Maximum Magin Classification

Farzaneh Mirzazadeh
Thanks to Prof Dale Schuurmans

University of California, Santa Cruz

Winter ’17

Part One

Maximum Margin Classification for Data with Separable Classes

Linearly separable classes

Assume data is linearly separable.
Consider entire set of consistent linear classifiers.

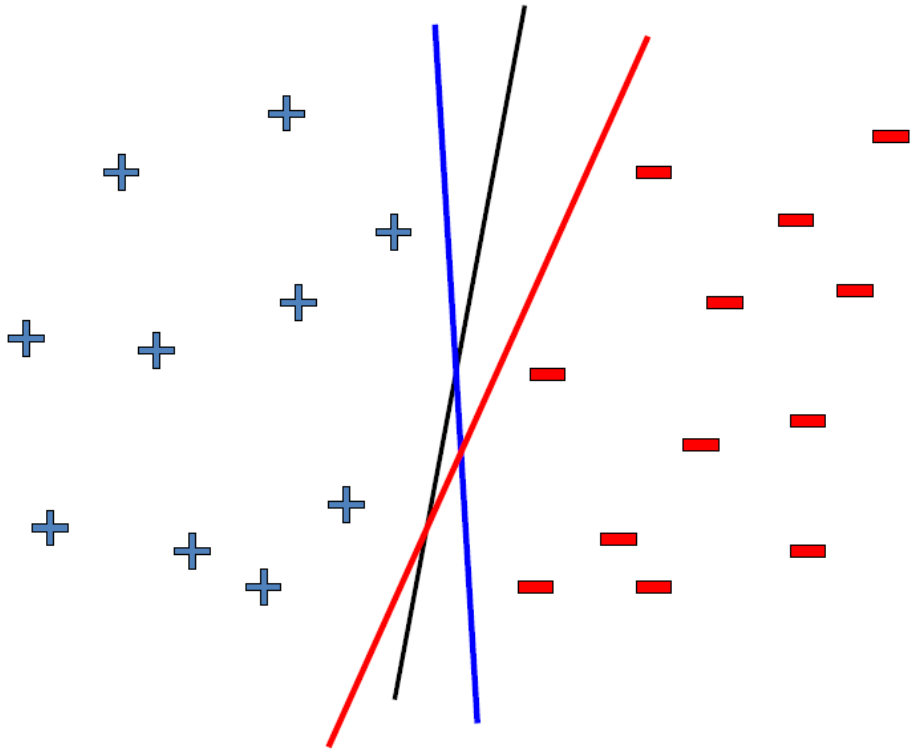


Figure from Prof Aarti Singh slides, CMU

Are some consistent linear classifiers better than others?
Which one we pick?

Linearly separable classes

- A simple but effective idea: Maximum Margi linear classifier

  Pick the one with the largest margin!

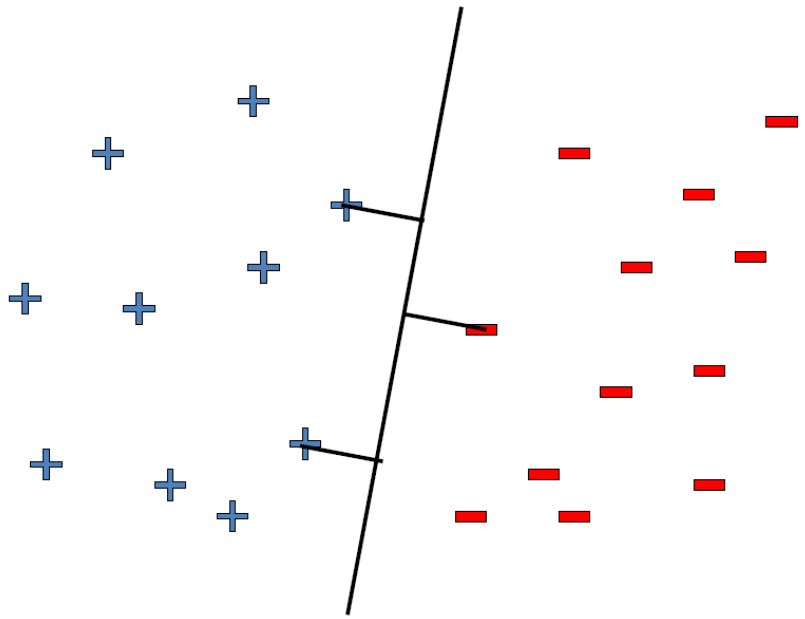


Figure from Prof Aarti Singh slides, CMU

- Valadimir Vapnik (1970s in Soviet Union), brought the idea to US in 1990.
- Choose $\boldsymbol{w}, b$ to maximize the minimum distance between data points and decision boundary.

How can this be computed efficiently?

Six steps to show how:

1. Parameterizing (writing the equation for) the decision boundary
2. Writing mathematical meaning of correct classification for a point
3. Measuring confidence of classification (aka functional margin)
4. Measuring the geometric margin
5. Maximizing the margin
6. Re-express it as a quadratic program (QP)

Step 1: Parameterizing the decision boundary

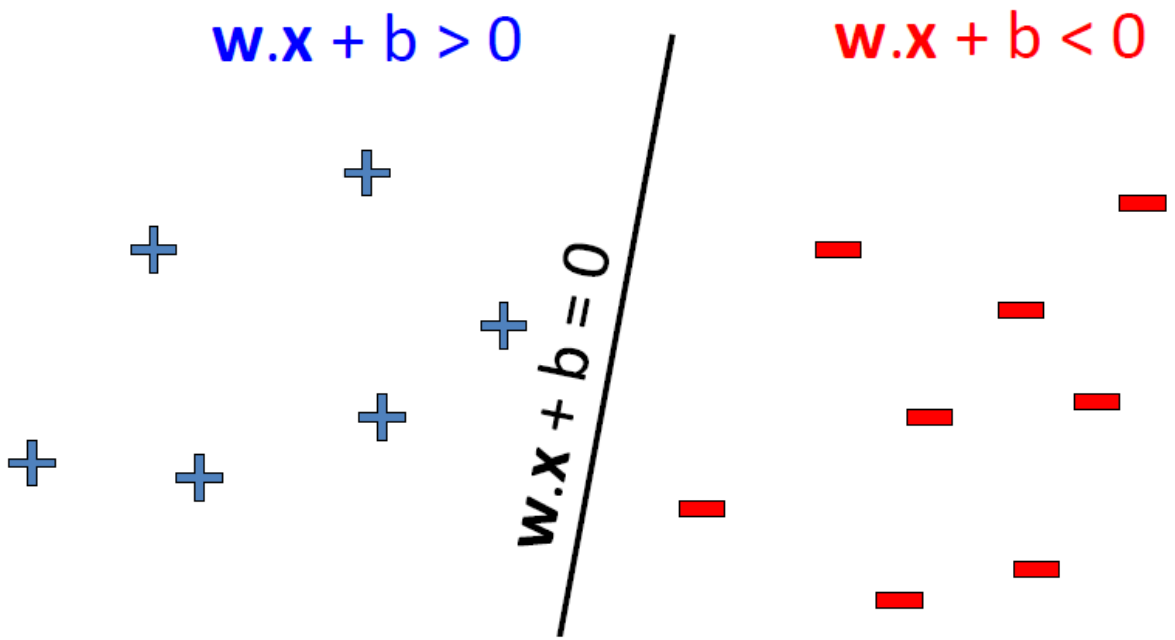
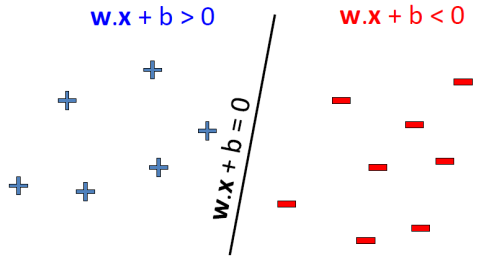


Figure from Prof Aarti Singh slides, CMU

- Equation of hyperplane: $\{\boldsymbol{x} : \boldsymbol{x}^\top \boldsymbol{w} + b = 0\}$

  Note: Dot product= inner product = $\boldsymbol{x} \bullet \boldsymbol{w} = \boldsymbol{w} \bullet \boldsymbol{x} := \boldsymbol{x}^\top \boldsymbol{w} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{w}_i$
- A hyperplane separates the space to two half spaces.
  - Positive side $\boldsymbol{x}^\top \boldsymbol{w} + b > 0$
  - Negative side $\boldsymbol{x}^\top \boldsymbol{w} + b < 0$

Step 2. Correct classification



For any point $\langle \boldsymbol{x}_i, y_i \rangle$, $y_i \in \{-1, +1\}$
the hyperplane $\boldsymbol{x}^\top \boldsymbol{w} + b$ classifies it correctly means:

$$\text{if} \quad y_i > 0 \quad \text{then} \quad \boldsymbol{x}_i^\top \boldsymbol{w} + b > 0$$
$$\text{if} \quad y_i < 0 \quad \text{then} \quad \boldsymbol{x}_i^\top \boldsymbol{w} + b < 0$$

In summary

Correct classification for all training points $i = 1, 2, \ldots, t$ requires

$$y_i(\boldsymbol{x}_i \boldsymbol{w}_i + b) > 0, \quad \forall i.$$

Step 3. Measuring confidence

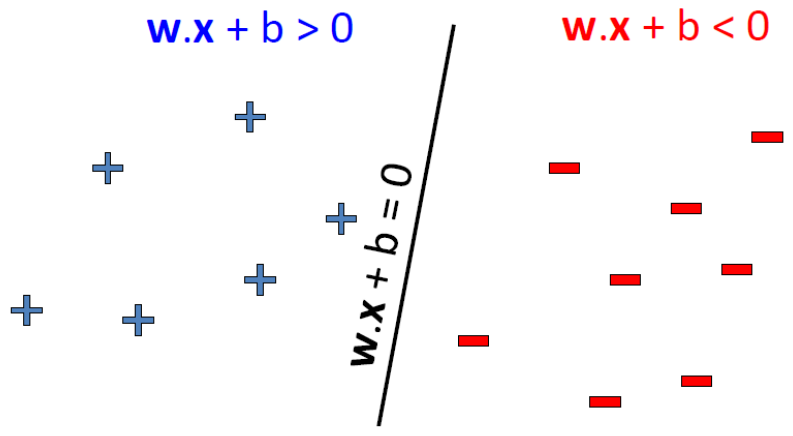


Figure from Prof Aarti Singh slides, CMU

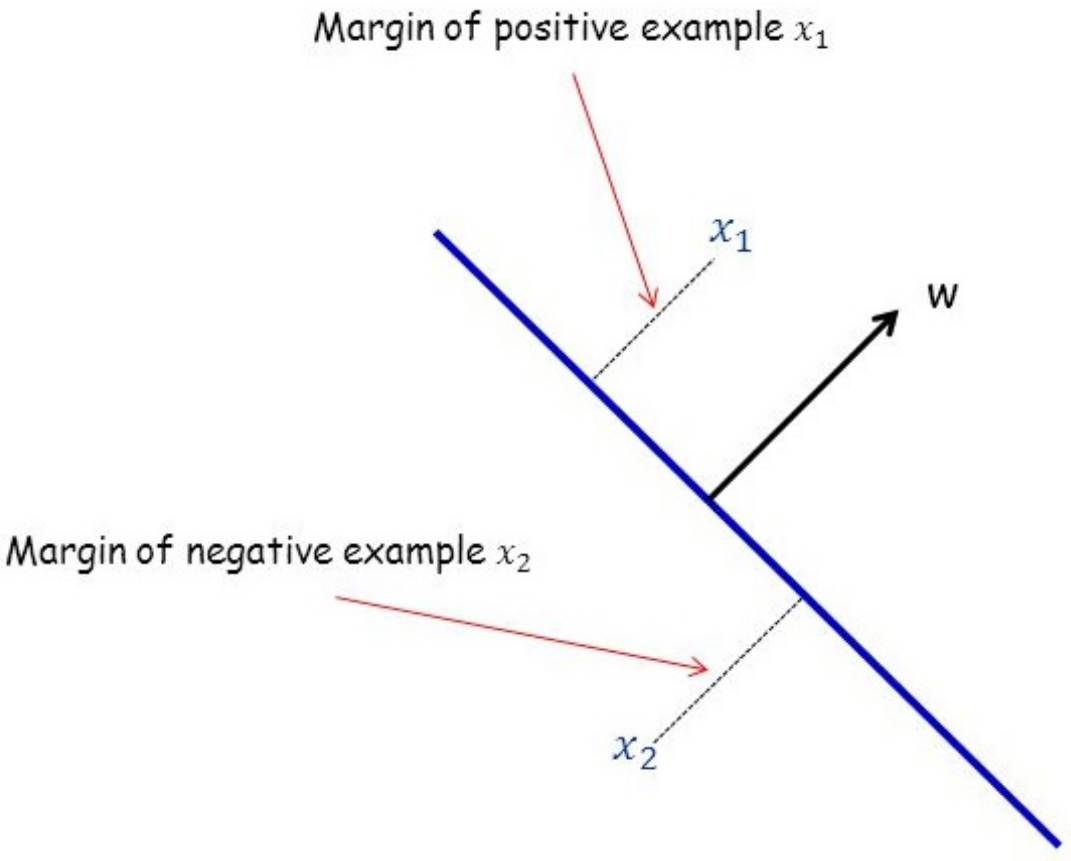
- Question: For a fixed hyperplane $\boldsymbol{x}^\top \boldsymbol{w} + b = 0$ what does the larger value of $(\boldsymbol{x}_i^\top \boldsymbol{w} + b) y_i$ show?
- Answer: Confidence of the classifier.
- 'Confidence' = $(\boldsymbol{x}_i^\top \boldsymbol{w} + b) y_i$.
- Call this confidence, functional margin.

Step 4. Computing the margin (geometric)

Confidence: functional margin $= y(\boldsymbol{x}^\top \boldsymbol{w} + b)$ depends on scaling

geometric margin $=$ distance between point and hyperplane

$$= \frac{y(\boldsymbol{x}^\top \boldsymbol{w} + b)}{||\boldsymbol{w}||}$$



Step 5. Maximizing the margin

Problem

Given a set of labeled examples, $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_t, y_t)$ where each $\boldsymbol{x}_i \in \mathbb{R}^n$ and each $y_i \in \{+1, -1\}$, find a weight vector $\boldsymbol{w}$ and intercept b such that $\text{sign}(\boldsymbol{x}_i^\top \boldsymbol{w} + b) = y_i$ for all i.
Assume linearly separable! Want to maximize the minimum *margin* (over training point).

Maximum margin training

$$\max_{\boldsymbol{w}, b} \quad \min_{i=1}^{t} \frac{y_i(\boldsymbol{x}^\top \boldsymbol{w}_i + b)}{\|\boldsymbol{w}\|}$$

How?
Recall the trick you have seen many times so far to convert min to a linear form. Use that. Then form a quadratic program (quadratic objective, linear constraints) with a little tweak. See next slides.

Step 6. Forming the quadratic program

$$\max_{\boldsymbol{w},b} \min_{i=1}^{t} \text{ (geometric) margin of training point i}$$

$$\equiv \max_{\boldsymbol{w},b} \min_{i=1}^{t} \frac{y_i(\boldsymbol{x}^\top \boldsymbol{w}_i + b)}{\|\boldsymbol{w}\|}$$

$$\equiv \max_{\delta,\boldsymbol{w},b} \delta \quad \text{subject to} \quad \frac{y_i(\boldsymbol{x}^\top \boldsymbol{w}_i + b)}{\|\boldsymbol{w}\|} \geqslant \delta, \quad i = 1, 2, \ldots, t$$

$$\equiv \max_{\delta,\boldsymbol{w},b} \delta \quad \text{subject to} \quad y_i(\boldsymbol{x}^\top \boldsymbol{w}_i + b) \geqslant \delta\|\boldsymbol{w}\|, \quad i = 1, 2, \ldots, t$$

$$\equiv \ldots$$

If data linearly separable, then there exists $\boldsymbol{w}, b$ s.t. $y_i(X_{i:}^\top \boldsymbol{w} - b) > 0, \forall i$. Thus any δ can be achieved by rescaling $\boldsymbol{w}$, b. So just rescale RHS to 1.
See next slide.

Step 6. Cont'd

$$\ldots$$

$$\equiv \max_{\boldsymbol{w},b} \frac{1}{\|\boldsymbol{w}\|} \quad \text{subject to} \quad y_i(\boldsymbol{x}^\top \boldsymbol{w}_i + b) \geqslant 1, \quad i = 1, 2, \ldots, t$$

$$\equiv \min_{\boldsymbol{w},b} \|\boldsymbol{w}\| \quad \text{subject to} \quad y_i(\boldsymbol{x}^\top \boldsymbol{w}_i + b) \geqslant 1, \quad i = 1, 2, \ldots, t$$

$$\equiv \min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 \quad \text{subject to} \quad y_i(\boldsymbol{x}^\top \boldsymbol{w}_i + b) \geqslant 1, \quad i = 1, 2, \ldots, t$$

Quadratic objective with t linear constraints. A Quadratic Program

Maximum Margin Classification Training Optimization

Linearly separable case

Quadratic Programming

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|_2^2 \quad \text{s.t.} \quad y_i(\boldsymbol{x}_i^\top \boldsymbol{w}_i + b) \geqslant 1, \quad i = 1, 2, \ldots, t$$

Matrix form

All t constraint combined.

$$\min_{\boldsymbol{w},b} \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} \quad \text{s.t.} \quad \Delta(\boldsymbol{y})(X\boldsymbol{w} + \mathbf{1}b) \geqslant \mathbf{1}$$

where $\Delta(.)$ is an operator that puts a vector on the diagonal of a diagonal matrix. ("diag" function in Matlab.)

Prediction

For a new unseen test point $\boldsymbol{x}_\circ$, predict as:

$$\hat{y}(\boldsymbol{x}_\circ) = \begin{cases} +1 & \boldsymbol{x}_\circ^\top \boldsymbol{w} \geqslant 0 \\ -1, & \boldsymbol{x}_\circ^\top \boldsymbol{w} < 0 \end{cases}$$

Support Vectors

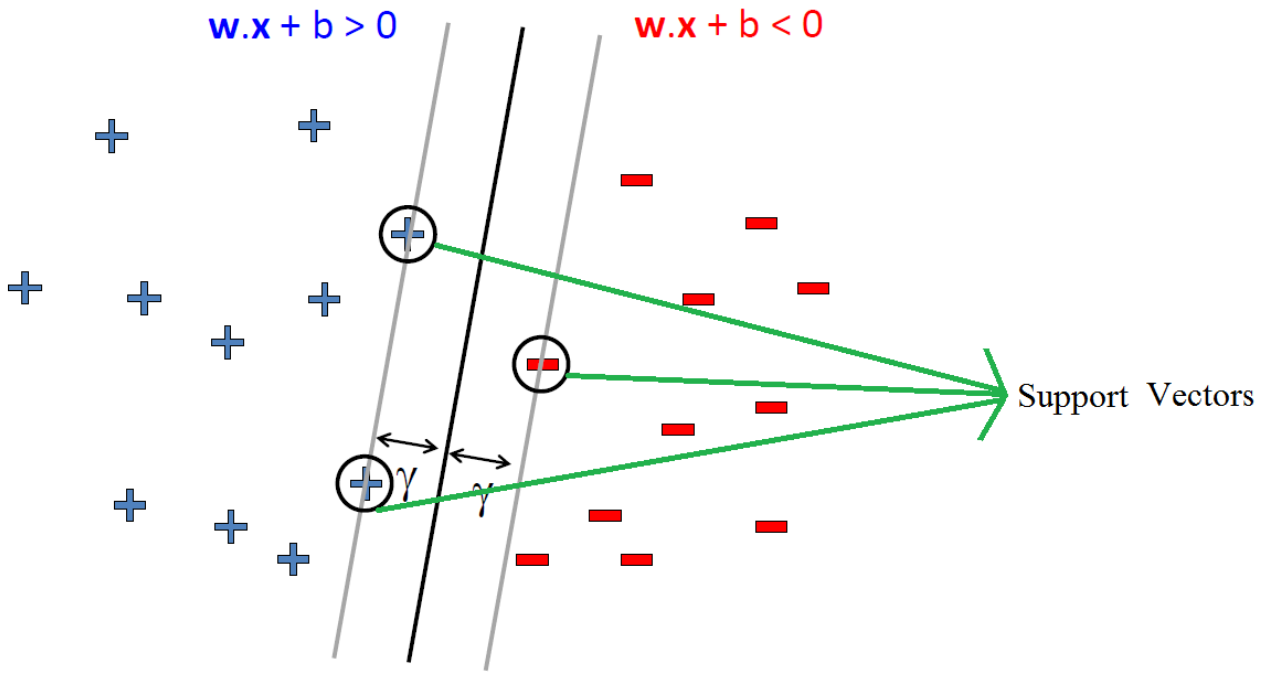


Figure from Prof Aarti Singh slides, CMU

Support Vectors: The subset of training examples that are closest to the optimal hyperplane defined by $\boldsymbol{w}$, b.

This give critical constraints! Removing other data doesn't change solution!

Support Vector Machines

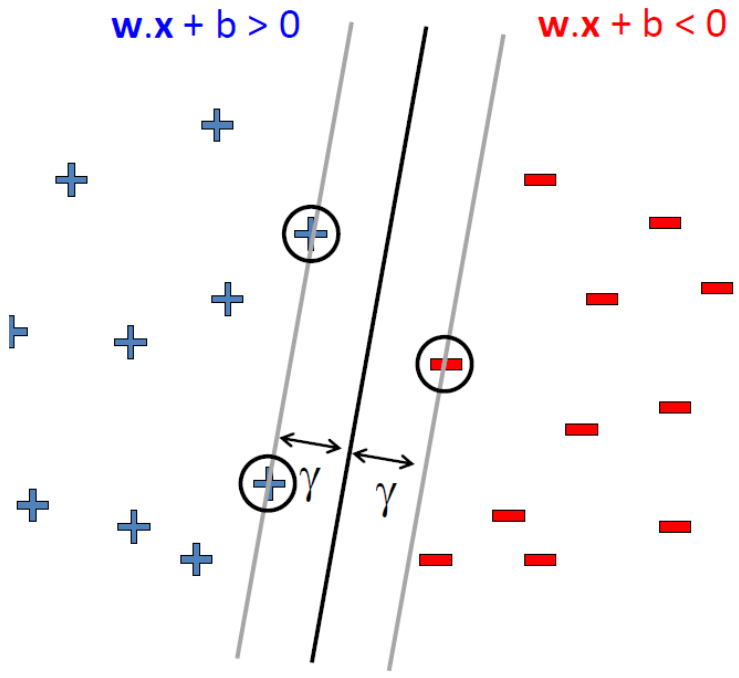


Figure from Prof Aarti Singh slides, CMU

- Data is summarized to its support vectors.
- Get a form of data compression.
- Provably good generalization if expected number of support vectors is much smaller than number of training examples.
- Thus called Support Vector Machines (SVMs).

Part Two

General Case: Maximum Margin Classification for Non-seprable Class Data

Non-separable case

- Noise and/or inconsistency in data
- Question: What if there is no consistent linear classifier
- Answer: Previous QP becomes infeasible. i.e will not have any solution
- Not good
- So what can one do?

Standard (heuristic) approach:

Add slack to margin

- Add a slack variable to each margin constraint.
- Try to maximize minimum margin minus a sum of slacks.
- Allow the maximum margin algorithm to give up on some bad training examples.

Much more robust against class noise!

Add slack to margins

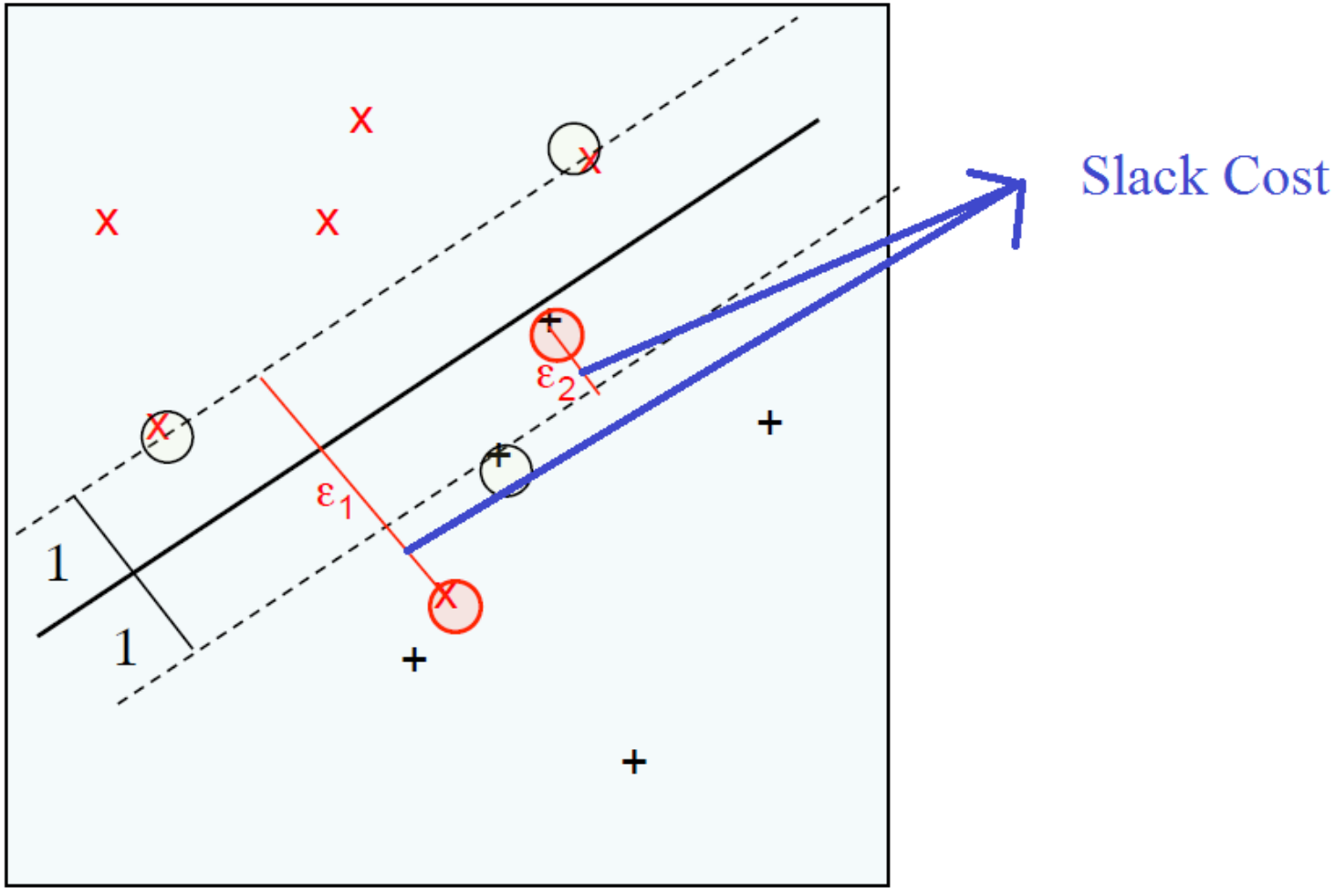


Figure from Prof David Helmbold Slides.

Quadratic Programming Algorithm with Slacks

$$\min_{\boldsymbol{w},b,\boldsymbol{\epsilon}} \quad \frac{\beta}{2}\|\boldsymbol{w}\|_2^2 + \mathbf{1}^\top \boldsymbol{\epsilon} \quad \text{s.t.} \quad \boldsymbol{\epsilon} \geqslant \mathbf{1} - \Delta(\boldsymbol{y})(X\boldsymbol{w} + \mathbf{1}b)$$

$$\boldsymbol{\epsilon} \geqslant \mathbf{0}$$

- Add a slack variable $\epsilon_i \geqslant 0$ for each training example $(X_{i:}, y_i)$
- Allow slack to reduce the margin, but at a cost (linear in mistake)
- β regularization (trade-off) parameter
- Called Soft margin Support Vector Machines

Hinge loss

- Note: The quadratic program that you get is exactly equivalent to minimizing a L_2 regularized form + a loss function

$$\min_{\boldsymbol{w},b} \frac{\beta}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i=1}^{t} \max(0, 1 - y_i(X_{i:}\boldsymbol{w} + b))$$

- The loss is called Hinge Loss

$$\text{hinge}(m) = \begin{cases} 1 - m & m \leqslant 1 \\ 0 & m > 1 \end{cases}$$

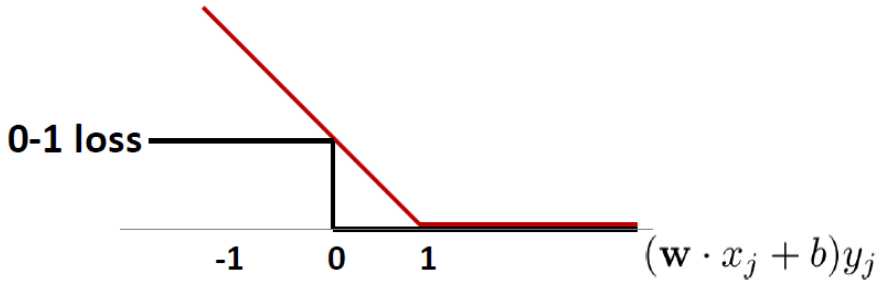


Figure from Prof Aarti Singh slides, CMU

- Again a loss + regularizer minimization as before

Summary

- We learned maximum margin classification:
  - Case 1: Linearly separable (hard margin sVM)
  - Case 2: Linearly non-separable (soft margin SVM)
- Showed that soft margin SVM is a regularized hinge loss.

Where to study?

- James et al (2013) Chap 9
- Hastie et al. 2nd ed (2009) Sec4.5.2, 12.2, 12.3
- Bishop (2006) Sec 7.1
- Duda et al. 2nd ed(2001) Chap 5
- Cherkassky & Mulier (1998) Chap 9
- Marsland (2009) Chap 5