Topics in Learning Theory

Lecture 6: Kernel Methods (I)

Topics

- From 2-norm regularization to kernel methods
- Mercel's Theorem Reproducing kernel Hilbert space (RKHS)
- Learning in RKHS
- Example Kernels and Corresponding Feature Representation

Empirical risk minimization with 2-norm regularization

- Consider 2-norm regularized empirical risk minimization formulation:
  - map input x to high-dimensional feature $\psi(x)$
  - scoring function $f(x) = w^T\psi(x) + b$ (b is optional) .

$$[\hat{w}, \hat{b}] = \arg\min_{w,b} \left[\sum_{i=1}^{n} \phi(w^T\psi(X_i) + b, Y_i) + \lambda w^T w \right] \qquad (I)$$

- Solution:

$$\hat{w} = -\frac{1}{2\lambda} \sum_{i=1}^{n} \phi_1'(\hat{w}^T\psi(X_i) + \hat{b}, Y_i)\psi(X_i).$$

- Scoring function:

$$f(x) = \hat{w}^T x + \hat{b} = \sum_{i=1}^{n} \hat{\alpha}_i \psi(X_i)^T \psi(x) + \hat{b} = \sum_{i=1}^{n} \hat{\alpha}_i k(x, X_i) + \hat{b},$$

where

$$\begin{aligned} \hat{\alpha}_i =& -\phi_1'(\hat{w}^T \psi(X_i) + \hat{b}, Y_i)/(2\lambda), \\ k(x, x') =& \psi(x)^T \psi(x'). \end{aligned}$$

and

$$\hat{w}^T \hat{w} = \sum_{i,j} \hat{\alpha}_i \hat{\alpha}_j k(x_i, x_j) = \hat{\alpha}^T K \hat{\alpha},$$

- $k(x, x')$: kernel, and K: kernel gram matrix.

Primal kernel learning formulation

- Primal kernel formulation:

$$[\hat{\alpha}, \hat{b}] = \arg\min_{\alpha, b} \left[\sum_{i=1}^{n} \phi\left([K\alpha]_i + b, Y_i\right) + \lambda \alpha^T K \alpha \right] \qquad (II)$$

(where $[K\alpha]_i = \sum_{j=1}^{n} \alpha_j k(X_i, X_j)$) with scoring function

$$f(x) = \sum_{i=1}^{n} \hat{\alpha}_i k(x, X_i) + b.$$

- If a kernel function can be represented as $k(x, x') = \psi(x)^T \psi(x')$, then (I) and (II) are equivalent.

Kernel learning formulation: interpretation

- Working with kernel or its implicit feature space is equivalent.
- Reduces high dimensional learning problem to problem in R^n.
- Each coefficients corresponding to a sample-point.
- 2-norm regularization in w to quadratic regularization in α.
  - K has to be positive (semi)-definite.
- Replaces linear combination in high dimensional features by linear combintation of kernel functions evaluated at the data points.

Sparsity of dual parameter

- Primal formulation: $[\hat{w}, \hat{b}] = \arg\min_{w,b} \left[\sum_{i=1}^{n} \phi(w^T \psi(X_i) + b, Y_i) + \lambda w^T w\right]$
- Solution: $f(x) = \sum_{i=1}^{n} \hat{\alpha}_i k(x, X_i) + \hat{b}$, $\hat{\alpha}_i = -\phi_1'(\hat{w}^T \psi(X_i) + \hat{b}, Y_i)/(2\lambda)$
- $\hat{\alpha}_i = 0$ when $\phi_1'(\hat{w}^T \psi(X_i) + \hat{b}, Y_i) = 0$
- SV classification with hinge loss $\phi(f_i, y_i) = (1 - f_i y_i)_+$
  - $\phi_1'(f_i, y_i) = 0$ when $f_i y_i > 1$.
- SV regression with ϵ-insensitive loss $\phi(f_i, y_i) = (|f_i - y_i| - \epsilon)_+$
  - $\phi_1'(f_i, y_i) = 0$ when $|f_i - y_i| < \epsilon$.

Comment: more general ways to use kernel

- Primal kernel formulation: $f(x) = \sum_{i=1}^{n} \hat{\alpha}_i k(x, X_i) + b$

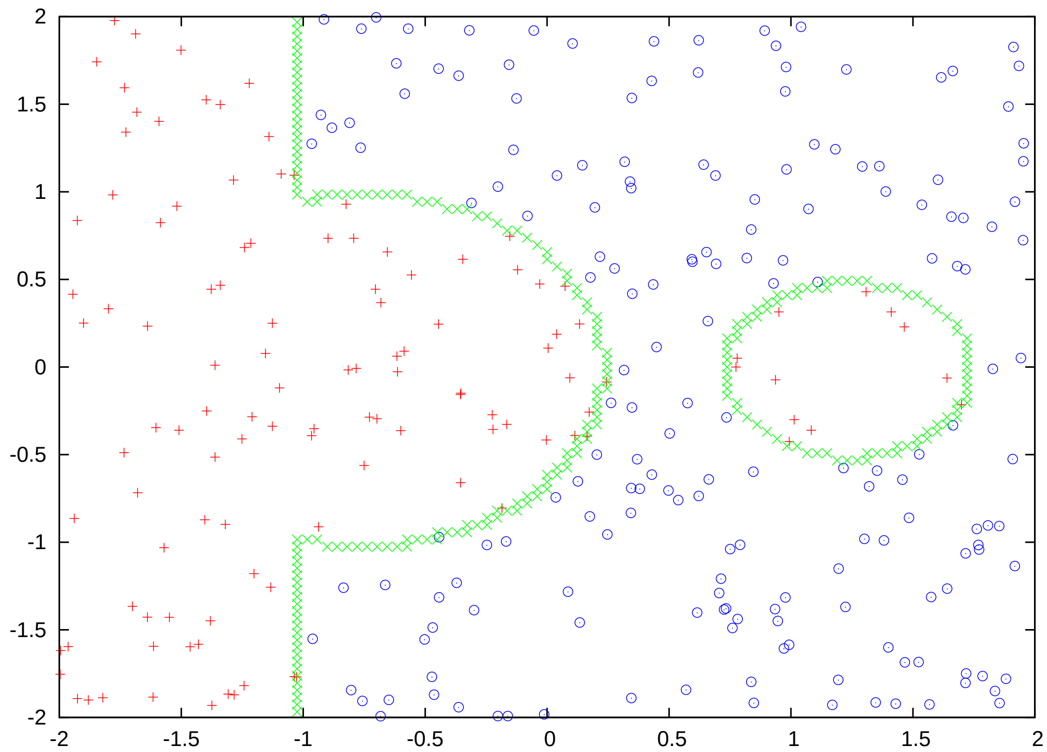
$$[\hat{\alpha}, \hat{b}] = \arg\min_{\alpha, b} \left[\sum_{i=1}^{n} \phi\left([K\alpha]_i + b, Y_i\right) + \lambda \alpha^T K \alpha \right]$$

- Treat kernel as features, and replace $\alpha^T K \alpha$ by other regularization condition on α: $\|\alpha\|_1$ or $\|\alpha\|_2$.
  - advantage: no need to require K positive definite.

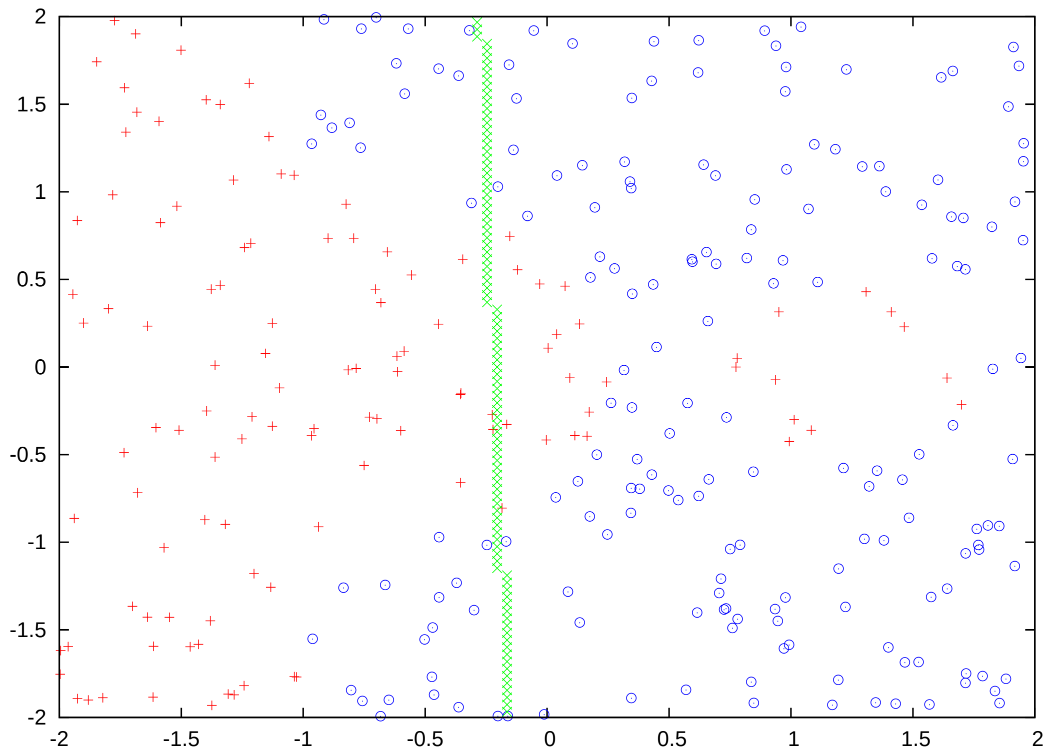
Two common kernels for high dimensional data

- Polynomial kernel with degree p: $k(x, x') = (1 + x^T x')^p$
- RBF (radial basis function) kernel: $k(x, x') = \exp(-\|x - x'\|_2^2 / 2\sigma^2)$.

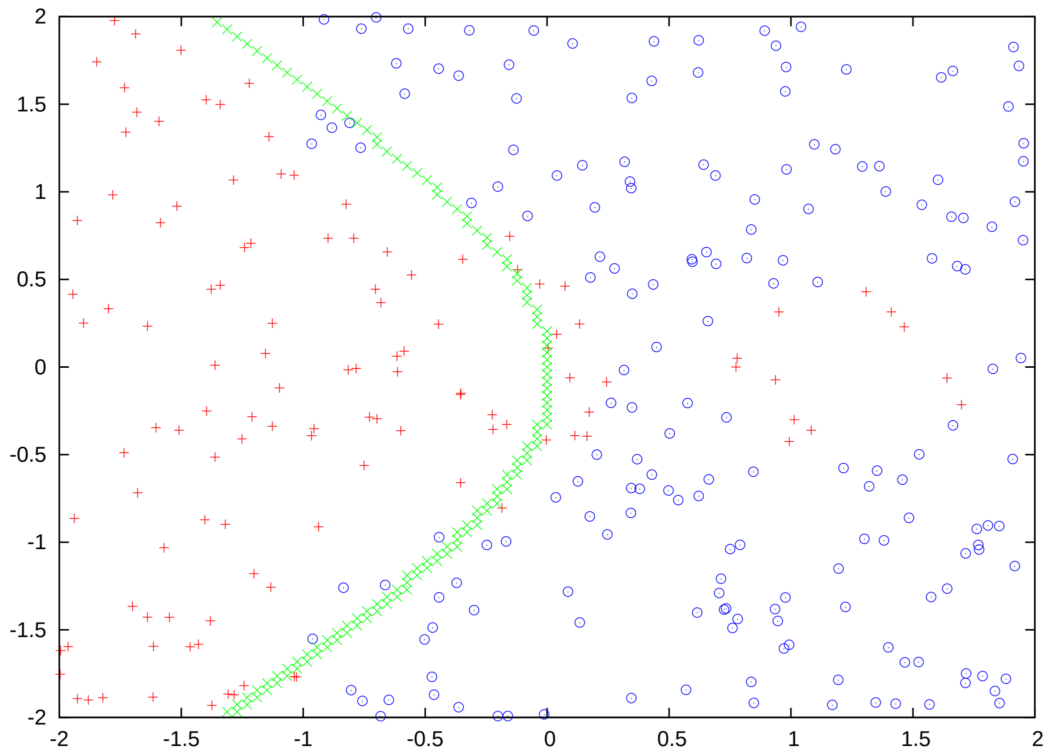
Example: true boundary



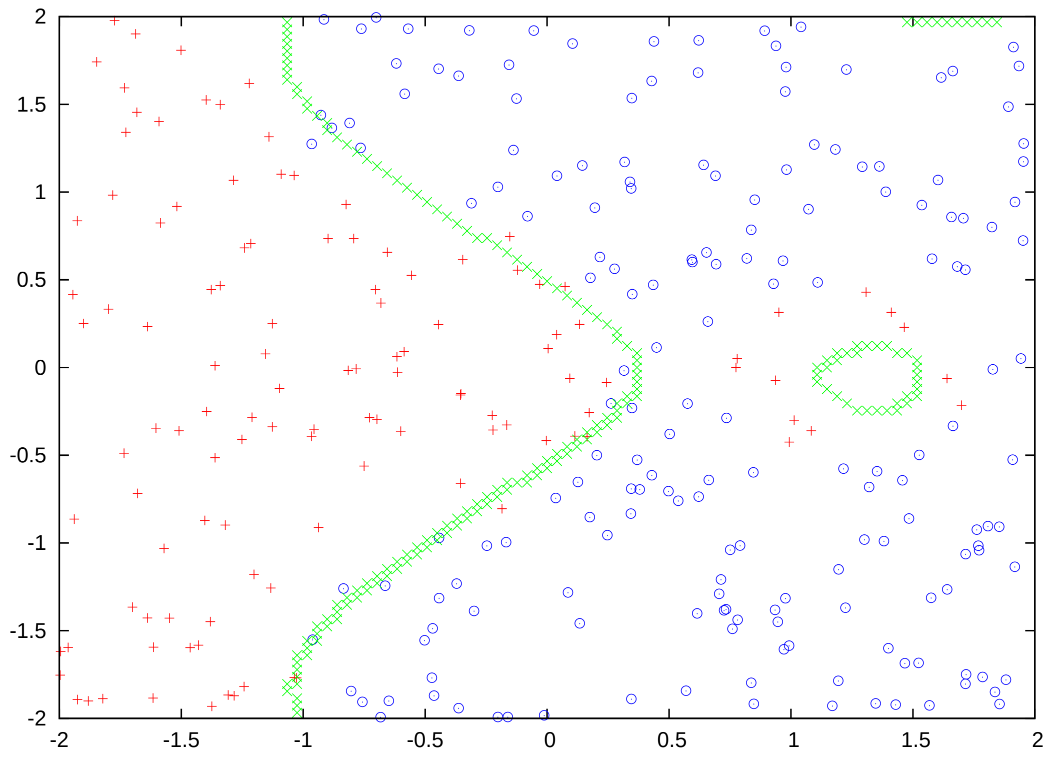
LS with Polynomail kernel: $p = 1$



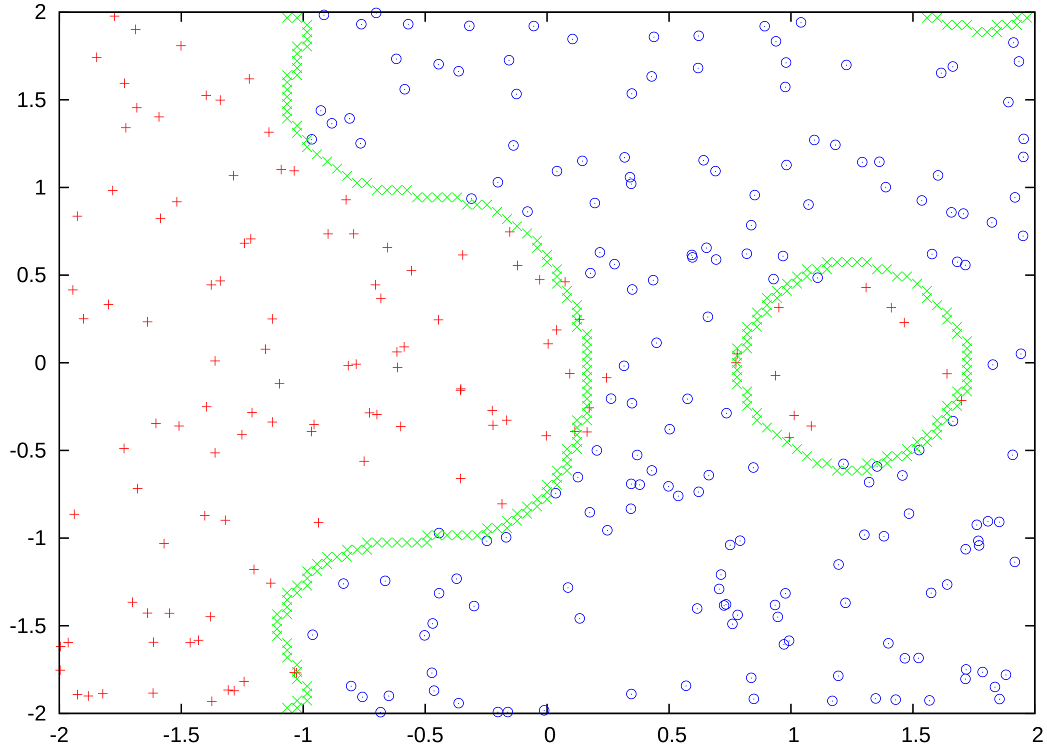
LS with Polynomail kernel: $p = 2$



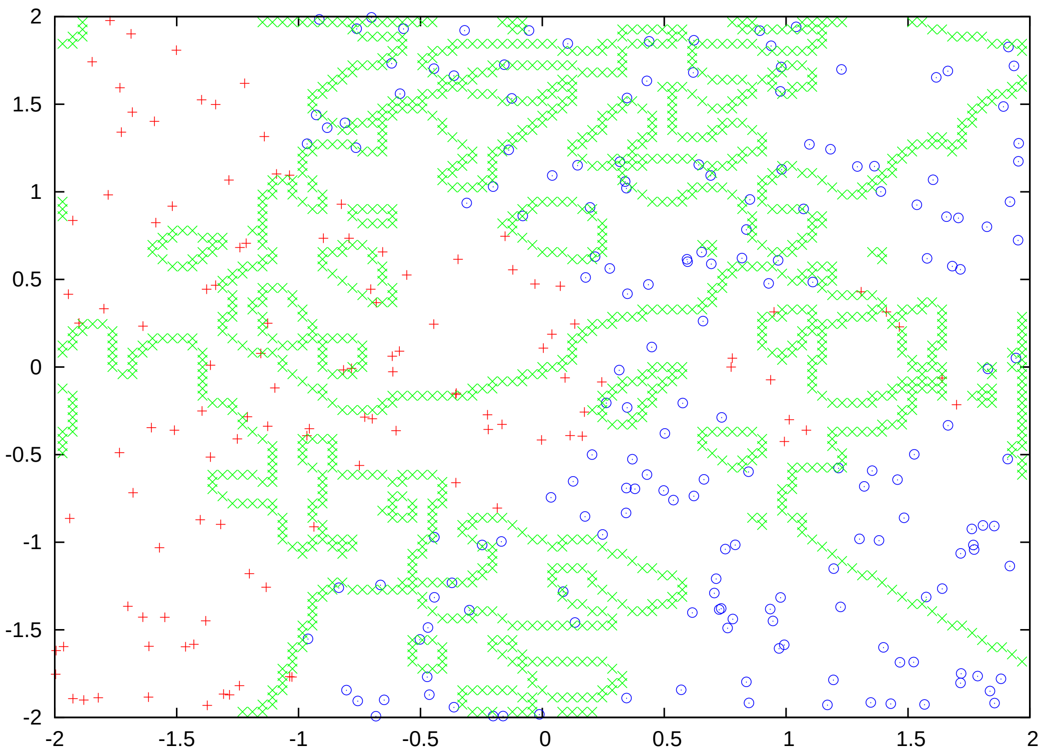
LS with Polynomail kernel: $p = 5$



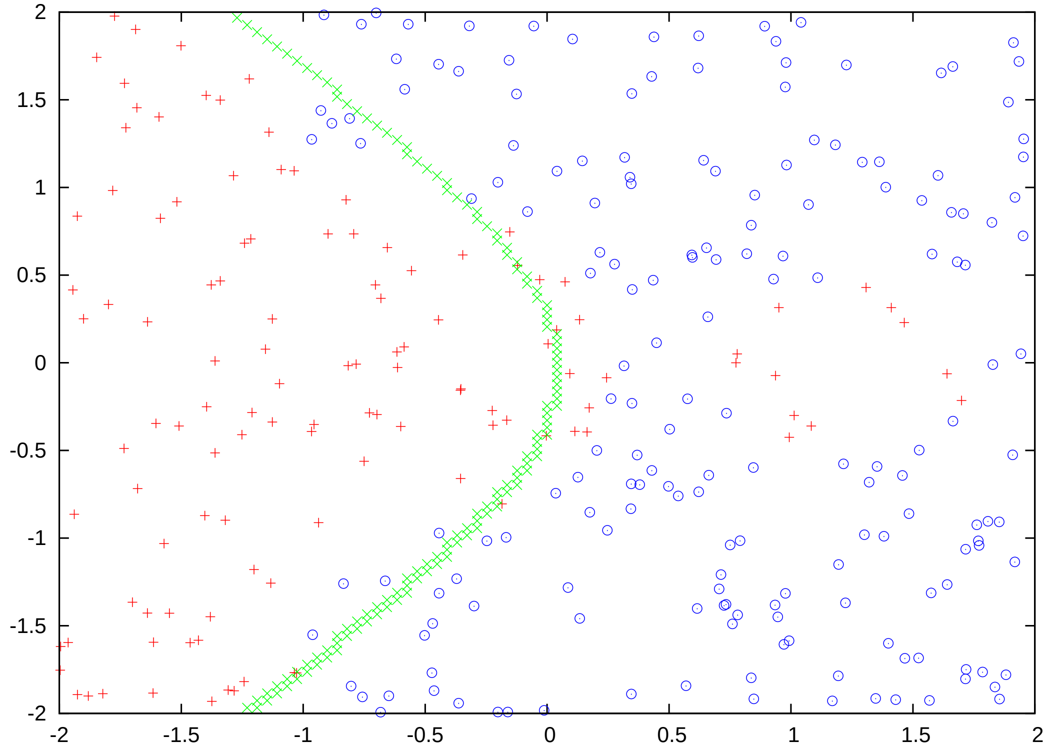
LS with Polynomail kernel: $p = 10$



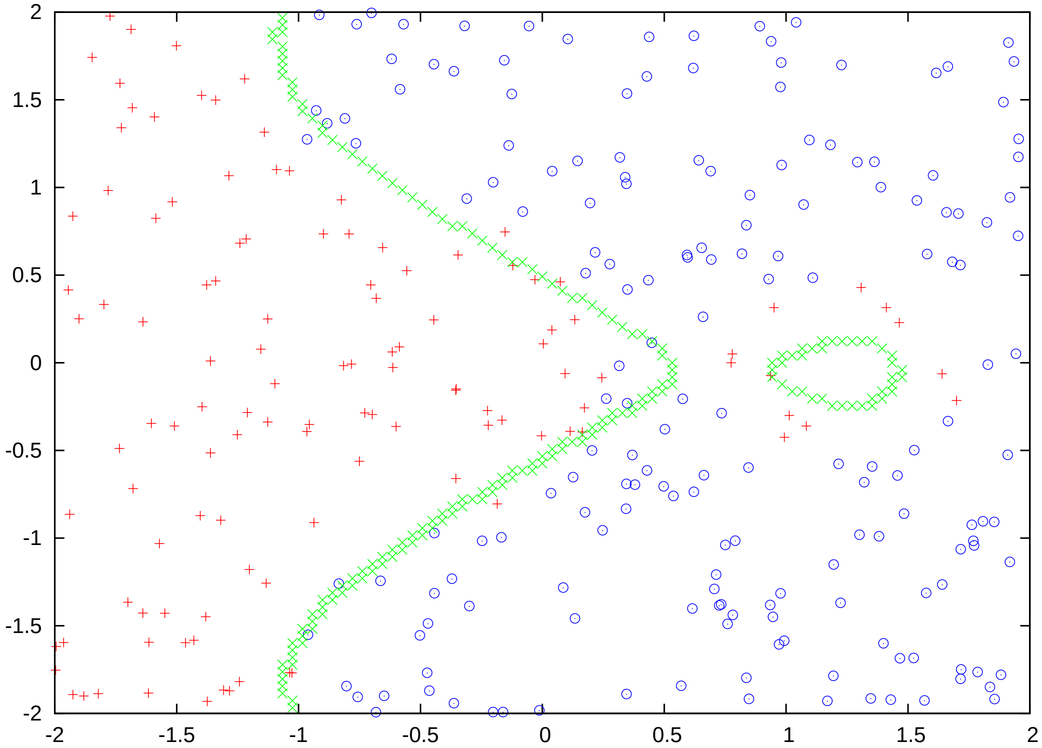
LS with Polynomail kernel: $p = 50$



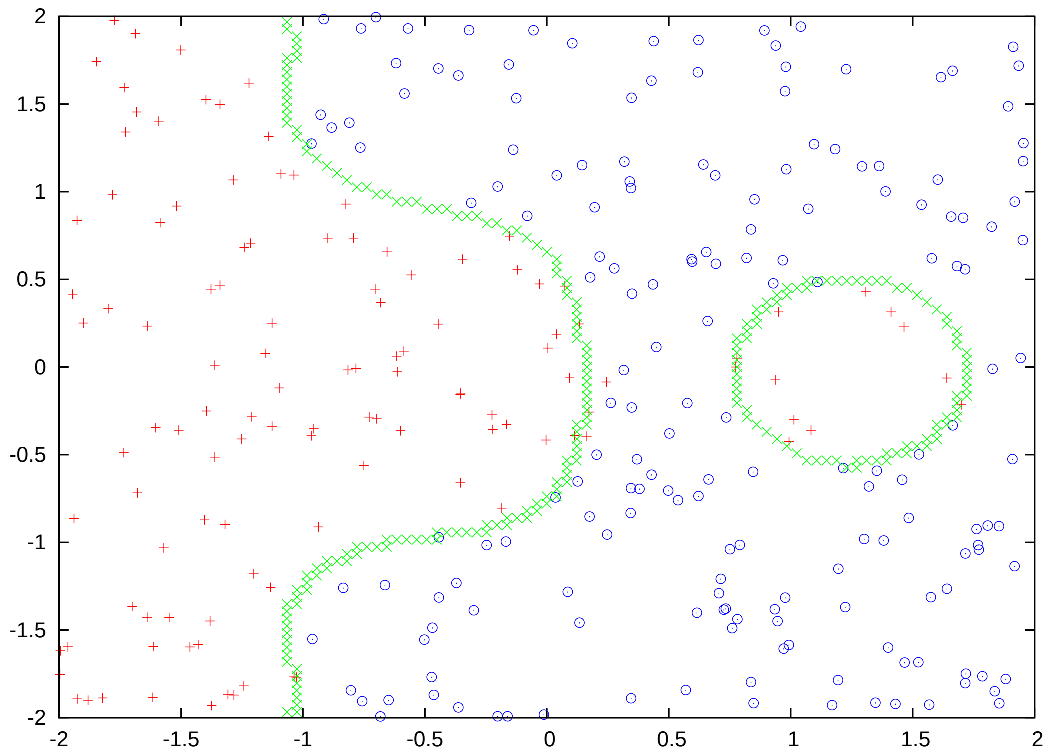
LS with RBF kernel $\exp(-\|x - x'\|_2^2/2\sigma^2)$: $\sigma = 10$



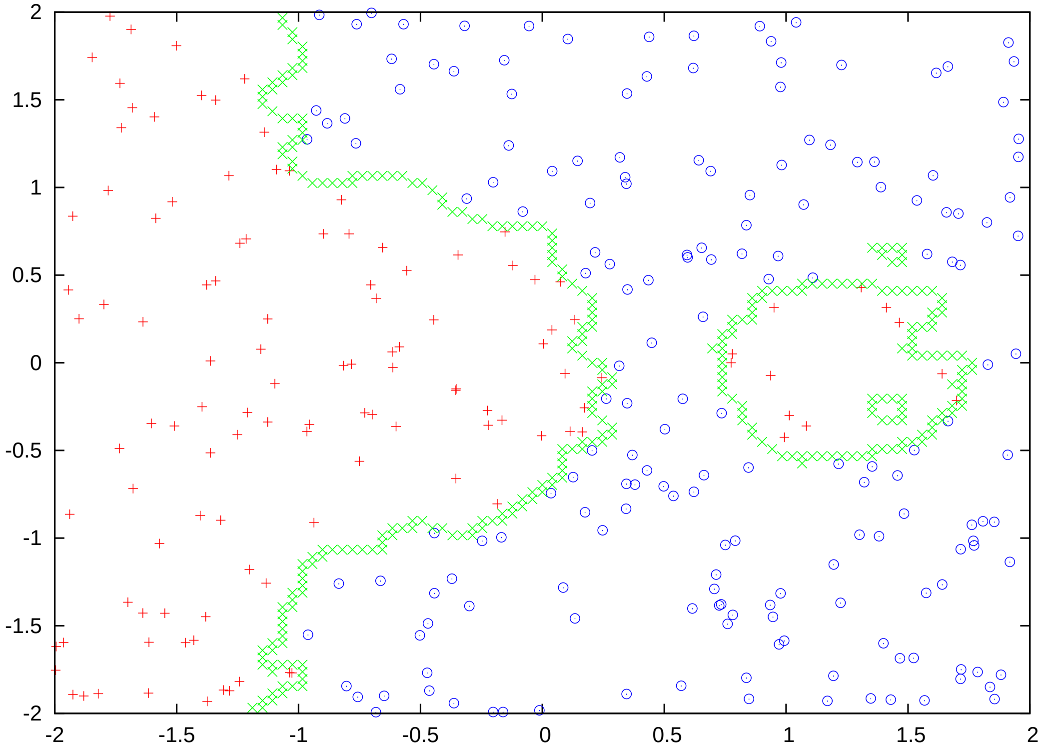
LS with RBF kernel $\exp(-\|x - x'\|_2^2/2\sigma^2)$: $\sigma = 3$



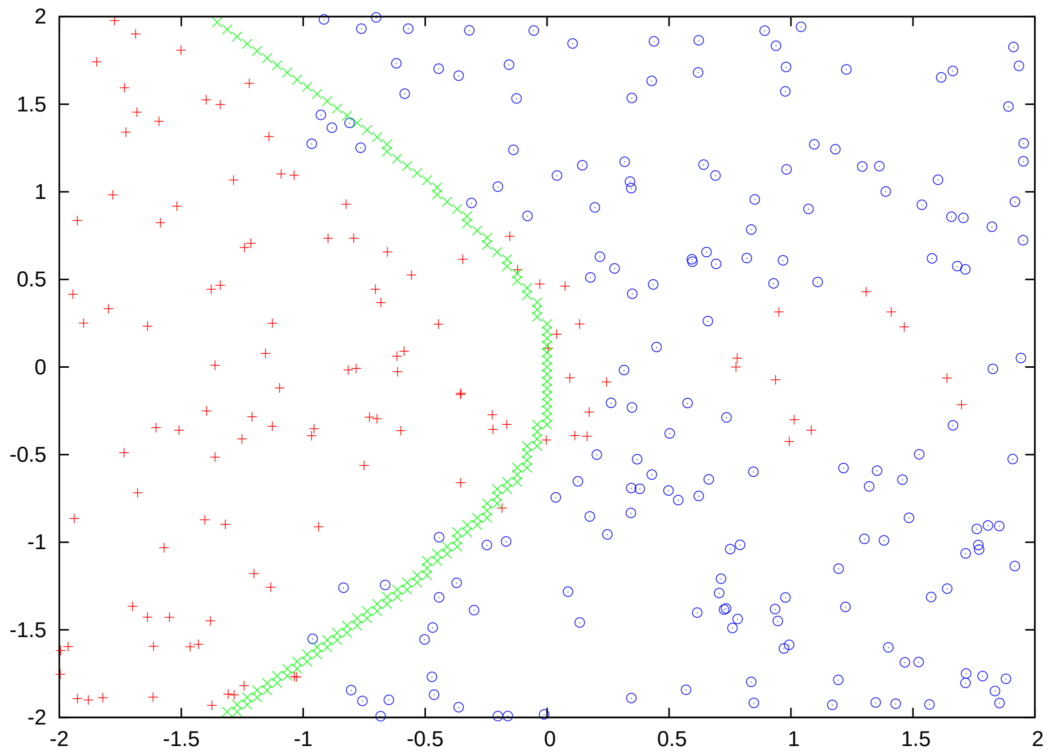
LS with RBF kernel $\exp(-\|x - x'\|_2^2/2\sigma^2)$: $\sigma = 1$



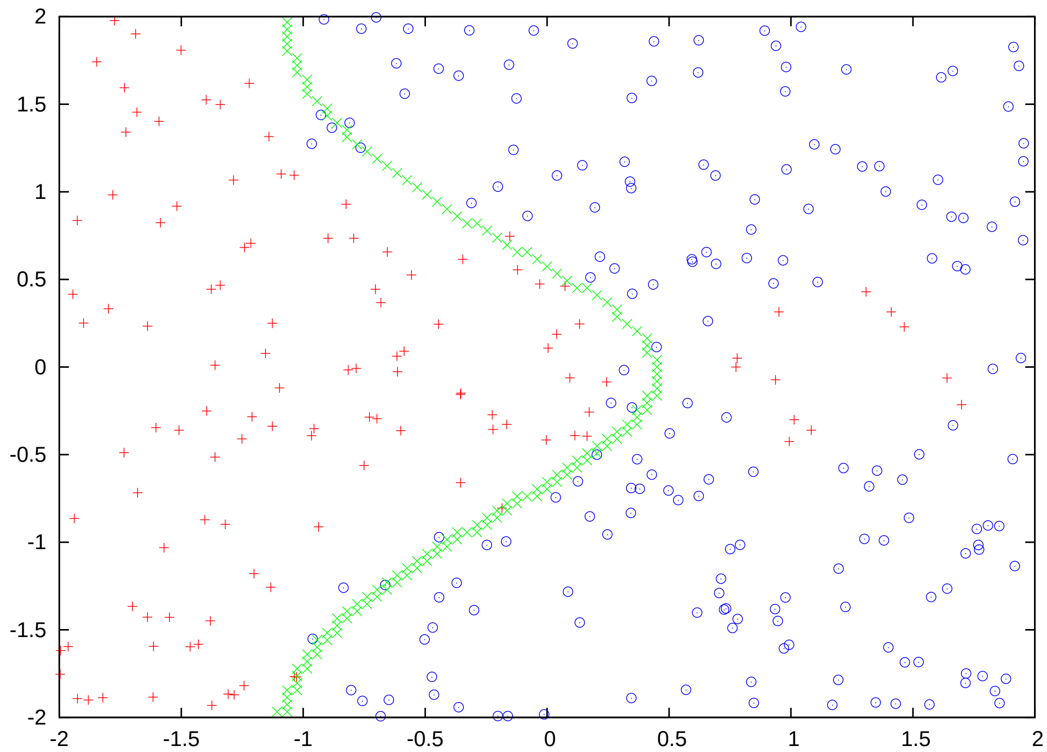
LS with RBF kernel $\exp(-\|x - x'\|_2^2 / 2\sigma^2)$: $\sigma = 0.1$



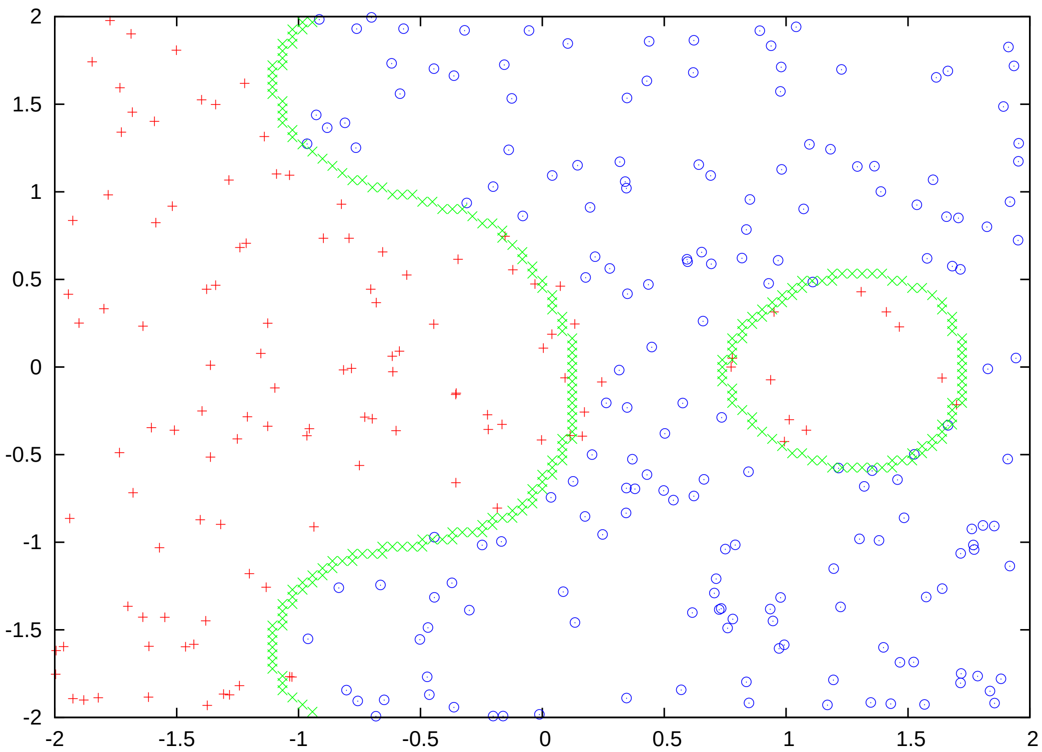
Ridge regression with RBF as feature: $\sigma = 10$



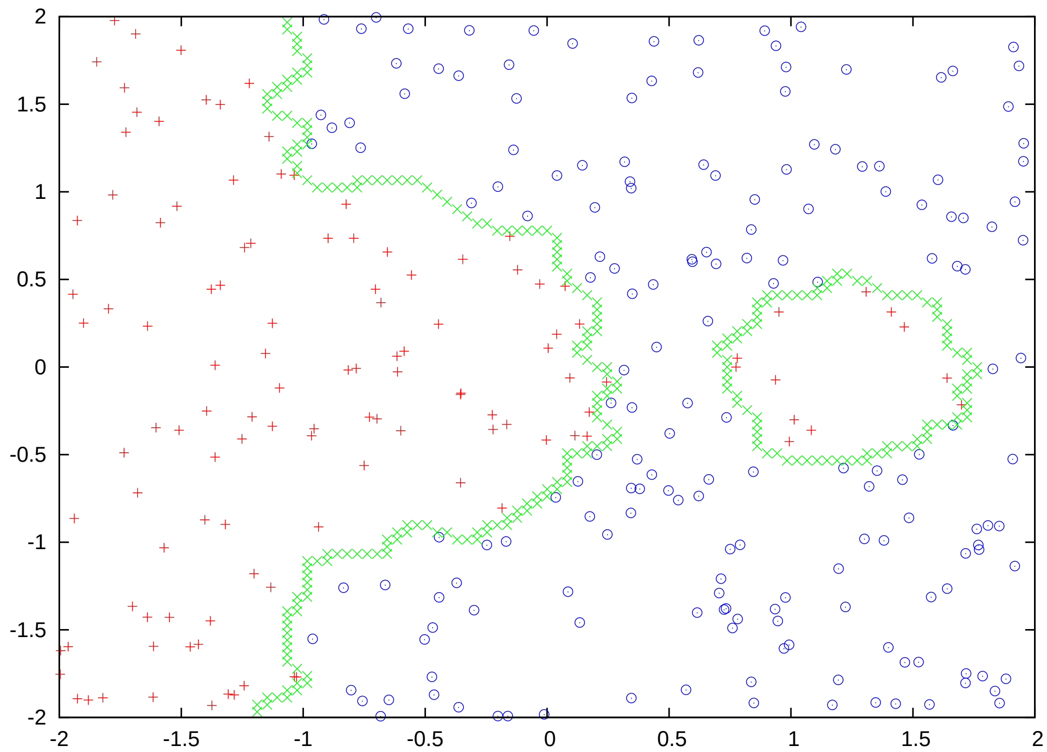
Ridge regression with RBF as feature: $\sigma = 3$



Ridge regression with RBF as feature: $\sigma = 1$



Ridge regression with RBF as feature: $\sigma = 0.1$



Mercer's Theorem

- Let $k(x, x')$ be a symmetric function. It's a (positive-definite) kernel if and only if $\forall x_i$ ($i = 1, \dots, n$), the gram matrix $K = [k(x_i, x_j)]_{i,j=1}^n$ is positive semi-definite.

- (Mercer's Theorem) Assume that $k(x, x')$ is a continuous symmetric function on $R^d \times R^d$ such that

$$\int k(x, x')f(x)f(x')dxdx' \geq 0$$

  for all $f \in L_2$. Then we can expand $k(x, x')$ in a uniformly convergence series in terms of eigen-functions v_j of operator $f \to \int k(x, x')f(x')dx'$: $k(x, x') = \sum_j \lambda_j v_j(x) v_j(x') = \psi(x)^T \psi(x')$.

Reproducing kernel Hilbert space (RKHS)

A Hilbert space H of functions spanned by functions of the form:

$$f(x) = \sum_i \alpha_i k(x_i, x),$$

with norm:

$$\|f\|_{\mathcal{H}}^2 = \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j),$$

where k is a kernel called reproducing kernel.

RKHS properties

- Some properties:
  - there is a feature space representation:
    - $f(x) = w^T \psi(x)$, and $\|f\|_{\mathcal{H}} = \|w\|_2$.
  - each x' maps to a vector $\psi(x')$ in the feature space, and corresponds to function $f_{x'}(x) = k(x', x) = \psi(x')^T \psi(x)$.
  - $f(x) \leq \|w\|_2 \|\psi(x)\|_2 = \|f\|_{\mathcal{H}} \sqrt{k(x, x)}$
    - $\forall x$, the linear functional $f \to f(x)$ is bounded.

- Given a Hilbert space $\mathcal{H}$ of real-valued functions $f(x)$ with norm $\|\cdot\|_{\mathcal{H}}$, such that $f \to f(x)$ is bounded for all x, then it is a RKHS.
  - Rietz representatin theorem implies $x' \to f_{x'}(x) = k(x', x) \in \mathcal{H}$
  - such that $f(x') = \langle f_{x'}, f \rangle_{\mathcal{H}}$, thus $f_{x'}(x) = \langle f_{x'}, f_x \rangle_{\mathcal{H}} = k(x', x)$.

Learning in Hilbert space

- Primal formulation on RKHS:

$$[\hat{f}, \hat{b}] = \arg \min_{f \in \mathcal{H}, b \in R} \left[\sum_{i=1}^{n} \phi \left(\sum_{j=1}^{n} f(X_i) + b, Y_i \right) + \lambda \|f\|_{\mathcal{H}}^2 \right] \qquad (III)$$

with scoring function $f(x) = \hat{f}(x) + \hat{b}$.

- If a reproducing kernel of an RKHS is $k(x, x') = \psi(x)^T \psi(x')$, then (I) and (II) are special representations of (III), thus equivalent to (III).

Different representations of RKHS norm

- $f(x) \in \mathcal{H}$ with RHKS $\mathcal{H}$ norm $\|f\|_{\mathcal{H}}$
  - feature: $\psi(x)$
  - kernel: $k(x, x') = \psi(x)^T \psi(x')$
- kernel representation: $f(x) = \sum_{i=1}^n \alpha_i k(X_i, x)$

$$\|f\|_{\mathcal{H}}^2 = \alpha^T K_m \alpha \leq a^2.$$

- feature space representation: $f(x) = w^T \psi(x)$

$$\|f\|_{\mathcal{H}}^2 = \|w\|_2^2$$

Some examples

- Linear kernel: $k(x,x') = x^T x' = \sum_j x_j x'_j$
  - features: $\psi_j(x) = x_j$.
  - RKHS functions: $f(x) = w^T x$.
  - norm: $\|f\|_{\mathcal{H}}^2 = \|w\|_2^2$.

- Polynomial kernel: $k(x,x') = (1 + x^T x')^p = \sum_s C_p^s \prod_j x_j^{s_j} {x'_j}^{s_j}$
  - features: $\prod_{j=1}^d x_j^{s_j}$: $s_0 + \sum_{j=1}^d s_j = p$ and $s_j \geq 0$.
  - RKHS functions: $f(x) = \sum_s w_s \prod_j x_j^{s_j}$
  - norm: $\|f\|_{\mathcal{H}}^2 = \sum_s w_s^2 / C_p^s$.

- Inner product exponential kernel: $k(x,x') = \exp(x^T x') = \sum_s \prod_{j=1}^d \frac{1}{s_j!} x_j^{s_j} {x'_j}^{s_j}$.

- features: $\prod_{j=1}^{d} x_j^{s_j}$ with $s_j \geq 0$.
- RKHS functions: $f(x) = \sum_s w_s \prod_j x_j^{s_j}$
- norm: $\|f\|_{\mathcal{H}}^2 = \sum_s w_s^2 \prod_{j=1}^{d} s_j!$

- RBF (radial basis function) exponential kernel:
$k(x,x') = \exp(-\|x-x'\|_2^2/2\sigma^2) = \sum_s \prod_{j=1}^{d} \frac{1}{s_j!\sigma^{2s_j}} x_j^{s_j} e^{-x^2/2\sigma^2} x_j'^{s_j} e^{-x'^2/2\sigma^2}$.
  - features: $\prod_{j=1}^{d} x_j^{s_j} e^{-x^2/2\sigma^2}$
  - RKHS functions: $f(x) = \sum_s w_s \prod_j x_j^{s_j} e^{-x^2/2\sigma^2}$
  - norm: $\|f\|_{\mathcal{H}}^2 = \sum_s w_s^2 \prod_{j=1}^{d} (s_j!\sigma^{2s_j})$

- Smoothing spline (1-d) with periodic boundary condition: $f(-\pi) = f(\pi)$.
  - RKHS functions $f(x) = \sum_{j>0} [a_j \cos(jx) + b_j \sin(jx)]$.
  - norm $\|f\|_{\mathcal{H}}^2 = \frac{1}{\pi} \int_{-\pi}^{\pi} (f^{(p)}(x))^2 dx = \sum_j j^{2p} (a_j^2 + b_j^2)$
  - features $\cos(jx)$ and $\sin(jx)$