Recitation 5

Kernels

Brett Bernstein

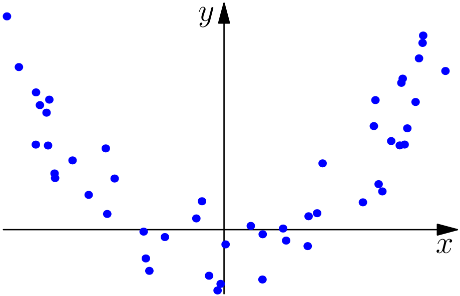
CDS at NYU

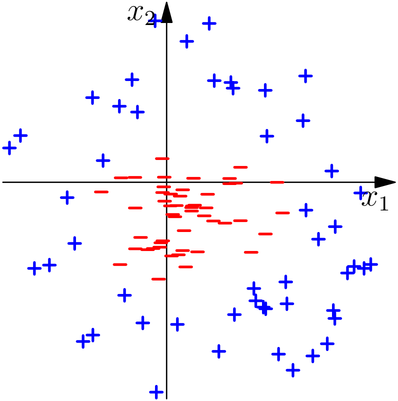
February 23, 2017

Intro Question

Question

Consider applying linear regression to the data set on the left, and an SVM to the data set on the right. What is the issue? Can it be improved?





Intro Solution

Regression Solution

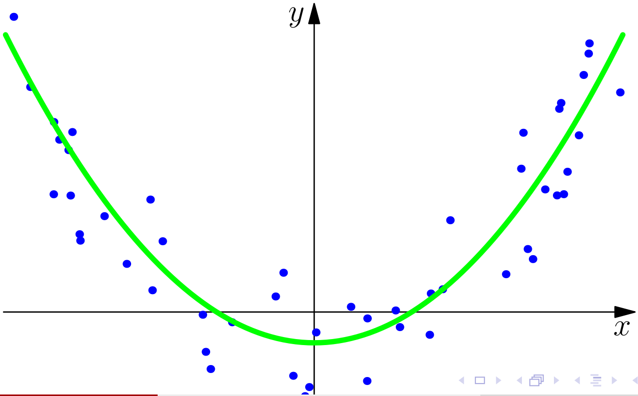
We want to allow for non-linear regression functions, but we would like to reuse the same fitting procedures we have already developed. To do this we will expand our feature set by adding non-linear functions of old features. We change our features from $(1, x)$ to $(1, x, x^2)$. That is

$$X = \begin{pmatrix} 1 & -1 \\ 1 & -.7 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \implies \Phi = \begin{pmatrix} 1 & -1 & (-1)^2 \\ 1 & -.7 & (-.7)^2 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 1^2 \end{pmatrix}.$$

Intro Solution

Regression Solution

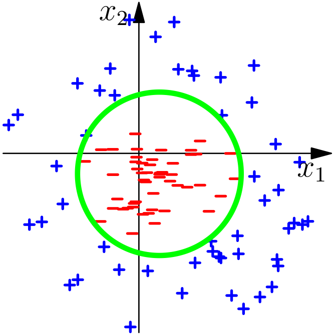
Using features $(1, x, x^2)$ and $w = (-.1, 0, 1)$ gives us $f_w(x) = -.1 + 0x + 1x^2 = x^2 - .1$. Our prediction function is quadratic but we obtained it through standard linear methods.



Intro Solution

SVM Solution

For the SVM we expand our feature vector from $(1, x_1, x_2)$ to $(1, x_1, x_2, x_1x_2, x_1^2, x_2^2)$. Using $w = (-1.875, 2.5, -2.5, 0, 1, 1)$ gives $-1.875 + 2.5x_1 - 2.5x_2 + x_1^2 + x_2^2 = (x_1 + 1.25)^2 + (x_2 - 1.25)^2 - 5 = 0$ as our decision boundary.



Cost of Adding Features

Question

Suppose we begin with d features (and a bias) $x = (1, x_1, \ldots, x_d)$. We add all monomials up to degree M. More precisely, all terms of the form $x_1^{p_1} \cdots x_d^{p_d}$ where $p_i \geq 0$ and $p_1 + \cdots + p_d \leq M$. How many features will we have in total?

Cost of Adding Features

Question

Suppose we begin with d features (and a bias) $x = (1, x_1, \ldots, x_d)$. We add all monomials up to degree M. More precisely, all terms of the form $x_1^{p_1} \cdots x_d^{p_d}$ where $p_i \geq 0$ and $p_1 + \cdots + p_d \leq M$. How many features will we have in total?

Solution

There will be $\binom{M+d}{M}$ terms total. If M is fixed and we let d grow, this behaves like $\frac{d^M}{M!}$. For example, if $d = 40$ and $M = 8$ we get $\binom{40+8}{8} = 377348994$. If we are training or predicting with a linear model $w^T x$, this product now takes $O(d^M)$ operations to evaluate.

Kernel Trick

Consider the polynomial kernel $k(x, y) = (1 + x^T y)^M$ where $x, y \in \mathbb{R}^d$. This computes the inner product of all monomials up to degree M in time $O(d)$. For example, if $M = 2$ we have

$$\begin{aligned}(1 + x^T y)^2 &= 1 + 2x^T y + x^T y x^T y \\ &= 1 + 2\sum_{i=1}^d x_i y_i + \sum_{i,j=1}^d x_i y_i x_j y_j.\end{aligned}$$

The resulting feature map is

$$\varphi(x) = (1, \sqrt{2}x_1, \ldots, \sqrt{2}x_d, x_1^2, \ldots, x_d^2, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3, \ldots, \sqrt{2}x_{d-1}x_d).$$

Then $k(x, y) = \varphi(x)^T \varphi(y)$.

Kernel Ridge Regression

- Recall that our ridge regression loss is given by

$$J(w) = \frac{1}{n}\sum_{i=1}^n (w^T x_i - y_i)^2 + \lambda\|w\|_2^2$$

- If we map to a larger feature space $\varphi : \mathbb{R}^d \to \mathbb{R}^D$ we get

$$J(\tilde{w}) = \frac{1}{n}\sum_{i=1}^n (\tilde{w}^T \varphi(x_i) - y_i)^2 + \lambda\|\tilde{w}\|_2^2.$$

- Using the kernel trick we can try to write this (incorrectly) as

$$J(w) = \frac{1}{n}\sum_{i=1}^n (k(w, x_i) - y_i)^2 + \lambda k(w, w).$$

- What are the issues with this?

Kernel Ridge Regression

Issues with

$$J(w) = \frac{1}{n}\sum_{i=1}^{n}(k(w,x_i) - y_i)^2 + \lambda k(w,w)$$

- Writing $\tilde{w}^T\varphi(x_i)$ isn't the same as $k(w,x_i) = \varphi(w)^T\varphi(x_i)$ since φ isn't onto. That is, $\varphi(w)$ is a very specific type of element of $\mathbb{R}^D$, the larger feature space.
- The $L(w)$ written above isn't a ridge regression problem any more, since $k(w,x_i)$ and $k(w,w)$ can be weird functions of w. Thus our previous code and theory for dealing with ridge regression doesn't immediately carry over.

Math Thought Experiment

- Suppose we have a standard ridge regression problem for $w \in \mathbb{R}^d$:

$$J(w) = \frac{1}{n}\sum_{i=1}^{n}(w^T x_i - y_i)^2 + \lambda\|w\|_2^2.$$

- Now suppose we add a new feature to x that is always zero: $\tilde{x} = (x, 0)$. For $w \in \mathbb{R}^{d+1}$ we now have

$$J(w) = \frac{1}{n}\sum_{i=1}^{n}(w^T \tilde{x}_i - y_i)^2 + \lambda\|w\|_2^2.$$

  Does the answer change? In other words, will the new minimizer $w_* \in \mathbb{R}^{d+1}$ have a non-zero value in its last coordinate?

Math Thought Experiment

Does the answer change for $w \in \mathbb{R}^{d+1}$ with

$$J(w) = \frac{1}{n}\sum_{i=1}^{n}(w^T\tilde{x}_i - y_i)^2 + \lambda\|w\|_2^2$$

No!

- Suppose $w_* = (v, a)$ where $v \in \mathbb{R}^d$ and $a \in \mathbb{R}$.
- Recall that $\tilde{x}_i = (x, 0)$.
- Then $w_*^T\tilde{x}_i = v^Tx$. In other words, a has **no** effect on the prediction $w_*^T\tilde{x}_i$.
- But if $a \neq 0$ then $\|(v, a)\|_2^2 > \|(v, 0)\|_2^2$.
- Can you think of the more general version of this phenomenon?

Representer Theorem (Baby Version)

Theorem ((Baby) Representer Theorem)

Suppose you have a loss function of the form

$$J(w) = L(w^T\varphi(x_1), \ldots, w^T\varphi(x_n)) + R(\|w\|_2)$$

where

- $w, \varphi(x_i) \in \mathbb{R}^D$.
- $L : \mathbb{R}^n \to \mathbb{R}$ *is an arbitrary function (loss term).*
- $R : \mathbb{R}_{\geq 0} \to \mathbb{R}$ *is increasing (regularization term).*

Assume J has at least one minimizer. Then J has a minimizer w^* of the form $w^* = \sum_{i=1}^n \alpha_i\varphi(x_i)$ for some $\alpha \in \mathbb{R}^n$. If R is strictly increasing, then all minimizers have this form.

Representer Theorem: Proof

Proof.

- Let $w^* \in \mathbb{R}^D$ and let $S = \text{Span}(\varphi(x_1), \ldots, \varphi(x_n))$.
- Write $w^* = u + v$ where $u \in S$ and $v \in S^\perp$. Here u is the orthogonal projection of w^* onto S, and $S^\perp$ is the subspace of all vectors orthogonal to S.
- Then $(w^*)^T\varphi(x_i) = (u+v)^T\varphi(x_i) = u^T\varphi(x_i) + v^T\varphi(x_i) = u^T\varphi(x_i)$.
- But $\|w^*\|_2^2 = \|u+v\|_2^2 = \|u\|_2^2 + \|v\|_2^2 + 2u^Tv = \|u\|_2^2 + \|v\|_2^2 \geq \|u\|_2^2$.
- Thus $R(\|w^*\|_2) \geq R(\|u\|_2)$ showing $J(w^*) \geq J(u)$.

$\square$

- Above we showed that $\|u+v\|_2^2 = \|u\|_2^2 + \|v\|_2^2$ when $u^Tv = 0$. This is called the Pythagorean theorem.

Representer Theorem: Meaning

- If your loss function only depends on w via its inner products with the inputs, and the regularization is an increasing function of the ℓ_2 norm, then we can write w^* as a linear combination of the training data.
- This applies to ridge regression and SVM.

Question

Suppose you have $n = 100$ samples, $d = 40$ features, and $M = 8$ degree monomial terms giving 377348994 features. This implies $w \in \mathbb{R}^{377348994}$ for ridge regression. What does the representer theorem say?

Representer Theorem: Meaning

- If your loss function only depends on w via its inner products with the inputs, and the regularization is an increasing function of the ℓ_2 norm, then we can write w^* as a linear combination of the training data.
- This applies to ridge regression and SVM.

Question

Suppose you have $n = 100$ samples, $d = 40$ features, and $M = 8$ degree monomial terms giving 377348994 features. This implies $w \in \mathbb{R}^{377348994}$ for ridge regression. What does the representer theorem say?

Solution

As $y \in \mathbb{R}^n$ varies, the solution w must lie in a 100-dimensional subspace of $\mathbb{R}^{377348994}$.

Representer Theorem: Ridge Regression

- By adding features to ridge regression we had

$$\begin{aligned} J(\tilde{w}) &= \frac{1}{n}\sum_{i=1}^{n}(\tilde{w}^T\varphi(x_i) - y_i)^2 + \lambda\|\tilde{w}\|_2^2 \\ &= \frac{1}{n}\|\Phi\tilde{w} - y\|_2^2 + \lambda\tilde{w}^T\tilde{w}, \end{aligned}$$

  where $\Phi \in \mathbb{R}^{n\times D}$ is the matrix with $\varphi(x_i)^T$ as its ith row.
- Representer Theorem applies giving $\tilde{w} = \sum_{j=1}^{n}\alpha_j\varphi(x_j) = \Phi^T\alpha$.
- Plugging in gives

$$J(\alpha) = \frac{1}{n}\left\|\Phi\Phi^T\alpha - y\right\|_2^2 + \lambda\alpha^T\Phi\Phi^T\alpha.$$

Representer Theorem: Ridge Regression

- Let $K \in \mathbb{R}^{n \times n}$ be given by $K = \Phi\Phi^T$. This is called the **Gram Matrix** and satisfies $K_{ij} = k(x_i, x_j) = \varphi(x_i)^T\varphi(x_j)$:

$$K = \begin{pmatrix} \varphi(x_1)^T\varphi(x_1) & \cdots & \varphi(x_1)^T\varphi(x_n) \\ \vdots & \ddots & \vdots \\ \varphi(x_n)^T\varphi(x_1) & \cdots & \varphi(x_n)^T\varphi(x_n) \end{pmatrix}.$$

- We can write ridge regression in the kernelized form by turning

$$J(\alpha) = \frac{1}{n}\left\|\Phi\Phi^T\alpha - y\right\|_2^2 + \lambda\alpha^T\Phi\Phi^T\alpha.$$

  into

$$J(\alpha) = \frac{1}{n}\|K\alpha - y\|_2^2 + \lambda\alpha^T K\alpha.$$

- Can derive the solution algebraically (see Homework 4).
- Prediction function is $f_\alpha(x) = (w^*)^T\varphi(x) = \sum_{i=1}^n \alpha_i k(x_i, x)$.

Representer Theorem: Primal SVM

- For a general linear model, the same derivation above shows
$$J(w) = L(\Phi w) + R(\|w\|_2)$$
becomes
$$J(\alpha) = L(K\alpha) + R(\sqrt{\alpha^T K\alpha}).$$
Here $\varphi(x_i)^T w$ became $(K\alpha)_i$.
- The primal SVM (bias in features) has loss function
$$J(w) = \frac{c}{n}\sum_{i=1}^n (1 - y_i(\varphi(x_i)^T w))_+ + \|w\|_2^2.$$
- This is kernelized to
$$J(\alpha) = \frac{c}{n}\sum_{i=1}^n (1 - y_i(K\alpha)_i)_+ + \|w\|_2^2.$$
- Positive decision made if $(w^*)^T\varphi(x) = \sum_{i=1}^n \alpha_i k(x_i, x) > 0$.

Dual SVM

- The dual SVM problem (with features) is given by

$$
\begin{aligned}
\text{maximize}_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \varphi(x_i)^T \varphi(x_j) \\
\text{subject to} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0 \\
& \alpha_i \in \left[0, \frac{c}{n}\right] \quad \text{for } i = 1, \ldots, n.
\end{aligned}
$$

- We can immediately kernelize (no representer theorem needed) by replacing $\varphi(x_i)^T \varphi(x_j) = k(x_i, x_j)$.
- Recall that we were able to derive the conclusion of the representer theorem using strong duality for SVMs.

RBF Kernel

- As we saw last time, the most frequently used kernel is the RBF kernel

$$k(w, x) = \exp\left(-\frac{\|w - x\|_2^2}{2\sigma^2}\right).$$

- Is there a corresponding feature map $\varphi : \mathbb{R}^d \to \mathbb{R}^D$ so that $k(w, x) = \varphi(w)^T \varphi(x)$?
- Unfortunately there is no finite D that will work.
- We will handle this by allowing infinite dimensional spaces called Hilbert spaces.

Rough Review of Abstract Linear Algebra

- Vector Spaces: Spaces where linear combinations (scaling and adding) make sense.
- A vector space has no sense of length or angle, but has concepts like subspace, basis, span, dimension, linear transformation, eigenvector.
- We can account for length and angle by looking at inner products.
- Recall that in 2d, there is a law of cosines that relates side lengths to the angles of a triangle. In other words, reasonable definitions of length and angle are not independent.

Definition of Inner Product

Definition (Inner Product)

Let V be a vector space. We say $\langle \cdot, \cdot \rangle : V^2 \to \mathbb{R}$ is an **inner product** if it satisfies the following:

1. Symmetry:
$$\langle v, w \rangle = \langle w, v \rangle$$
for $v, w \in V$.
2. Bilinearity:
$$\langle \alpha v_1 + \beta v_2, w \rangle = \alpha \langle v_1, w \rangle + \beta \langle v_2, w \rangle$$
for $v_1, v_2, w \in\in V$ and $\alpha, \beta \in \mathbb{R}$. Same holds for second argument by symmetry.
3. Positive-Definiteness: $\langle v, v \rangle \geq 0$ with $\langle v, v \rangle = 0$ if and only if $v = 0$.

If we associate an inner product with a vector space we call the pair an **inner product space**.

More on Inner Products and Norms

- On $\mathbb{R}^n$ the "dot-product" $v^T w$ is an inner product.
- Using an inner product we can obtain the norm (length) of a vector: $\|v\| = \sqrt{\langle v, v\rangle}$. We obtain the angle via $\cos(\angle(v, w)) := \frac{\langle v, w\rangle}{\|v\|\|w\|}$.
- We say v, w are orthogonal if $\langle v, w\rangle = 0$.
- Not all norms we have seen have an associated inner product. For example, $\|v\|_1$ has no associated inner product.
- Roughly, certain norms (like ℓ_1) do not lead to a consistent way of measuring angles. More precisely:

Theorem (Parallelogram Law)

A norm $\|\cdot\|$ has an associated inner product if and only if

$$2\|v\|^2 + 2\|w\|^2 = \|v - w\|^2 + \|v + w\|^2.$$

Hilbert Space

- Once we have a norm, we can talk about the distance between two vectors: $\|v - w\|$. This allows one to introduce concepts such as convergence and continuity.
- Certain facts we need about projections are false in infinite dimensions unless we impose an extra constraint on our inner product spaces: completeness.
- A space is complete if all Cauchy sequences converge.

Definition (Hilbert Space)

An inner product space V with inner product $\langle \cdot, \cdot \rangle$ is called a **Hilbert space** if it is complete with respect to the associated norm.

- All finite dimensional inner product spaces are Hilbert space.

Projection Theorem

Theorem (Projection Theorem)

Let H be a Hilbert space and let S be a finite dimensional subspace. Then any vector $w \in H$ can be written uniquely as $w = u + v$ where $u \in S$ and $v \in S^{\perp}$.

More can be said:

- u is the projection of w onto S: $u = \arg\min_{x \in S} \|x - w\|$.
- Can extend theorem to hold for any closed subspace (finite dimensional subspaces are examples of closed subspaces).
- A variant of the theorem holds for non-empty closed convex subsets.

Representer Theorem (Adult Version)

Theorem (Representer Theorem)

Suppose you have a loss function of the form

$$J(w) = L(\langle w, \varphi(x_1)\rangle, \ldots, \langle w, \varphi(x_n)\rangle) + R(\|w\|)$$

where

- $w, \varphi(x_i) \in H$ *for some Hilbert space* H*.*
- $L : \mathbb{R}^n \to \mathbb{R}$ *is an arbitrary function (loss term).*
- $R : \mathbb{R}_{\geq 0} \to \mathbb{R}$ *is increasing (regularization term).*
- $\|\cdot\|$ *is the norm associated with* H*.*

Assume J *has at least one minimizer. Then* J *has a minimizer* w^* *of the form* $w^* = \sum_{i=1}^n \alpha_i \varphi(x_i)$ *for some* $\alpha \in \mathbb{R}^n$*. If* R *is strictly increasing, then all minimizers have this form.*

Representer Theorem (Adult Version)

- Same proof as before, but apply the projection theorem and use $\langle v, w\rangle$ in place of $v^T w$.
- Now we have a representer theorem that works for infinite dimensional Hilbert spaces. Why do we care?
- We will show that kernels, such as the RBF Kernel, correspond to inner products in Hilbert spaces.

Positive Semi-Definite

Definition (Positive Semi-Definite)

A matrix $A \in \mathbb{R}^{n \times n}$ is **positive semi-definite** if it is symmetric and

$$x^T A x \geq 0$$

for all $x \in \mathbb{R}^n$.

- Equivalent to saying the matrix is symmetric with non-negative eigenvalues.

Mercer's Theorem

Theorem (Mercer's Theorem)

Fix a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. There is a Hilbert space H and a feature map $\varphi : \mathcal{X} \to H$ such that $k(x, y) = \langle \varphi(x), \varphi(y) \rangle_H$ if and only if for any $x_1, \ldots, x_n \in \mathcal{X}$ the associated matrix K is positive semi-definite:

$$K = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix}.$$

Such a kernel k is called **positive semi-definite**.

Finding Your Own Kernels

Let $k_1, k_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be positive semi-definite kernels. Then so are the following:

- $k_3(w, x) = k_1(w, x) + k_2(w, x)$
- $k_4(w, x) = \alpha k_1(w, x)$ for $\alpha \geq 0$
- $k_5(w, x) = f(w)f(x)$ for any function $f : \mathcal{X} \to \mathbb{R}$
- $k_6(w, x) = k_1(w, x)k_2(w, x)$

Furthermore, if $k_1, k_2, \ldots$ is a sequence of positive semi-definite kernels then

$$k(w, x) = \lim_{n \to \infty} k_n(w, x)$$

is also positive semi-definite (assuming the limit exists for all w, x).

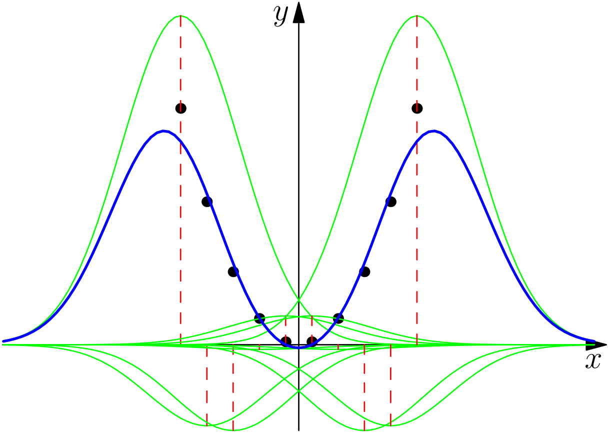
Representer Theorem for RBF Kernels

- As we saw earlier for ridge regression and SVM classification, the decision function has the form $f_\alpha(x) = \sum_{i=1}^n \alpha_i k(x_i, x)$.
- For ridge regression, this means that using the RBF kernel amounts to approximating our data by a linear combination of Gaussian bumps.
- For SVM classification, each $k(x_i, x) = \exp\left(-\|x_i - x\|_2^2/(2\sigma^2)\right)$ represents a exponentially decaying distance between x_i and x. Thus our decisions depend on our proximities to data points.

RBF Regression

- Below we use 10 uniformly spaced x-values between -2 and 2, with $y_i = x_i^2$. We fit kernelized ridge regression with the RBF kernel using $\sigma = 1$ and $\lambda = .1$.
- Each green curve is $g(x) = \alpha_i k(x_i, x)$. The predicted function is drawn in blue.
- As you might expect, extrapolating outside of $[-2, 2]$ can have poor results.
- People will often normalize the RBF kernel (see Hastie, Tibshirani, Friedman p. 213).

RBF Regression

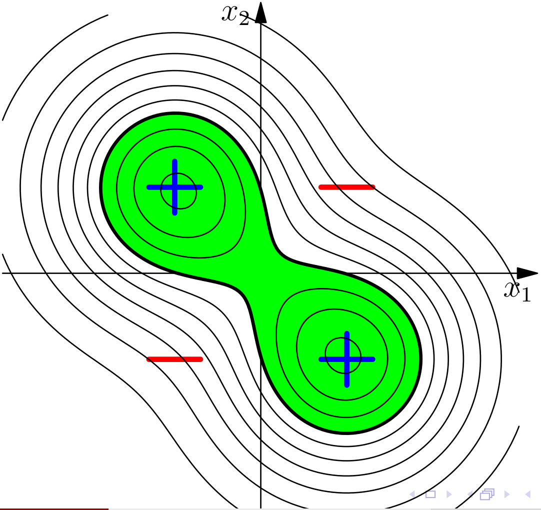


RBF Classification

- Next we show 4 points placed on the corners of a square with positive and negative points on each diagonal.

RBF Classification

Contours of $f(x) = k(x_1, x) + k(x_2, x)$ where x_1, x_2 are positive examples, and $\sigma = 1$.



RBF Classification

Contours of $f(x) = k(x_1, x) + k(x_2, x) - k(x_3, x) - k(x_4, x)$ where x_1, x_2 are positive examples, and $\sigma = 1$.