Two examples of kernels

UNIVERSITY *of* WASHINGTON

How do we deal with high-dimensional lifts/data?

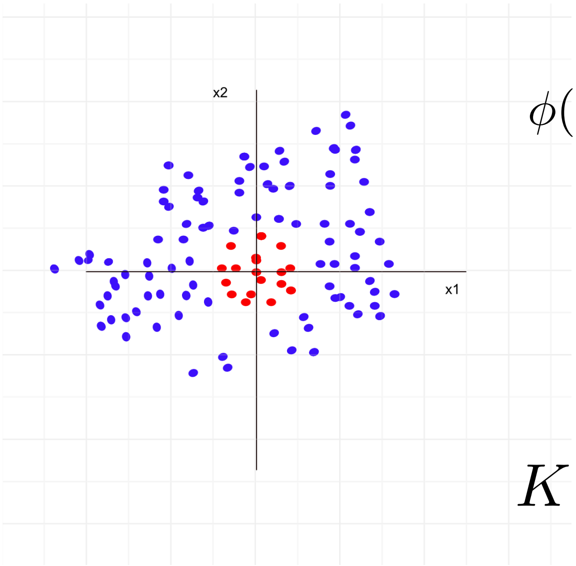
A fundamental trick in ML: use kernels

A function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a *kernel* for a map ϕ if $K(x, x') = \phi(x) \cdot \phi(x')$ for all x, x'.

So, if we can represent our algorithms/decision rules as dot products
and we can find a kernel for our feature map
then we can avoid explicitly dealing with $\phi(x)$.

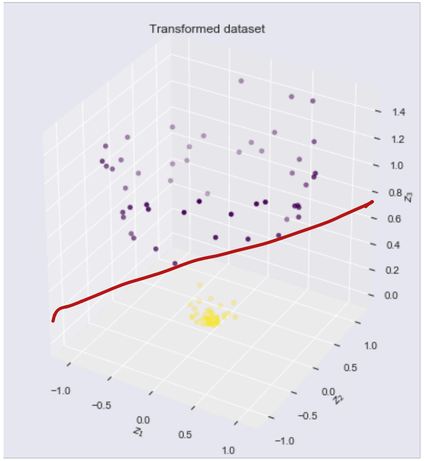
Why use a kernel?

- **Many decision boundaries aren't linearly separable in original feature space**



$$\phi(x) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

$$\longrightarrow$$



$$K(x, x') = (x^T x')^2$$

Photo credit:

https://xavierbourretsicotte.github.io/Kernel_feature_map.html

Examples of Kernels

- **Polynomials of degree exactly d**

$$K(\mathbf{u}, \mathbf{v}) = (\mathbf{u} \cdot \mathbf{v})^p$$

- **Polynomials of degree up to d**

$$K(\mathbf{u}, \mathbf{v}) = (\mathbf{u} \cdot \mathbf{v} + 1)^p$$

- **Gaussian (squared exponential) kernel**

$$K(\mathbf{u}, \mathbf{v}) = \exp\left(-\frac{||\mathbf{u} - \mathbf{v}||^2}{2\sigma^2}\right)$$

- **Sigmoid**

$$K(u, v) = \tanh(\gamma \cdot u^T v + r)$$

Don't forget regularization…

(Formulation of kernelized, regularized least squares)

Typically, $\mathbf{K} \succ 0$. What if $\lambda = 0$?

$$\widehat{\alpha} = \arg\min_{\alpha} ||\mathbf{y} - \mathbf{K}\alpha||_2^2 + \lambda\alpha^T\mathbf{K}\alpha$$

Don't forget regularization…

(Formulation of kernelized, regularized least squares)

Typically, $\mathbf{K} \succ 0$. What if $\lambda = 0$?

$$\widehat{\alpha} = \arg\min_{\alpha} ||\mathbf{y} - \mathbf{K}\alpha||_2^2 + \boxed{\lambda \alpha^T \mathbf{K} \alpha}$$

Unregularized kernel least squares can (over) fit **any data**!

$$\boxed{\widehat{\alpha} = \mathbf{K}^{-1}\mathbf{y}}$$

RBF Kernel

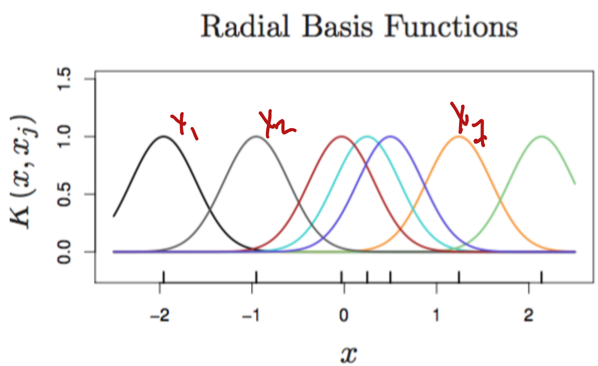
$$K(\mathbf{u},\mathbf{v}) = \exp\left(-\frac{||\mathbf{u}-\mathbf{v}||_2^2}{2\sigma^2}\right)$$

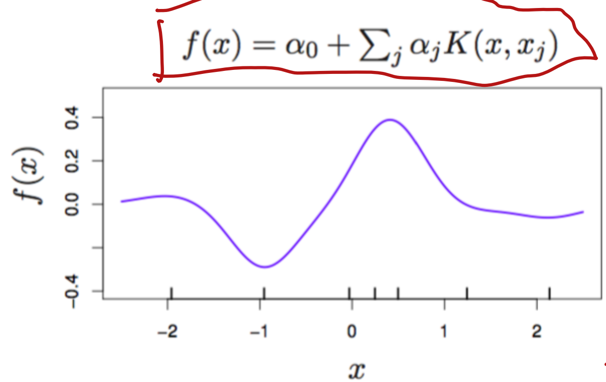
Not an "obvious" map for which this is the kernel

ϕ

Like weighting "bumps" on each point like kernel smoothing but now we learn the weights

$$\widehat{f}(x) = \sum_{i=1}^{n} \widehat{\alpha}_i K(x_i, x)$$



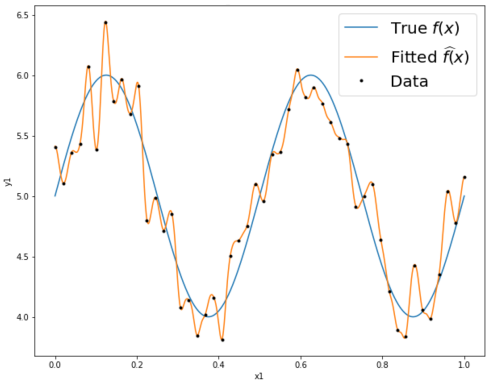


RBF Kernel

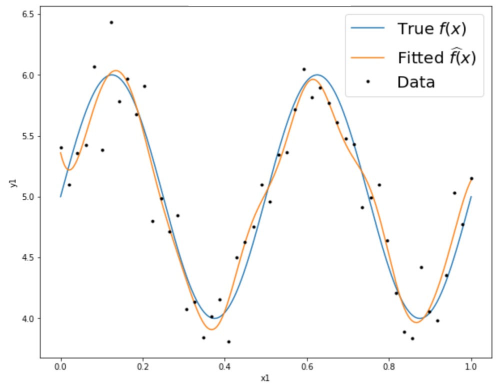
$$K(\mathbf{u},\mathbf{v}) = \exp\left(-\frac{||\mathbf{u}-\mathbf{v}||_2^2}{2\sigma^2}\right)$$

The bandwidth sigma has an enormous effect on fit:

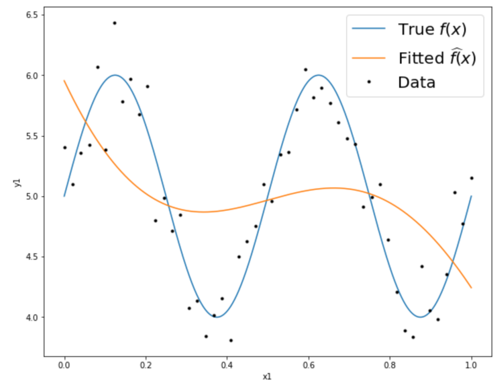
.01 $\sigma = 10^{-2}$ $\lambda = 10^{-4}$



.1 $\sigma = 10^{-1}$ $\lambda = 10^{-4}$



1 $\sigma = 10^{-0}$ $\lambda = 10^{-4}$



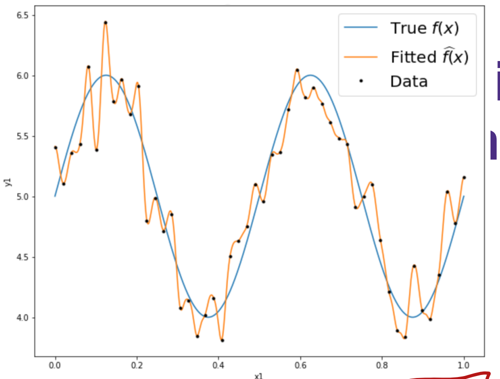
$$\widehat{f}(x) = \sum_{i=1}^{n} \widehat{\alpha}_i K(x_i, x)$$

RBF Kernel

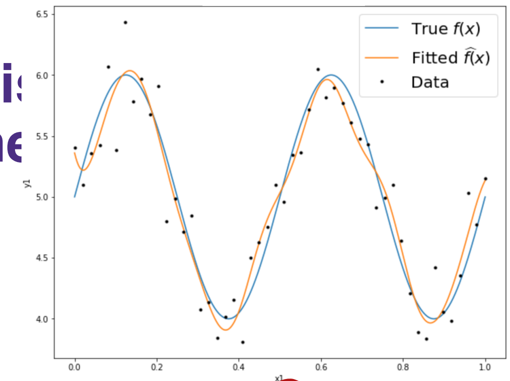
$$K(\mathbf{u},\mathbf{v}) = \exp\left(-\frac{||\mathbf{u}-\mathbf{v}||_2^2}{2\sigma^2}\right)$$

The bandwidth sigma has an enormous effect on fit:

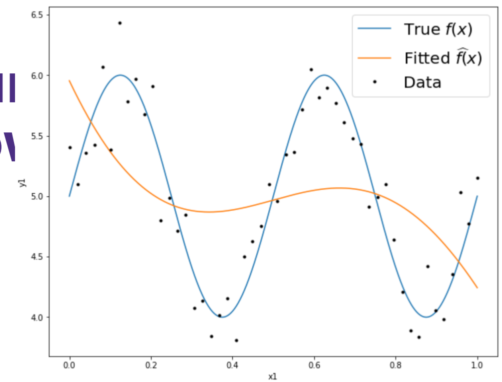
$\sigma = 10^{-2} \quad \lambda = 10^{-4}$



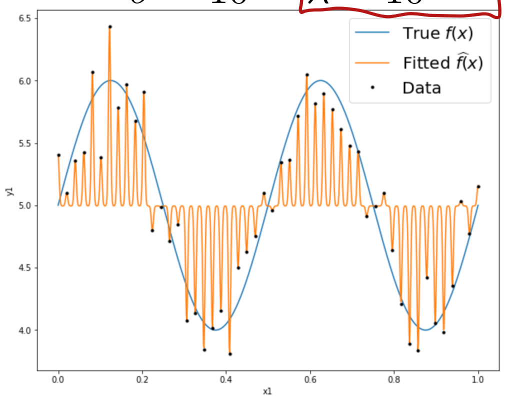
$\sigma = 10^{-1} \quad \lambda = 10^{-4}$



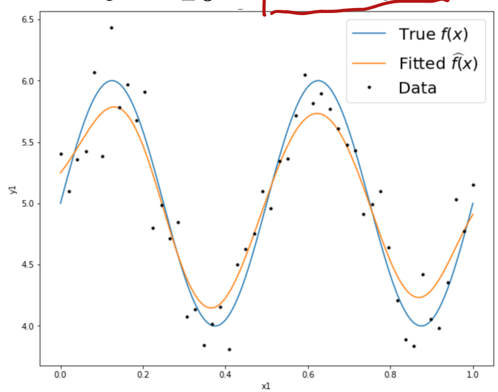
$\sigma = 10^{-0} \quad \lambda = 10^{-4}$



$\sigma = 10^{-3} \quad \boxed{\lambda = 10^{-4}}$



$\sigma = 10^{-1} \quad \boxed{\lambda = 10^{-0}}$

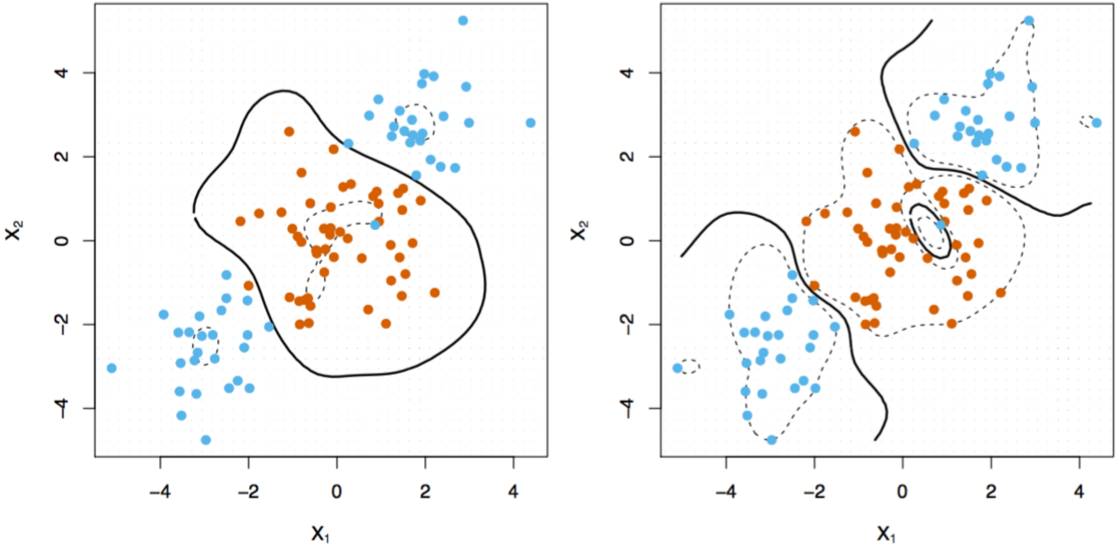


$$\widehat{f}(x) = \sum_{i=1}^{n} \widehat{\alpha}_i K(x_i, x)$$

RBF Classification

$$\widehat{w} = \sum_{i=1}^{n} \max\{0, 1 - y_i(b + x_i^T w)\} + \lambda ||w||_2^2$$

$$\min_{\alpha,b} \sum_{i=1}^{n} \max\{0, 1 - y_i(b + \sum_{j=1}^{n} \alpha_j \langle x_i, x_j \rangle)\} + \lambda \sum_{i,j=1}^{n} \alpha_i \alpha_j \langle x_i, x_j \rangle$$



RBF Kernel

$$K(\mathbf{u},\mathbf{v}) = \exp\left(-\frac{||\mathbf{u}-\mathbf{v}||_2^2}{2\sigma^2}\right)$$

$\in \mathbb{R}^d$

$K(u,v) = \phi(u)\cdot\phi(v)$

$\mathbb{R}^\infty$

Basis representation in 1d?

$$\boxed{[\phi(x)]_i = \frac{1}{\sqrt{i!}}e^{-\frac{x^2}{2}}x^i \quad \text{for } i = 0, 1, \ldots}$$

$$\phi(x)^T\phi(x') = \sum_{i=0}^{\infty}\left(\frac{1}{\sqrt{i!}}e^{-\frac{x^2}{2}}x^i\right)\left(\frac{1}{\sqrt{i!}}e^{-\frac{(x')^2}{2}}(x')^i\right)$$

$$= e^{-\frac{x^2+(x')^2}{2}}\sum_{i=0}^{\infty}\frac{1}{i!}(xx')^i$$

$$= e^{-|x-x'|^2/2}$$

Wait, infinite dimensions?

- Isn't everything separable there? How are we not overfitting?

$x \to \phi(x)$

$\max \|x\|_2^2$

- Regularization! Fat shattering (R/margin)^2