Intro to ML

November 10th, 2021

Logistic Discrimination (logistic regression)

Two classes: Assume log likelihood ratio is linear

$$\log \frac{p(\mathbf{x}|C_1)}{p(\mathbf{x}|C_2)} = \mathbf{w}^T\mathbf{x} + w_0^o$$

$$\text{logit}(P(C_1|\mathbf{x})) = \log \frac{P(C_1|\mathbf{x})}{1 - P(C_1|\mathbf{x})} = \log \frac{p(\mathbf{x}|C_1)}{p(\mathbf{x}|C_2)} + \log \frac{P(C_1)}{P(C_2)}$$

$$= \mathbf{w}^T\mathbf{x} + w_0$$

$$\text{where } w_0 = w_0^o + \log \frac{P(C_1)}{P(C_2)}$$

$$y = \hat{P}(C_1|\mathbf{x}) = \frac{1}{1 + \exp\left[-\left(\mathbf{w}^T\mathbf{x} + w_0\right)\right]}$$

Training: Two Classes

Model label given x with probability y

$$\mathcal{X} = \{\mathbf{x}^t, r^t\}_t \qquad \boxed{r^t \mid \mathbf{x}^t \sim \text{Bernoulli}(y^t)}$$

Note the difference to likelihood method

$$y = P(C_1 \mid \mathbf{x}) = \frac{1}{1 + \exp\left[-\left(\mathbf{w}^T\mathbf{x} + w_0\right)\right]}$$

Maximize this function label/data condition likelihood based on data we have

$$l(\mathbf{w}, w_0 \mid \mathcal{X}) = \prod_t \left(y^t\right)^{\left(r^t\right)} \left(1 - y^t\right)^{\left(1 - r^t\right)}$$

$$E = -\log l$$

Minimize this

$$E(\mathbf{w}, w_0 \mid \mathcal{X}) = -\sum_t r^t \log y^t + \left(1 - r^t\right) \log\left(1 - y^t\right)$$

What is this? This is a function that we call ‘cross entropy’

Training: Gradient-Descent

$$E(\mathbf{w}, w_0 \mid \mathcal{X}) = -\sum_t r^t \log y^t + \left(1 - r^t\right) \log\left(1 - y^t\right)$$

$$\text{If } y = \text{sigmoid}(\text{a}) \quad \frac{dy}{da} = y(1-y)$$

$$\Delta w_j = -\eta \frac{\partial E}{\partial w_j} = \eta \sum_t \left(\frac{r^t}{y^t} - \frac{1-r^t}{1-y^t}\right) y^t \left(1-y^t\right) x_j^t$$

$$\boxed{= \eta \sum_t \left(r^t - y^t\right) x_j^t, j = 1, \ldots, d}$$

$$\boxed{\Delta w_0 = -\eta \frac{\partial E}{\partial w_0} = \eta \sum_t \left(r^t - y^t\right)}$$

Good practice: Z-normalize features

For $j = 0, \ldots, d$
  $w_j \leftarrow$ rand(-0.01,0.01)
Repeat
  For $j = 0, \ldots, d$
    $\Delta w_j \leftarrow 0$
  For $t = 1, \ldots, N$
    $o \leftarrow 0$
    For $j = 0, \ldots, d$
      $o \leftarrow o + w_j x_j^t$
    $y \leftarrow \text{sigmoid}(o)$
    $\Delta w_j \leftarrow \Delta w_j + (r^t - y) x_j^t$
  For $j = 0, \ldots, d$
    $w_j \leftarrow w_j + \eta \Delta w_j$
Until convergence

Keep initial close to zero

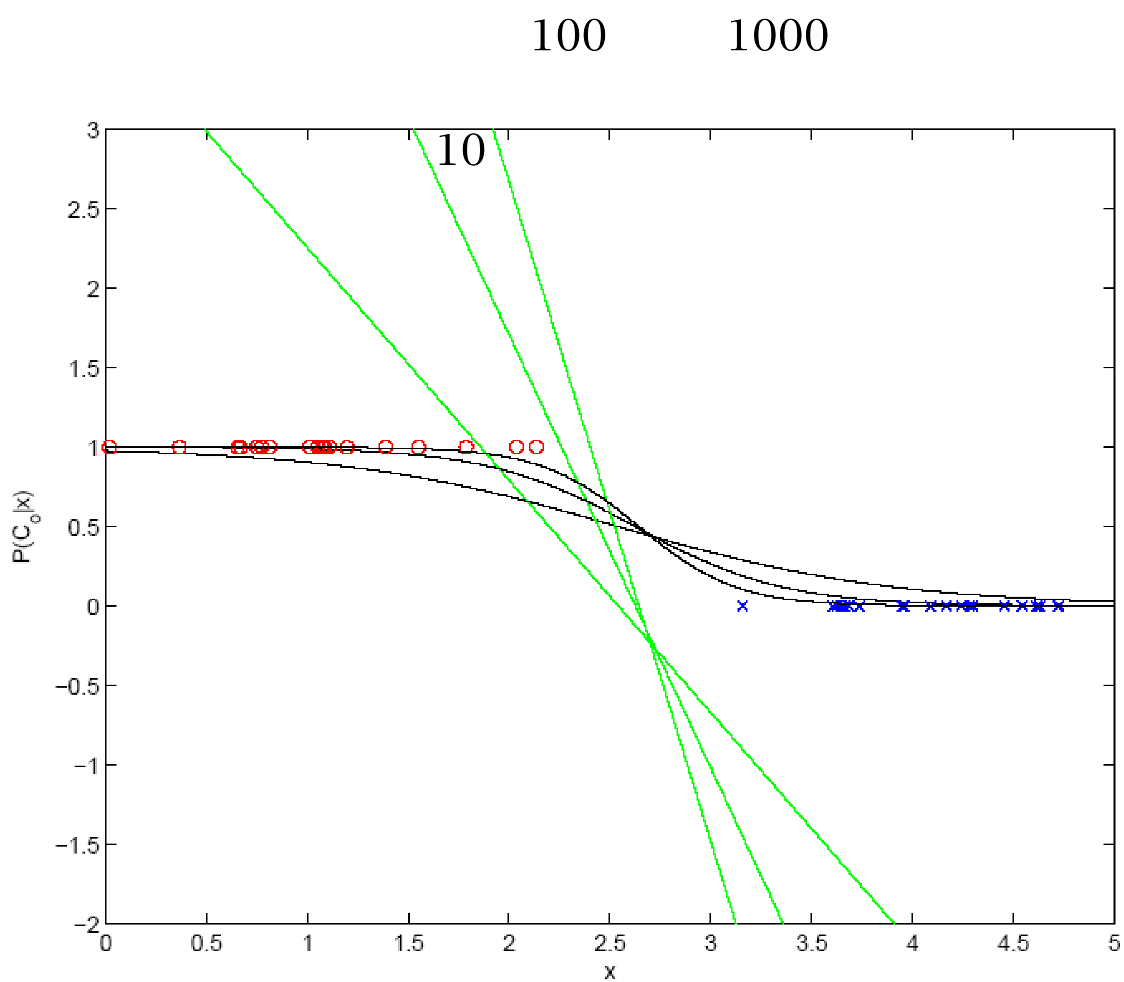
Notes

- Gradient does not change anymore then converged

- In this case, we assume log ratio of class density is linear to perform this learning (but we never explicitly estimate $p(x|Ci)$ or $P(Ci)$

- Training effectively takes data of a class to result in either $y<0.5$ or $y>0.5$

1d Example



Keep iteration without stopping, make the sigmoid function harden (quickly takes the sample to close to 0 or 1)

But does not change misclassification rate

Early stopping

CHAPTER 14:

Kernel Machines

Kernel Machines

- Discriminant-based: No need to estimate densities first
- Define the discriminant in terms of support vectors
  - Support vectors: subset of training instances
- The use of kernel functions, application-specific measures of similarity
- Convex optimization problems with a unique solution

Optimal Separating Hyperplane

$$\mathcal{X} = \{\mathbf{x}^t, r^t\}_t \text{ where } r^t = \begin{cases} +1 & \text{if } \mathbf{x}^t \in C_1 \\ -1 & \text{if } \mathbf{x}^t \in C_2 \end{cases}$$

find $\mathbf{w}$ and w_0 such that

$$\mathbf{w}^T \mathbf{x}^t + w_0 \geq +1 \text{ for } r^t = +1$$

$$\mathbf{w}^T \mathbf{x}^t + w_0 \leq +1 \text{ for } r^t = -1$$

which can be rewritten as

$$r^t \left(\mathbf{w}^T \mathbf{x}^t + w_0 \right) \geq \boxed{+1}$$

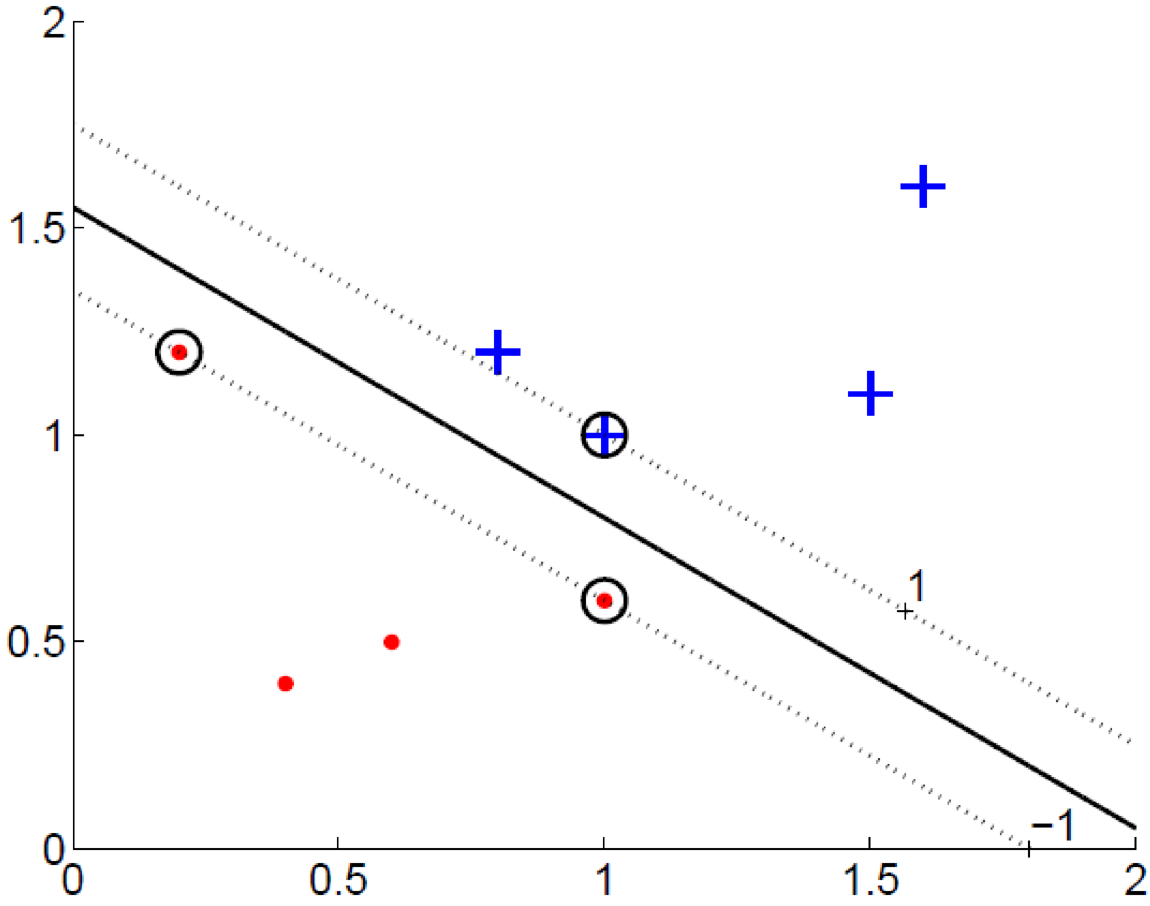
Not simply >0
With some distance

(Cortes and Vapnik, 1995; Vapnik, 1995)

Margin

- Distance from the discriminant to the closest instances on either side
- Distance of x to the hyperplane is
$$\frac{\left|\mathbf{w}^T\mathbf{x}^t + w_0\right|}{\|\mathbf{w}\|}$$
- We require
$$\frac{r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right)}{\|\mathbf{w}\|} \geq \rho, \forall t$$
Like to maximize this distance
- For a unique sol'n, fix $\rho||w||=1$, and to max margin equal minimize w
$$\min \frac{1}{2}\|\mathbf{w}\|^2 \text{ subject to } r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) \geq +1, \forall t$$

Margin



$$\min \frac{1}{2}\|\mathbf{w}\|^2 \text{ subject to } r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) \geq +1, \forall t$$

Use LaGrange multiplier

$$L_p = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{t=1}^{N} \alpha^t \left[r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) - 1\right]$$

$$= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{t=1}^{N} \alpha^t r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) + \sum_{t=1}^{N} \alpha^t$$

$$\frac{\partial L_p}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{t=1}^{N} \alpha^t r^t \mathbf{x}^t$$

$$\frac{\partial L_p}{\partial w_0} = 0 \Rightarrow \sum_{t=1}^{N} \alpha^t r^t = 0$$

The reason to rewrite this

- Original complexity depends on d
- Turn the solution into complexity depends on N

$$L_d = \frac{1}{2}\left(\mathbf{w}^T\mathbf{w}\right) - \mathbf{w}^T\sum_t \alpha^t r^t \mathbf{x}^t - w_0 \sum_t \alpha^t r^t + \sum_t \alpha^t$$

$$= -\frac{1}{2}\left(\mathbf{w}^T\mathbf{w}\right) + \sum_t \alpha^t$$

$$= -\frac{1}{2}\sum_t\sum_s \alpha^t \alpha^s r^t r^s \left(\mathbf{x}^t\right)^T \mathbf{x}^s + \sum_t \alpha^t$$

Need to solve for alpha

subject to $\sum_t \alpha^t r^t = 0$ and $\alpha^t \geq 0, \forall t$

Most α^t are 0 and only a small number have $\alpha^t > 0$;

they are the support vectors they satisfy $r^t(w^T x^t + w_0) = 1$

These are support vector machines, it only cares those on the boundaries not within the decision regions

testing

- $g(x)=w^Tx + w_0$
- Choose the results according to sign

Soft Margin Hyperplane

- Not linearly separable

$$r^t\left(\mathbf{w}^T x^t + w_0\right) \ge 1 - \xi^t$$

Slack variable

$$0 < \xi^t < 1$$

Error in the margin

$$\xi^t \ge 1$$

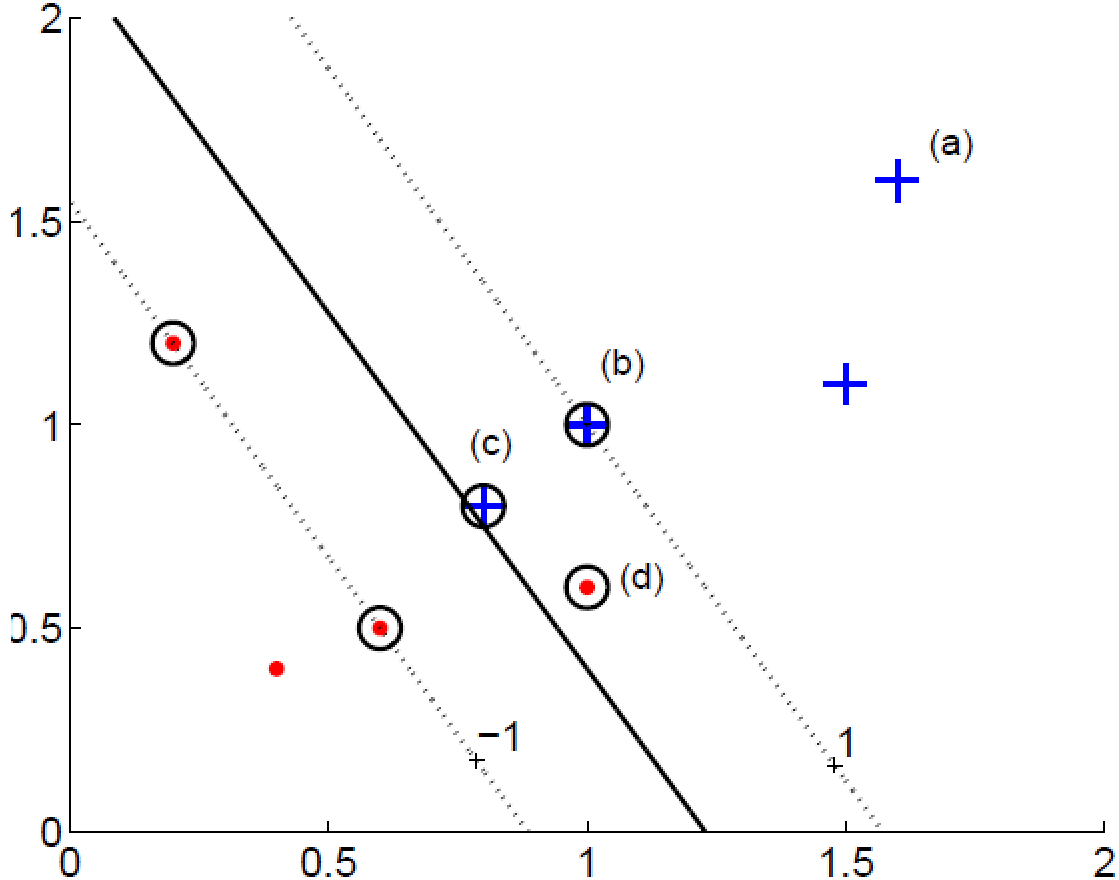
misclassified

- Soft error $\sum_t \xi^t$

- New primal is

Lagrange to ensure positivity of error

$$L_p = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_t \xi^t - \sum_t \alpha^t\left[r^t\left(\mathbf{w}^T x^t + w_0\right) - 1 + \xi^t\right] - \sum_t \mu^t \xi^t$$

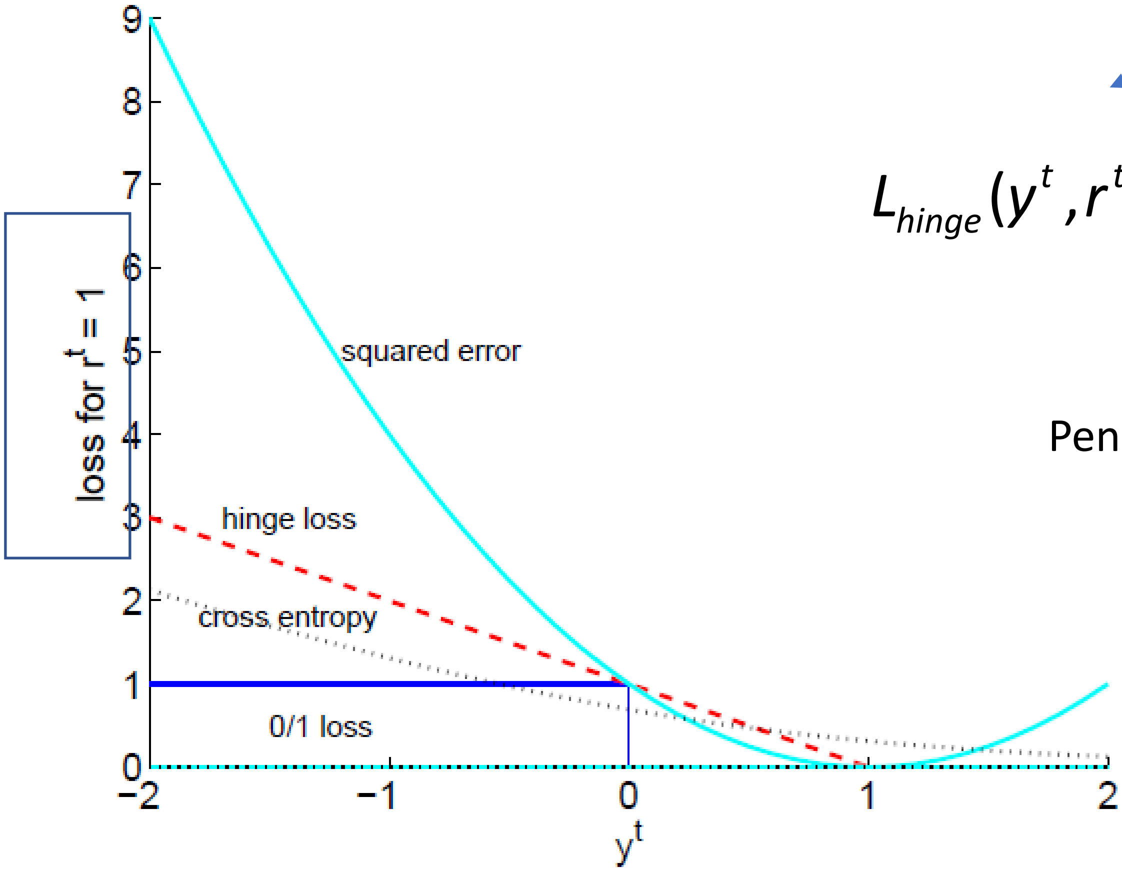


C is a tunable parameter
Trade off between margin maximization and error minimization
Too large -> high penalty for error -> may overfit
Too small -> not enough penalty for error -> may underfit

Hinge Loss

$$r^t\left(\mathbf{w}^T x^t + w_0\right) \geq \boxed{1-\xi^t}$$

$$L_{hinge}(y^t, r^t) = \begin{cases} 0 & \text{if } y^t r^t \geq 1 \\ 1 - y^t r^t & \text{otherwise} \end{cases}$$



Penalizes instance in the margin

Kernel Trick

- Preprocess input ***x*** by basis functions

$$\boldsymbol{z} = \boldsymbol{\varphi}(\boldsymbol{x}) \qquad g(\boldsymbol{z}) = \boldsymbol{w}^T\boldsymbol{z}$$

$$g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x})$$

- The SVM solution

$$\mathbf{w} = \sum_t \alpha^t r^t \mathbf{z}^t = \sum_t \alpha^t r^t \boldsymbol{\varphi}(\mathbf{x}^t)$$

$$g(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}) = \sum_t \alpha^t r^t \boxed{\boldsymbol{\varphi}(\mathbf{x}^t)^T \boldsymbol{\varphi}(\mathbf{x})}$$

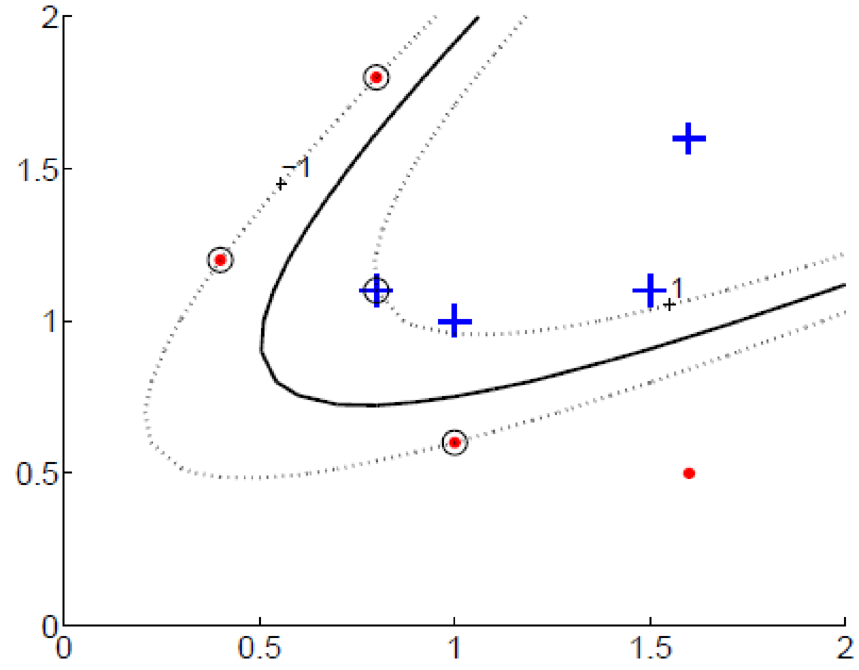
$$g(\mathbf{x}) = \sum_t \alpha^t r^t \boxed{K(\mathbf{x}^t, \mathbf{x})}$$

Dot product in z space replace by a kernel machine

Polynomial Kernels

Polynomials of degree q:

$$K\left(\mathbf{x}^t, \mathbf{x}\right) = \left(\mathbf{x}^T \mathbf{x}^t + 1\right)^q$$



$$K(\mathbf{x}, \mathbf{y}) = \left(\mathbf{x}^T \mathbf{y} + 1\right)^2$$
$$= \left(x_1 y_1 + x_2 y_2 + 1\right)^2$$
$$\boxed{= 1 + 2x_1 y_1 + 2x_2 y_2 + 2x_1 x_2 y_1 y_2 + x_1^2 y_1^2 + x_2^2 y_2^2}$$

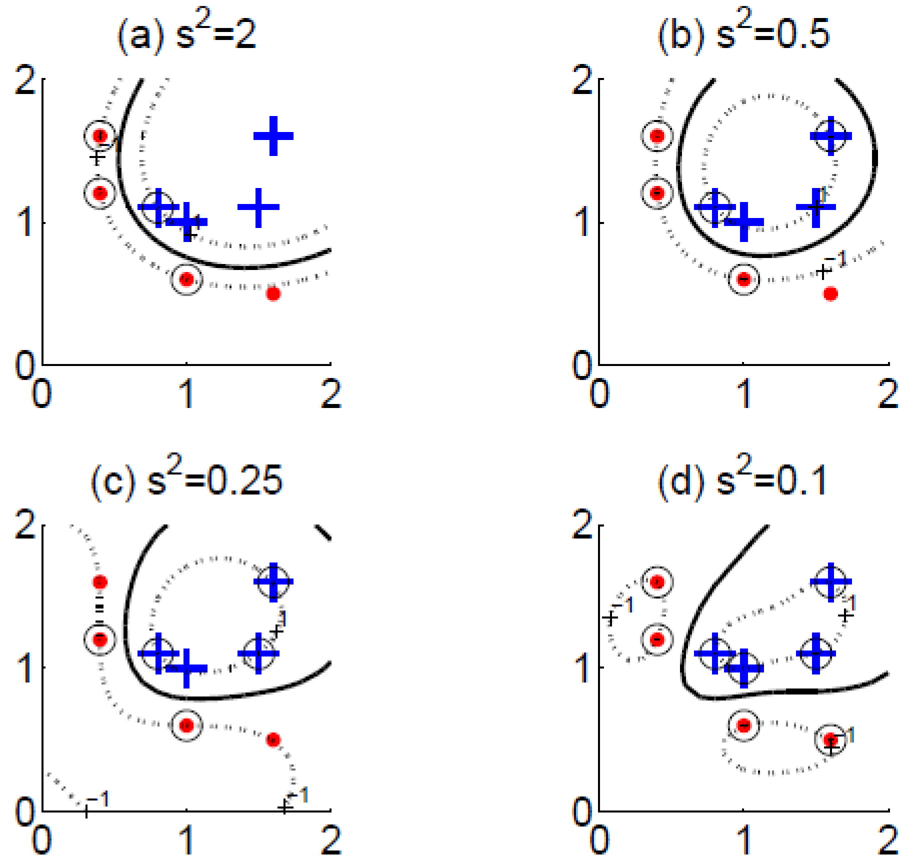
$$\phi(\mathbf{x}) = \left[1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1 x_2, x_1^2, x_2^2\right]^T$$

inner product of this basis function

RBF (Gaussian) Kernel

Radial-basis functions:

$$K\left(\mathbf{x}^t, \mathbf{x}\right) = \exp\left[-\frac{\left\|\mathbf{x}^t - \mathbf{x}\right\|^2}{2s^2}\right]$$



Defining Kernels

- Kernel “engineering”
- Defining good measures of similarity
- String kernels, graph kernels, image kernels, ...
- Kernel can be ‘designed’

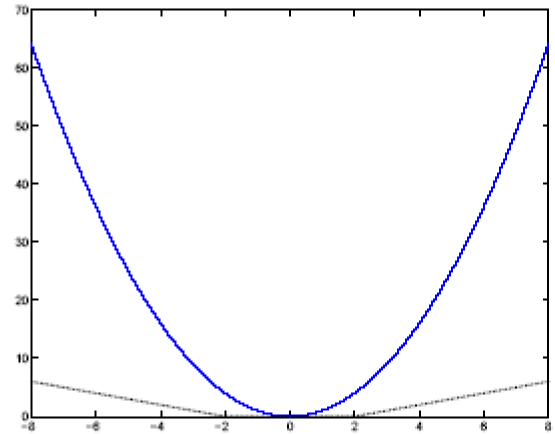
SVM for Regression

- Use a linear model (possibly kernelized)

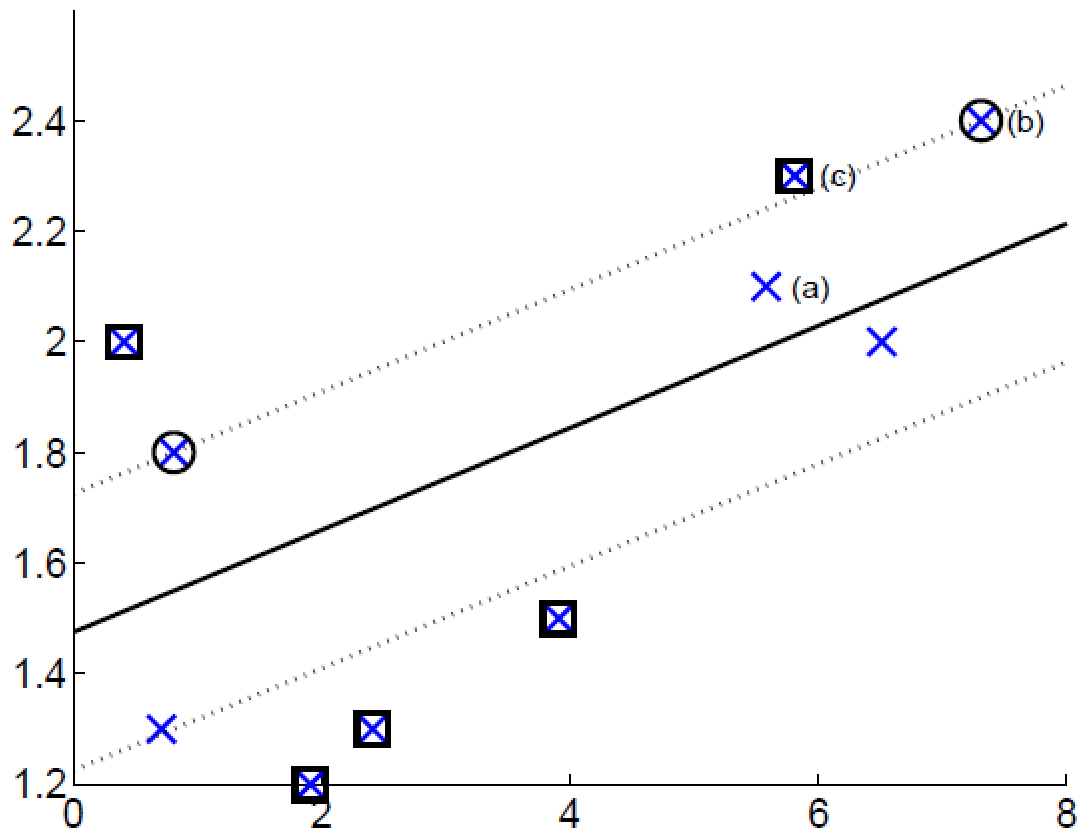
$$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + w_0$$

- Use the ϵ-sensitive error function

$$e_{\varepsilon}\left(r^t, f\left(\mathbf{x}^t\right)\right) = \begin{cases} 0 & \text{if } \left|r^t - f\left(\mathbf{x}^t\right)\right| < \varepsilon \\ \left|r^t - f\left(\mathbf{x}^t\right)\right| - \varepsilon & \text{otherwise} \end{cases}$$



$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_t \left(\xi_+^t + \xi_-^t\right) \quad \begin{aligned} r^t - \left(\mathbf{w}^T\mathbf{x} + w_0\right) &\leq \varepsilon + \xi_+^t \\ \left(\mathbf{w}^T\mathbf{x} + w_0\right) - r^t &\leq \varepsilon + \xi_-^t \\ \xi_+^t, \xi_-^t &\geq 0 \end{aligned}$$


2.4
2.2
2
1.8
1.6
1.4
1.2
0
2
4
6
8
(a)
(b)
(c)

Kernel Regression

- Polynomial kernel
- Gaussian kernel