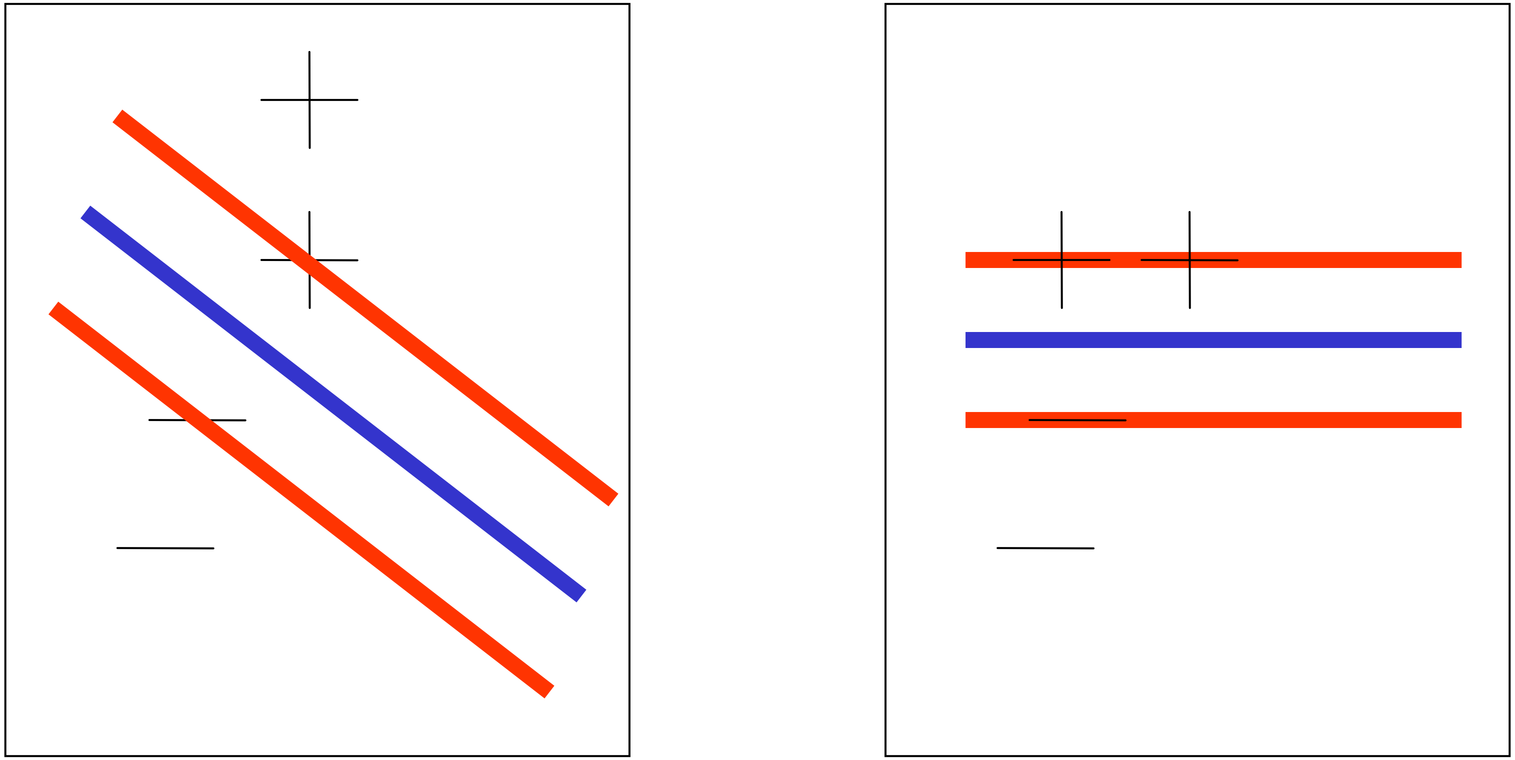
Foundations: fortunate choices

- Unusual choice of separation strategy:
  > Maximize "street" between groups
- Attack maximization problem:
  > Lagrange multipliers + hairy mathematics
- New problem is a quadratic minimization:
  > Susceptible to fancy numerical methods
- Result depends on dot products only
  > Enables use of kernel methods.

Key idea: find widest separating “street”



Classifier form is given and constrained

- Classify unknown **u** as plus if:

$$f(\mathbf{u}) = \mathbf{w} \cdot \mathbf{u} + b > 0$$

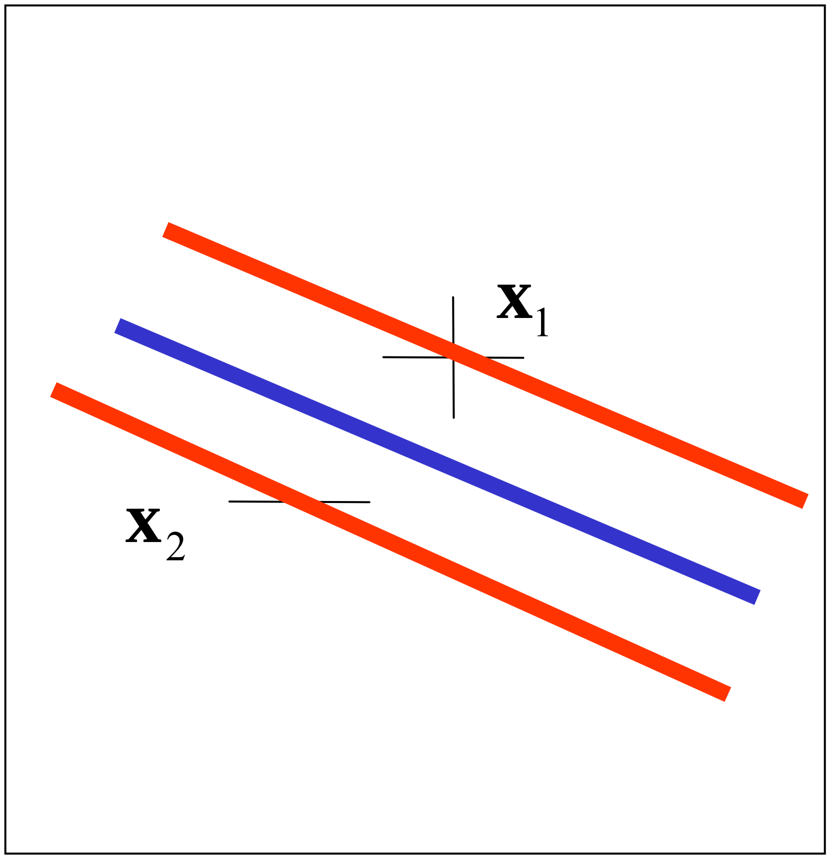
- Then, constrain, for all plus sample vectors:

$$f(\mathbf{x}_+) = \mathbf{w} \cdot \mathbf{x}_+ + b \geq 1$$

- And for all minus sample vectors

$$f(\mathbf{x}_-) = \mathbf{w} \cdot \mathbf{x}_- + b \leq -1$$

Distance between street's gutters



- The constraints require:

$$\mathbf{w} \cdot \mathbf{x}_1 + b = +1$$

$$\mathbf{w} \cdot \mathbf{x}_2 + b = -1$$

- So, subtracting:

$$\mathbf{w} \cdot (\mathbf{x}_1 - \mathbf{x}_2) = 2$$

- Dividing by the length of **w** produces the distance between the lines:

$$\frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot (\mathbf{x}_1 - \mathbf{x}_2) = \frac{2}{\|\mathbf{w}\|}$$

From maximizing to minimizing…

- So, to maximize the width of the street, you need to "wiggle" w until the length of w is minimum, *while still honoring constraints on gutter values*:

$$\frac{2}{\|\mathbf{w}\|} = \text{separation}$$

- One possible approach to finding the minimum is to use the method devised by Lagrange. Working through the method reveals how the sample data figures into the classification formula.

From maximizing to minimizing…

- A step toward solving the problem using LaGrange’s method is to note that maximizing street width is ensured if you minimize the following, *while still honoring constraints on gutter values.*

$$\frac{1}{2}\left\|\mathbf{w}\right\|^2$$

- Translation of the previous formula into this one, with ½ and squaring, is a mathematical convenience.

…while honoring constraints

- Remember, the minimization is constrained
- You can write the constraints as:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1$$

Where y*i* is 1 for plusses and –1 for minuses.

Dependence on dot products

- Using LaGrange's method, and working through some mathematics, you get to the following problem. When solved for the alphas, you then have what you need for the classification formula.

Maximize $\sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i,j=1}^{l} a_i a_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j$

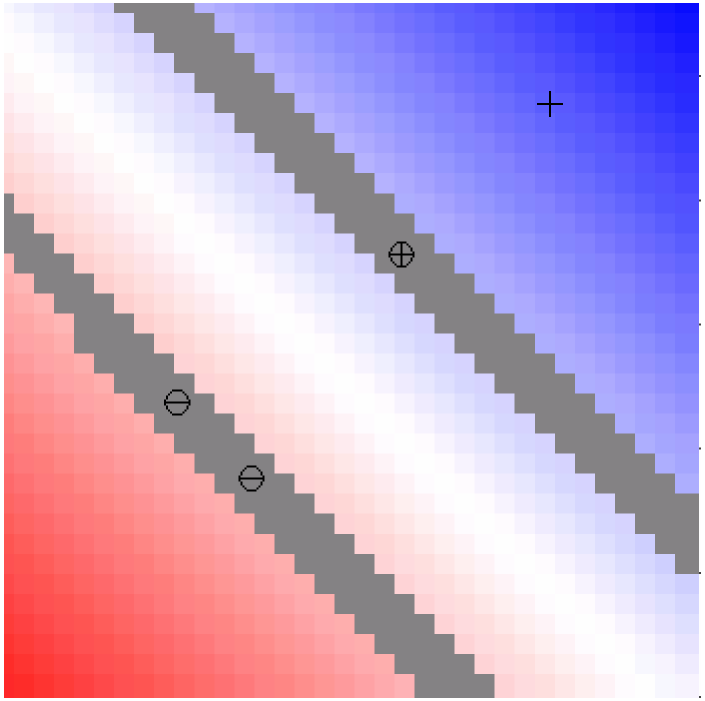
Subject to $\sum_{i}^{l} a_i y_i = 0_i$ and $a_i \geq 0$

Then check sign of $f(\mathbf{u}) = \mathbf{w} \cdot \mathbf{u} + b = (\sum_{i,j=1}^{l} a_i y_i \mathbf{x}_i \cdot \mathbf{u}) + b$

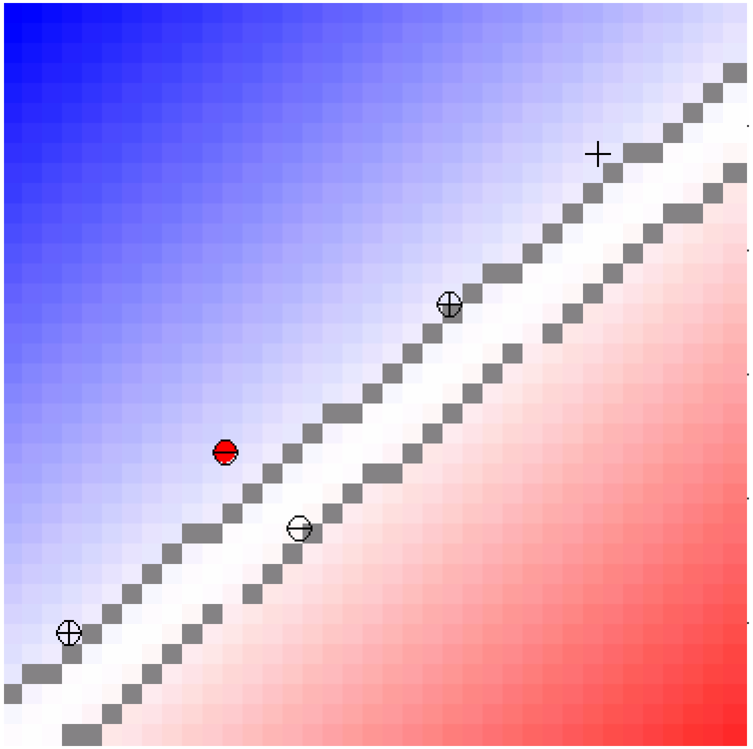
Key to importance

- Learning depends only on dot products of sample pairs.
- Classification depends only on dot products of unknown with samples.
- Exclusive reliance on dot products enables approach to problems in which samples cannot be separated by a straight line.

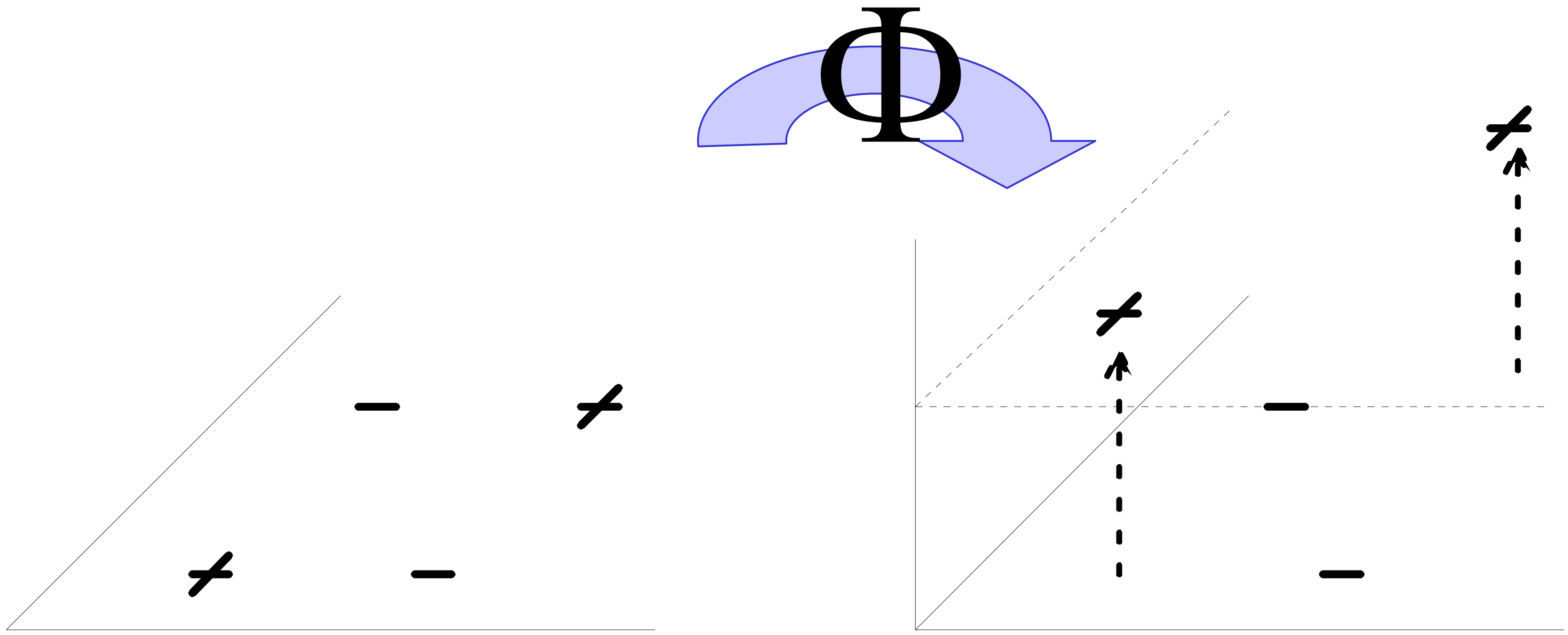
Example



Another example



Not separable?
Try another space!
Using some mapping, Φ



Problem starts here, 2D

Dot products computed here, 3D

What you need

- To get $\mathbf{x}_1$ into the high-dimensional space, you use

$\Phi(\mathbf{x}_1)$

- To optimize, you need

$\Phi(\mathbf{x}_1) \cdot \Phi(\mathbf{x}_2)$

- To use, you need

$\Phi(\mathbf{x}_1) \cdot \Phi(\mathbf{u})$

- So, all you need is a way to compute dot products in high-dimensional space as a function of vectors in original space!

What you don't need

- Suppose dot products are supplied by

$$\Phi(\mathbf{x}_1) \cdot \Phi(\mathbf{x}_2) = K(\mathbf{x}_1, \mathbf{x}_2)$$

- Then, all you need is

$$K(\mathbf{x}_1, \mathbf{x}_2)$$

- Evidently, you don't need to know what Φ is; having K is enough!

Standard choices

- No change

$$\Phi(\mathbf{x}_1)\cdot\Phi(\mathbf{x}_2) = K(\mathbf{x}_1,\mathbf{x}_2) = \mathbf{x}_1\cdot\mathbf{x}_2$$

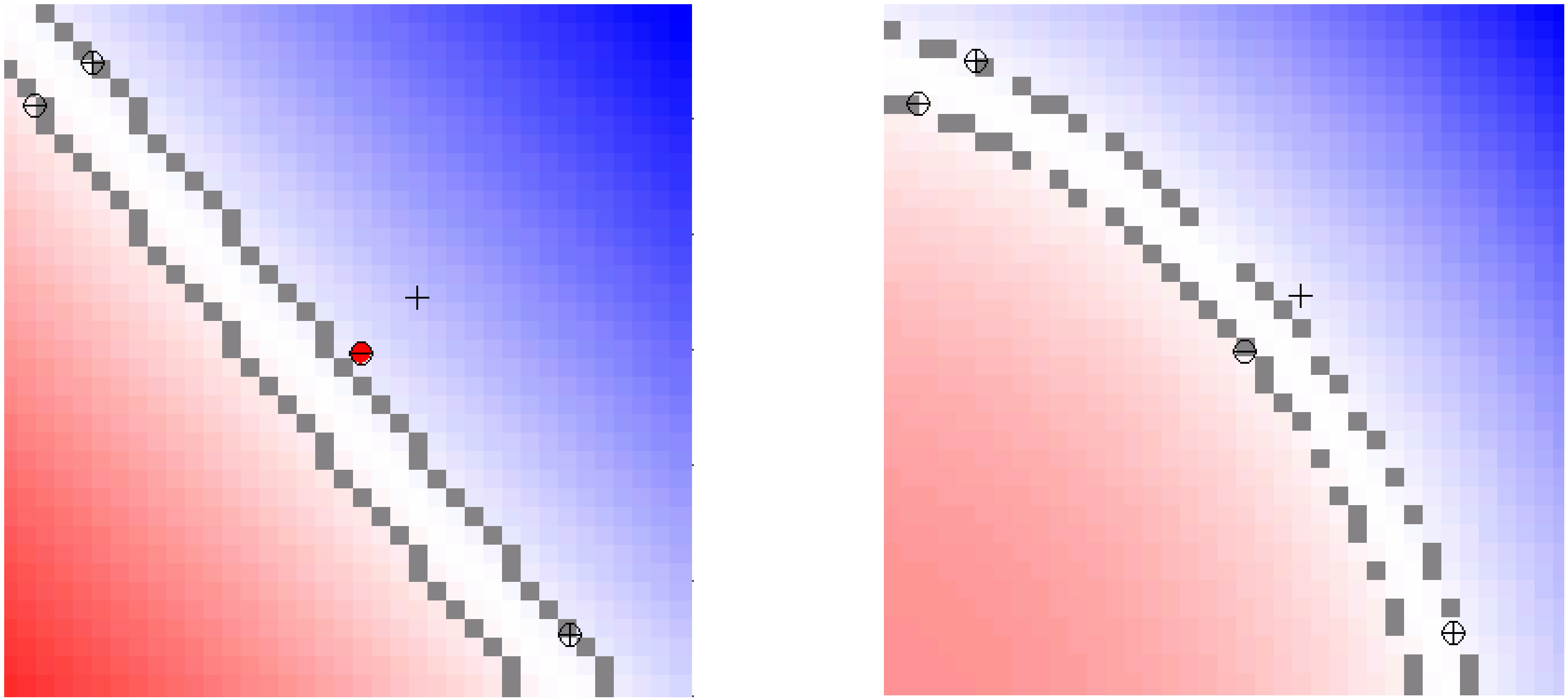
- Polynomial

$$K(\mathbf{x}_1,\mathbf{x}_2) = (\mathbf{x}_1\cdot\mathbf{x}_2 + 1)^n$$

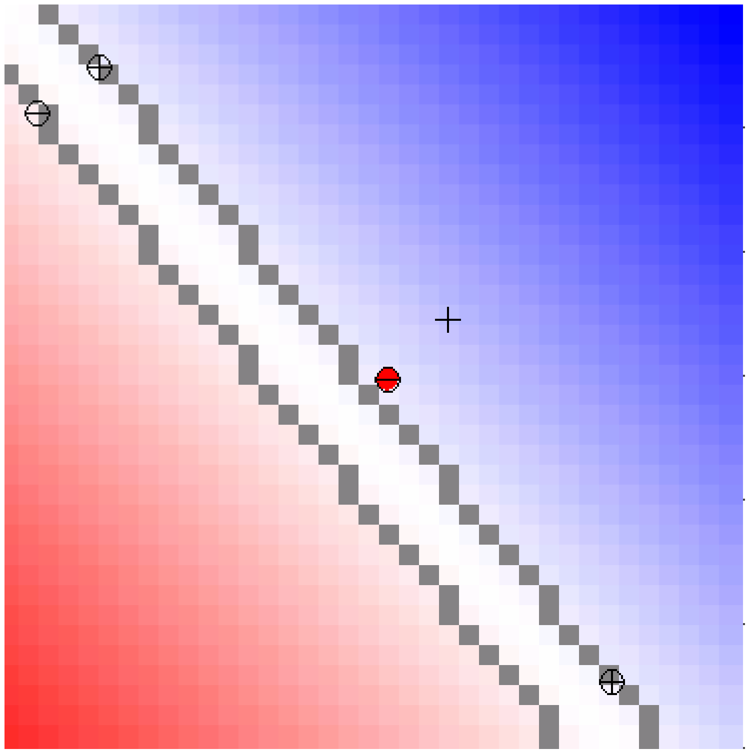
- Radial basis function

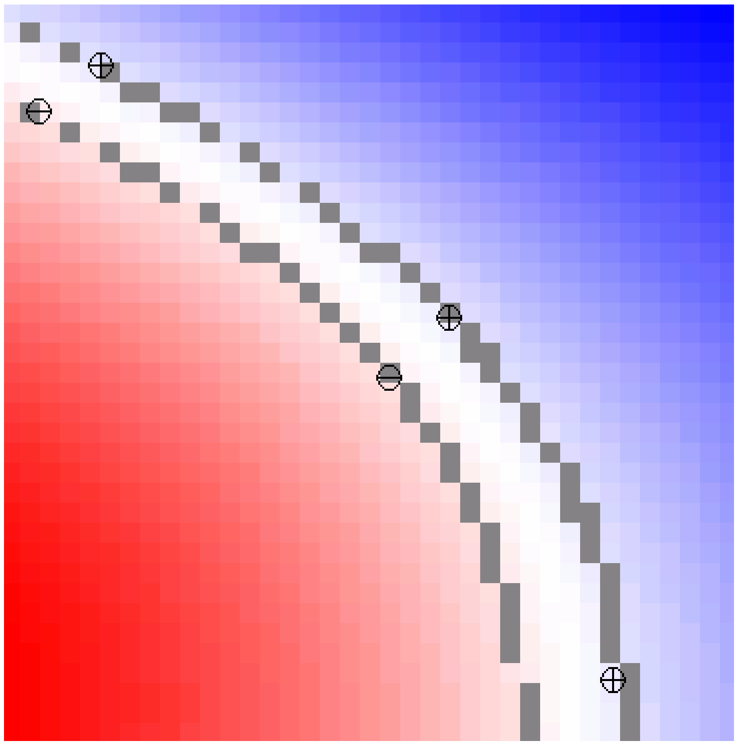
$$K(\mathbf{x}_1,\mathbf{x}_2) = e^{\frac{-\|\mathbf{x}_1-\mathbf{x}_2\|^2}{2\sigma^2}}$$

Polynomial Kernel

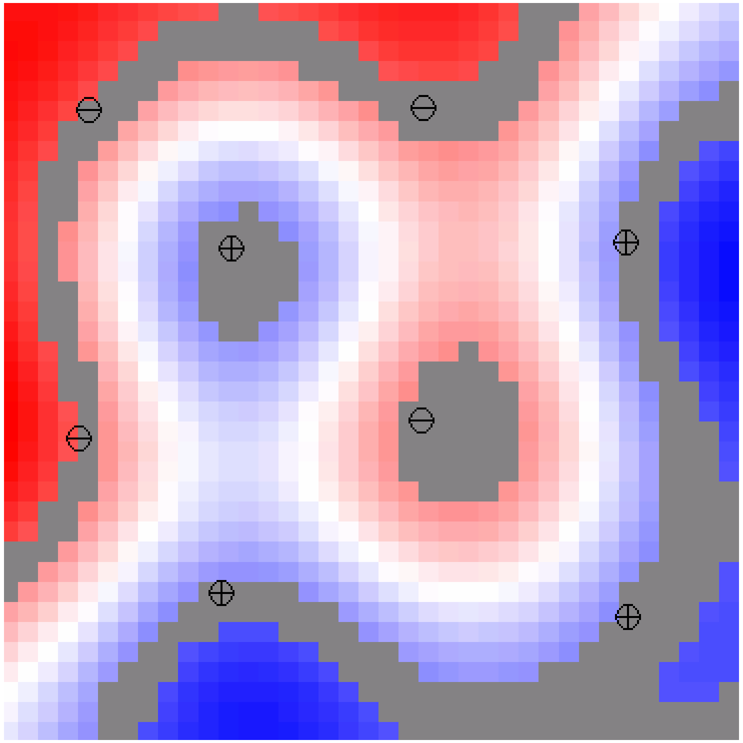


Radial-basis kernel





Another radial-basis example



Aside: about the hairy mathematics

- Step 1: Apply method of Lagrange multipliers

To minimize $\frac{1}{2}\|\mathbf{w}\|^2$ subject to constraints $y_i(\mathbf{x}_i \cdot \mathbf{w} + b) \geq 1$

Find places where $L = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{l} a_i(y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1)$

has zero derivatives

Aside: about the hairy mathematics

Step 2: remember how to differentiate vectors

$$\frac{\partial \|\mathbf{w}\|^2}{\partial \mathbf{w}} = 2\mathbf{w} \quad \text{and} \quad \frac{\partial \mathbf{x} \cdot \mathbf{w}}{\partial \mathbf{w}} = \mathbf{x}$$

Step 3: find derivatives of the Lagrangian L

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{l} a_i y_i \mathbf{x}_i = 0$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{l} a_i y_i = 0$$

Aside: about the hairy mathematics

- Step 4: do the algebra, then ask a numerical analyst to write a program to find the values of alpha that produce an extreme value for:

$$L = \sum_{i=1}^{l} a_i - \frac{1}{2} \sum_{i,j=1}^{l} a_i a_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j$$

But, note that

- Quadratic minimization depends on only on dot products of sample vectors
- Recognition depends only on dot products of unknown vector with sample vectors
- Reliance on only dot products key to remaining magic