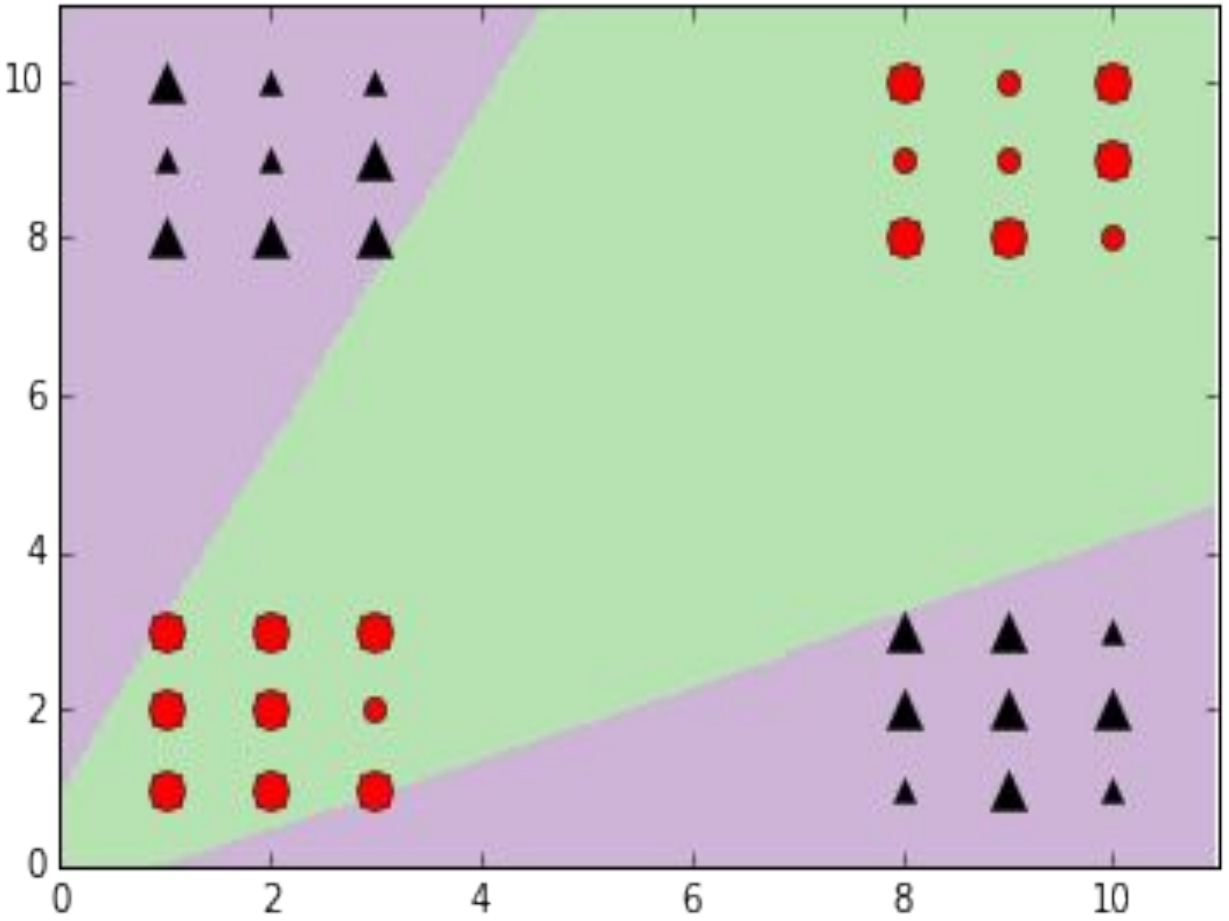
- $\alpha = 0$
- while there is some i with $y^{(i)} \left(\sum_j \alpha_j y^{(j)} \underbrace{\Phi(x^{(j)}) \cdot \Phi(x^{(i)})}_{\text{efficient}} \right) < 0$:
  - $\alpha_i = \alpha_i + 1$

To classify a new point x: Return sign $\left(\sum_j \alpha_j y^{(j)} \Phi(x^{(j)}) \cdot \Phi(x) \right)$.

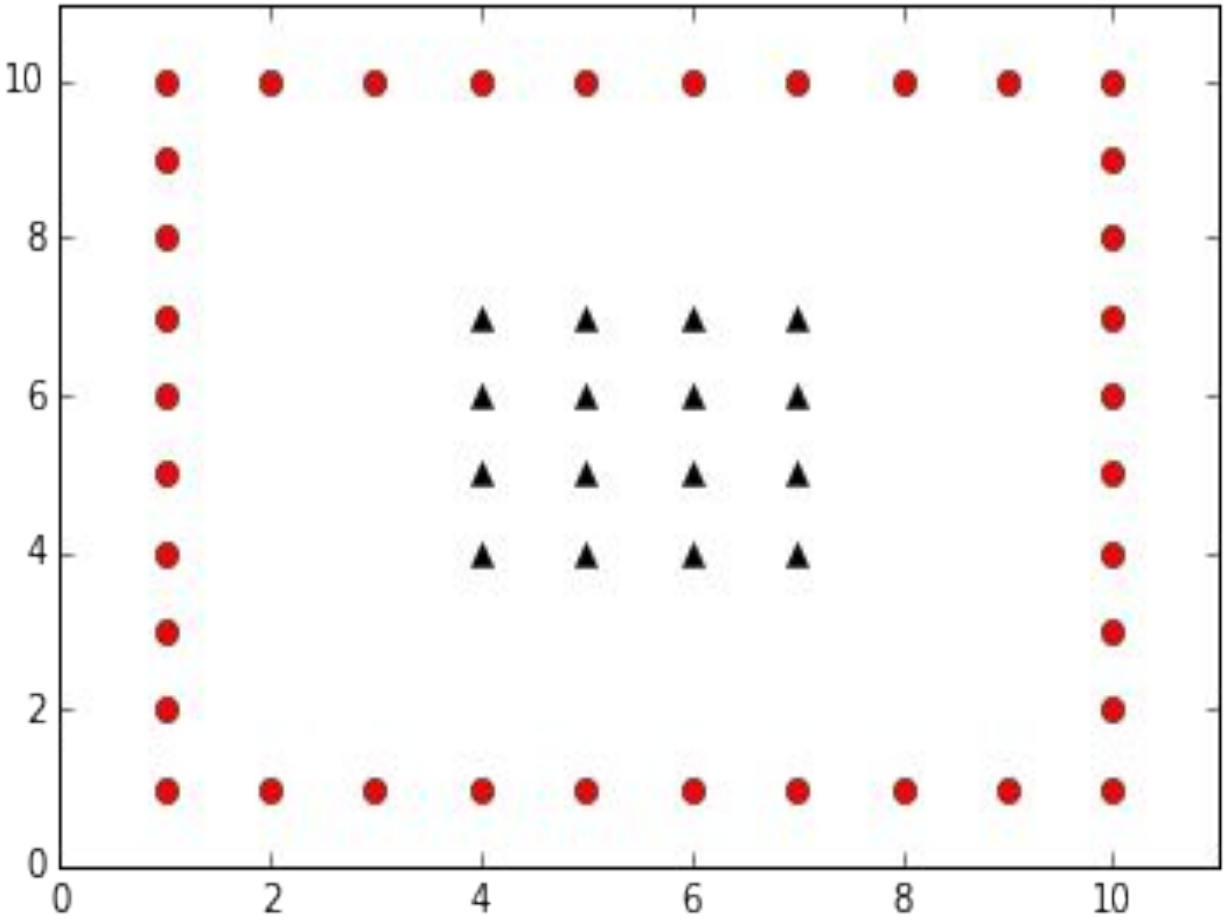
Kernel Perceptron: Examples



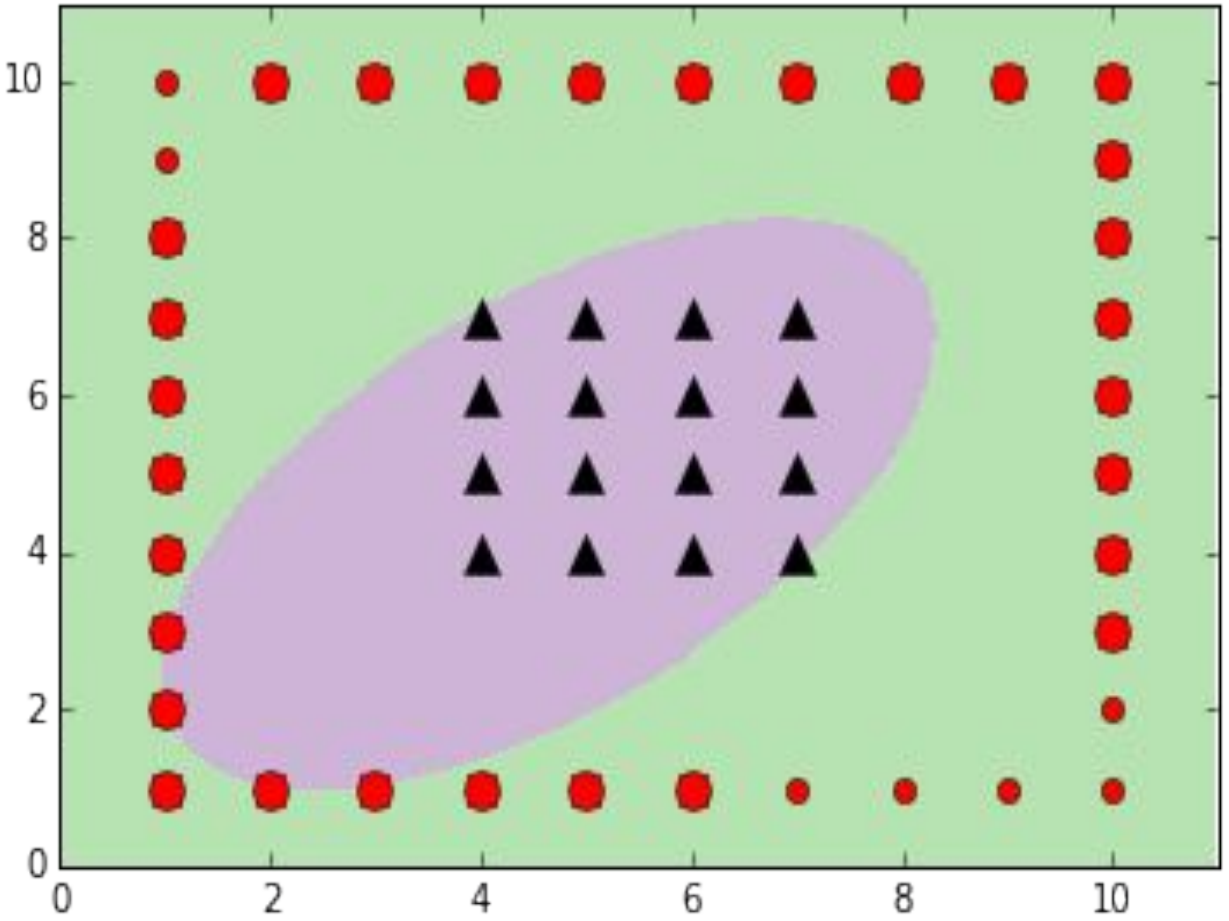
Kernel Perceptron: Examples



Kernel Perceptron: Examples



Kernel Perceptron: Examples



Quick quiz

Recall the kernel perceptron algorithm:

- $\alpha = 0$
- while there is some i with $y^{(i)}(\sum_j \alpha_j y^{(j)} \Phi(x^{(j)}) \cdot \Phi(x^{(i)})) < 0$:
  - $\alpha_i = \alpha_i + 1$

Suppose we run it on a data set and find that it converges after k updates.

1. How many dot product computations are needed to classify a new point?

Quick quiz

Recall the kernel perceptron algorithm:

- $\alpha = 0$
- while there is some i with $y^{(i)}(\sum_j \alpha_j y^{(j)} \Phi(x^{(j)}) \cdot \Phi(x^{(i)})) < 0$:
  - $\alpha_i = \alpha_i + 1$

Suppose we run it on a data set and find that it converges after k updates.

1. How many dot product computations are needed to classify a new point?

Solution: At most k

Quick quiz

Recall the kernel perceptron algorithm:

- $\alpha = 0$
- while there is some i with $y^{(i)}(\sum_j \alpha_j y^{(j)} \Phi(x^{(j)}) \cdot \Phi(x^{(i)})) < 0$:
  - $\alpha_i = \alpha_i + 1$

Suppose we run it on a data set and find that it converges after k updates.

1. How many dot product computations are needed to classify a new point?

   Solution: At most k

2. What is the total number of dot product computations during learning?

Quick quiz

Recall the kernel perceptron algorithm:

- $\alpha = 0$
- while there is some i with $y^{(i)}(\sum_j \alpha_j y^{(j)} \Phi(x^{(j)}) \cdot \Phi(x^{(i)})) < 0$:
  - $\alpha_i = \alpha_i + 1$

Suppose we run it on a data set and find that it converges after k updates.

1. How many dot product computations are needed to classify a new point?

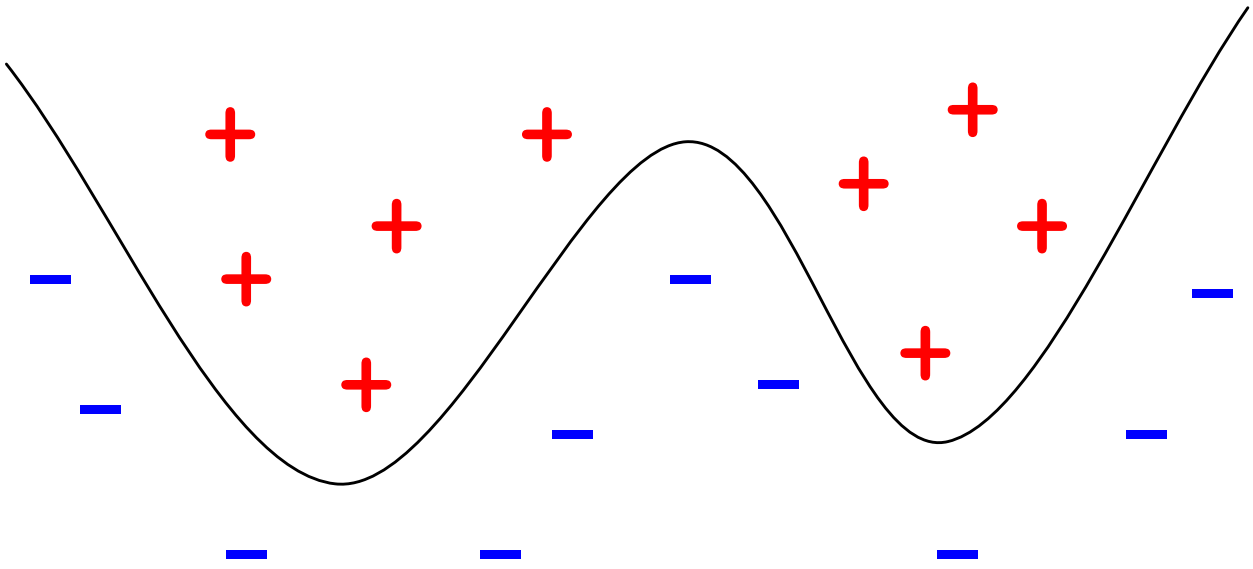
Solution: At most k

2. What is the total number of dot product computations during learning?

Solution: 1 term after first update, 2 terms after second update and so on till k updates. Therefore, in order of (1 + 2 + … + k) = **at least $\binom{k}{2}$** updates.

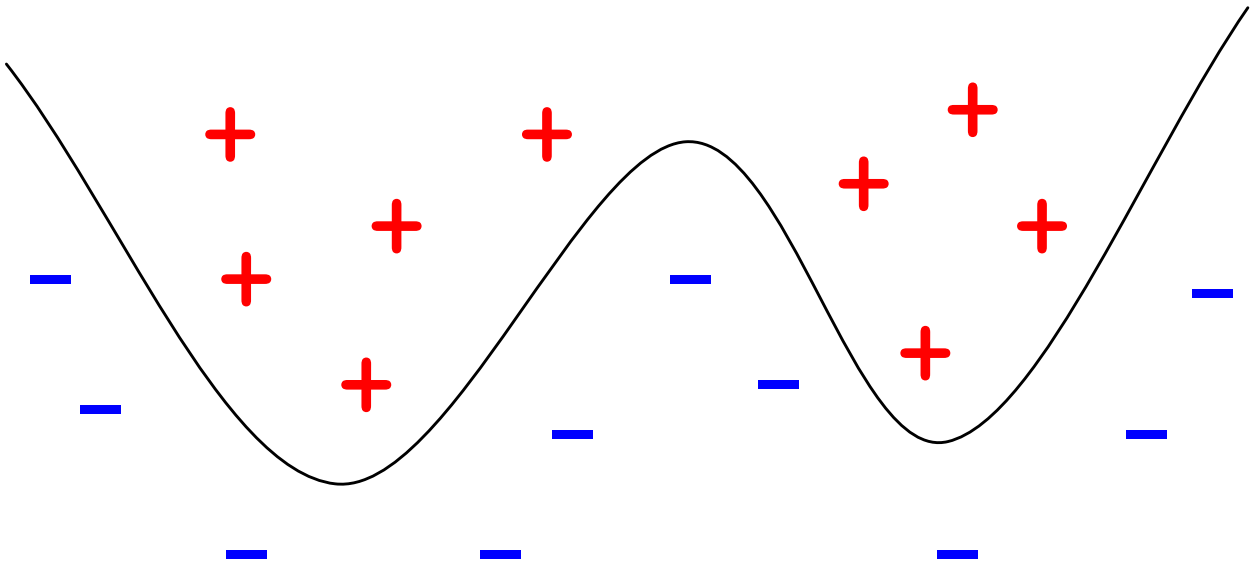
Polynomial decision boundaries

To get a decision surface which is an arbitrary polynomial of order *d* :



Polynomial decision boundaries

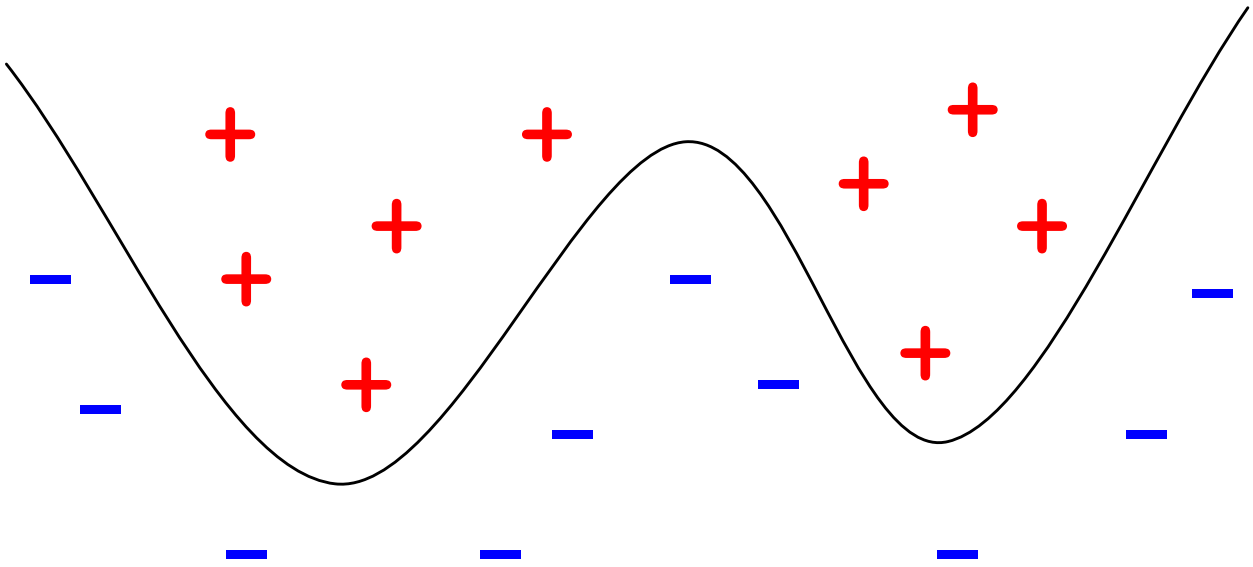
To get a decision surface which is an arbitrary polynomial of order d :



- Let $\Phi(x)$ consist of all terms of order $\leq d$, such as $x_1 x_2^2 x_3^{d-3}$. (How many such terms are there, roughly?)

Polynomial decision boundaries

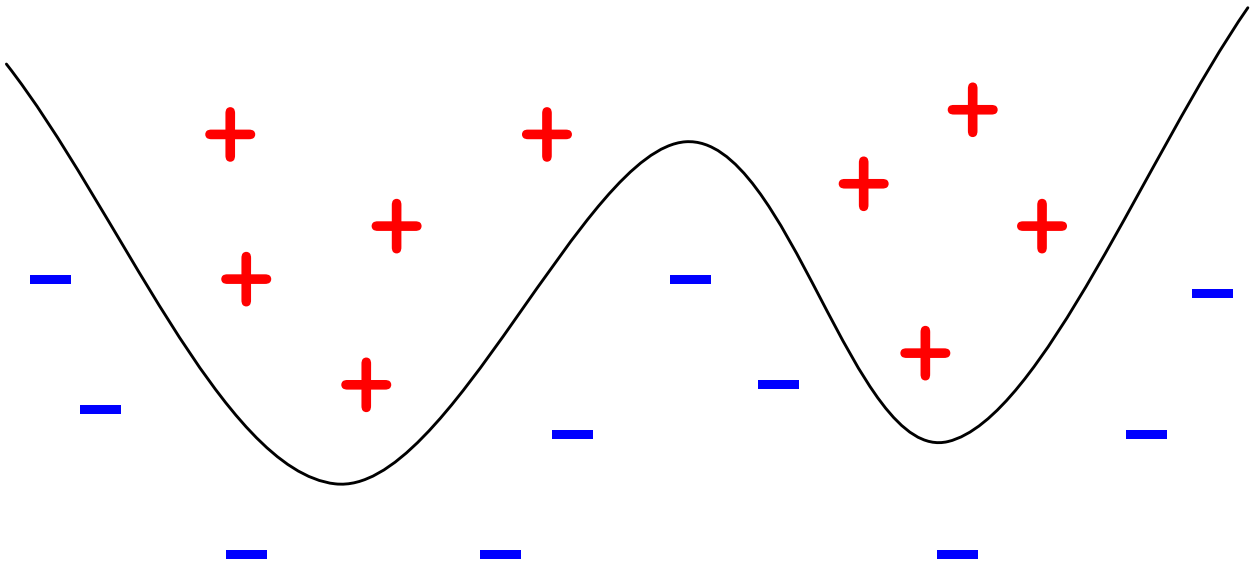
To get a decision surface which is an arbitrary polynomial of order d :



- Let $\Phi(x)$ consist of all terms of order $\leq d$, such as $x_1 x_2^2 x_3^{d-3}$. (How many such terms are there, roughly?)
- Same trick works: $\Phi(x) \cdot \Phi(z) = (1 + x \cdot z)^d$.

Polynomial decision boundaries

To get a decision surface which is an arbitrary polynomial of order d :



- Let $\Phi(x)$ consist of all terms of order $\leq d$, such as $x_1 x_2^2 x_3^{d-3}$. (How many such terms are there, roughly?)
- Same trick works: $\Phi(x) \cdot \Phi(z) = (1 + x \cdot z)^d$.

The **kernel function**, a measure of similarity:

$$k(x, z) = \Phi(x) \cdot \Phi(z).$$

The kernel function

Now shift attention: away from the embedding $\Phi(x)$, which we never explicitly construct, towards the thing we actually use:

- the **similarity measure** $k(x, z)$

Rewrite learning algorithm and final classifier in terms of k .

The kernel function

Now shift attention: away from the embedding $\Phi(x)$, which we never explicitly construct, towards the thing we actually use:

- the **similarity measure** $k(x, z)$

Rewrite learning algorithm and final classifier in terms of k.

The classifier $w \cdot \Phi(x)$, for

$$w = \alpha_1 y^{(1)} \Phi(x^{(1)}) + \cdots + \alpha_n y^{(n)} \Phi(x^{(n)}),$$

becomes a **similarity-weighted vote**,

$$F(x) = \alpha_1 y^{(1)} k(x^{(1)}, x) + \cdots + \alpha_n y^{(n)} k(x^{(n)}, x).$$

The kernel function

Now shift attention: away from the embedding $\Phi(x)$, which we never explicitly construct, towards the thing we actually use:

- the **similarity measure** $k(x, z)$

Rewrite learning algorithm and final classifier in terms of k.

The classifier $w \cdot \Phi(x)$, for

$$w = \alpha_1 y^{(1)} \Phi(x^{(1)}) + \cdots + \alpha_n y^{(n)} \Phi(x^{(n)}),$$

becomes a **similarity-weighted vote**,

$$F(x) = \alpha_1 y^{(1)} k(x^{(1)}, x) + \cdots + \alpha_n y^{(n)} k(x^{(n)}, x).$$

Can we choose k to be any similarity function suitable for the application domain?

Quick quiz

For each of the following kernel functions, find a corresponding embedding $x \to \Phi(x)$.

1. $k(x, z) = \cos(\text{angle between } x \text{ and } z)$.

Quick quiz

For each of the following kernel functions, find a corresponding embedding $x \rightarrow \Phi(x)$.

1. $k(x, z) = \cos(\text{angle between } x \text{ and } z)$.

1. We know, $x.z = |x||z|cos\theta \Rightarrow cos\theta = \frac{x.z}{|x||z|} \Rightarrow \phi(x) = \frac{x}{|x|}$

Quick quiz

For each of the following kernel functions, find a corresponding embedding $x \rightarrow \Phi(x)$.

1. $k(x, z) = \cos(\text{angle between } x \text{ and } z)$.

1. We know, $x.z = |x||z|cos\theta \Rightarrow cos\theta = \frac{x.z}{|x||z|} \Rightarrow \phi(x) = \frac{x}{|x|}$

2. $k(x, z) = x^T Az$, where A is a symmetric positive semidefinite matrix.

Quick quiz

For each of the following kernel functions, find a corresponding embedding $x \rightarrow \Phi(x)$.

1. $k(x, z) = \cos(\text{angle between } x \text{ and } z)$.

1. We know, $x.z = |x||z|cos\theta \Rightarrow cos\theta = \frac{x.z}{|x||z|} \Rightarrow \phi(x) = \frac{x}{|x|}$

2. $k(x, z) = x^T Az$, where A is a symmetric positive semidefinite matrix.

2. Since A is PSD, we can re-write it as UU^T

$$x^T Az = x^T UU^T z = (U^T x)^T (zU)^T$$
$$\Rightarrow \phi(x) = U^T x$$

Kernels: postscript

1. Customized kernels
   - For many different domains (NLP, biology, speech, ...)
   - Over many different structures (sequences, sets, trees, graphs, ...)

Kernels: postscript

1. Customized kernels
   - For many different domains (NLP, biology, speech, ...)
   - Over many different structures (sequences, sets, trees, graphs, ...)
2. Learning the kernel function

   Given a set of plausible kernels, find a linear combination of them that works well.

Kernels: postscript

1. Customized kernels
   - For many different domains (NLP, biology, speech, ...)
   - Over many different structures (sequences, sets, trees, graphs, ...)
2. Learning the kernel function

   Given a set of plausible kernels, find a linear combination of them that works well.
3. Speeding up learning and prediction

   The $n \times n$ kernel matrix $k(x^{(i)}, x^{(j)})$ is a bottleneck for large n.

   One idea:
   - Go back to the primal space!
   - Replace the embedding Φ by a low-dimensional mapping $\widetilde{\Phi}$ such that

$$\widetilde{\Phi}(x) \cdot \widetilde{\Phi}(z) \approx \Phi(x) \cdot \Phi(z).$$