Lab 2-1 Support Vector Machine

Tutorial for Optimization

DMKM Course

Xinyu WANG

Wangxinyu.xenia@gmial.com

24-Oct-2013

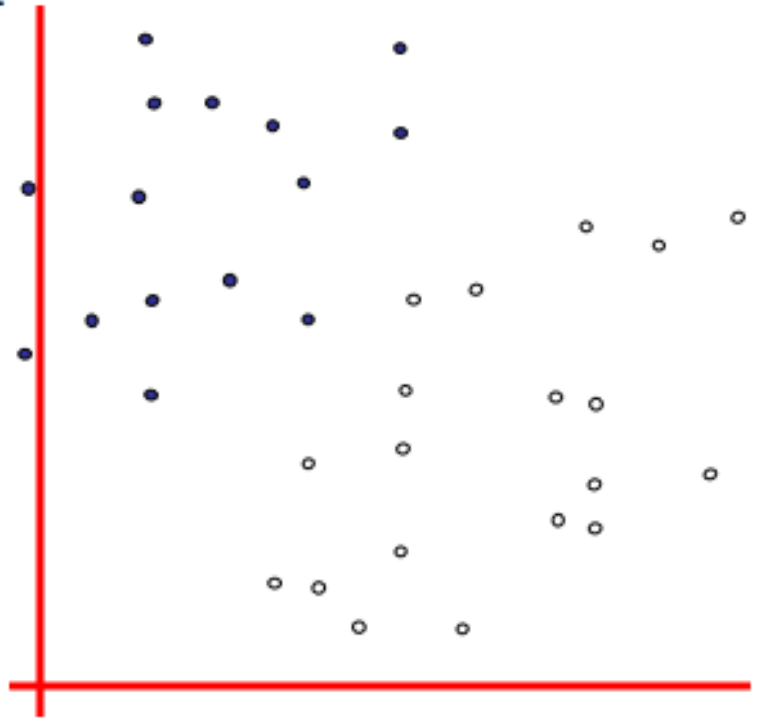
SVM

- Many believe it is “the best” supervised learning algorithm
- This tutorial: basic introduction + lab requirements
- Next tutorial: extension, duality, Kernal
- Goals: better understanding of SVM

Intuition

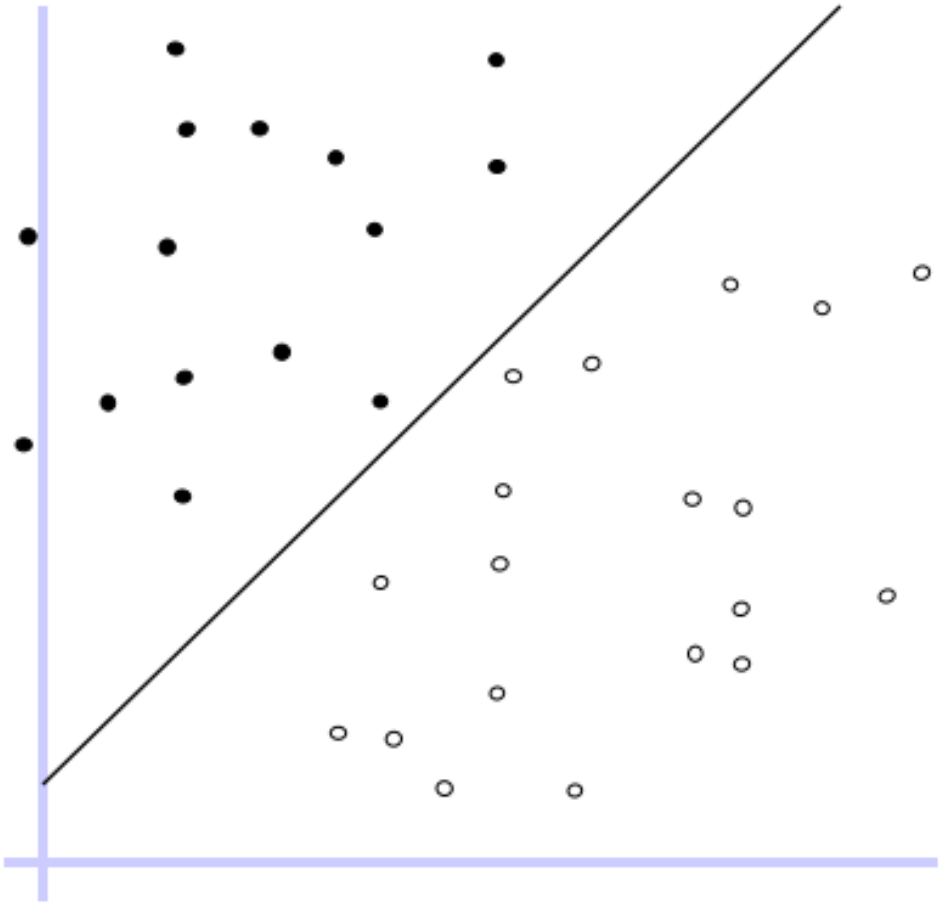
Class labels

- denotes +1
- denotes -1



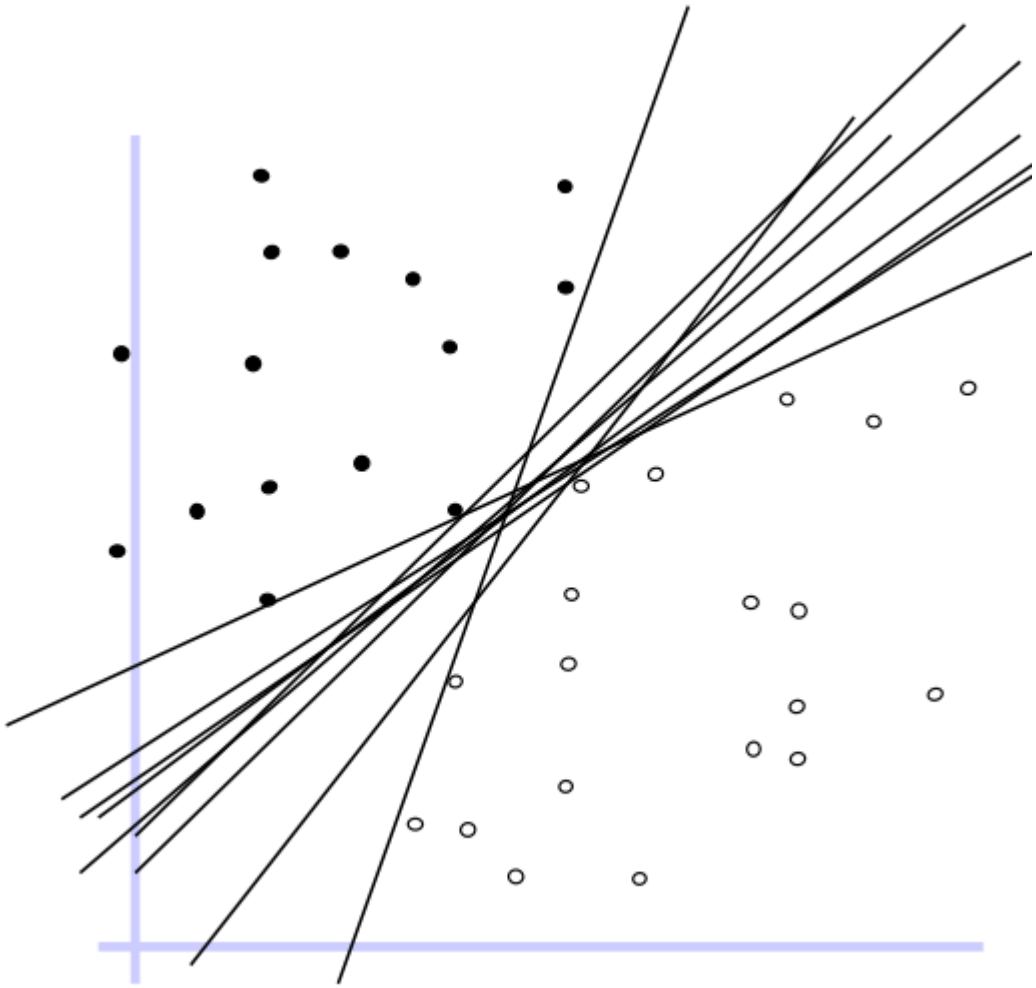
How would you classify this data?

Intuition



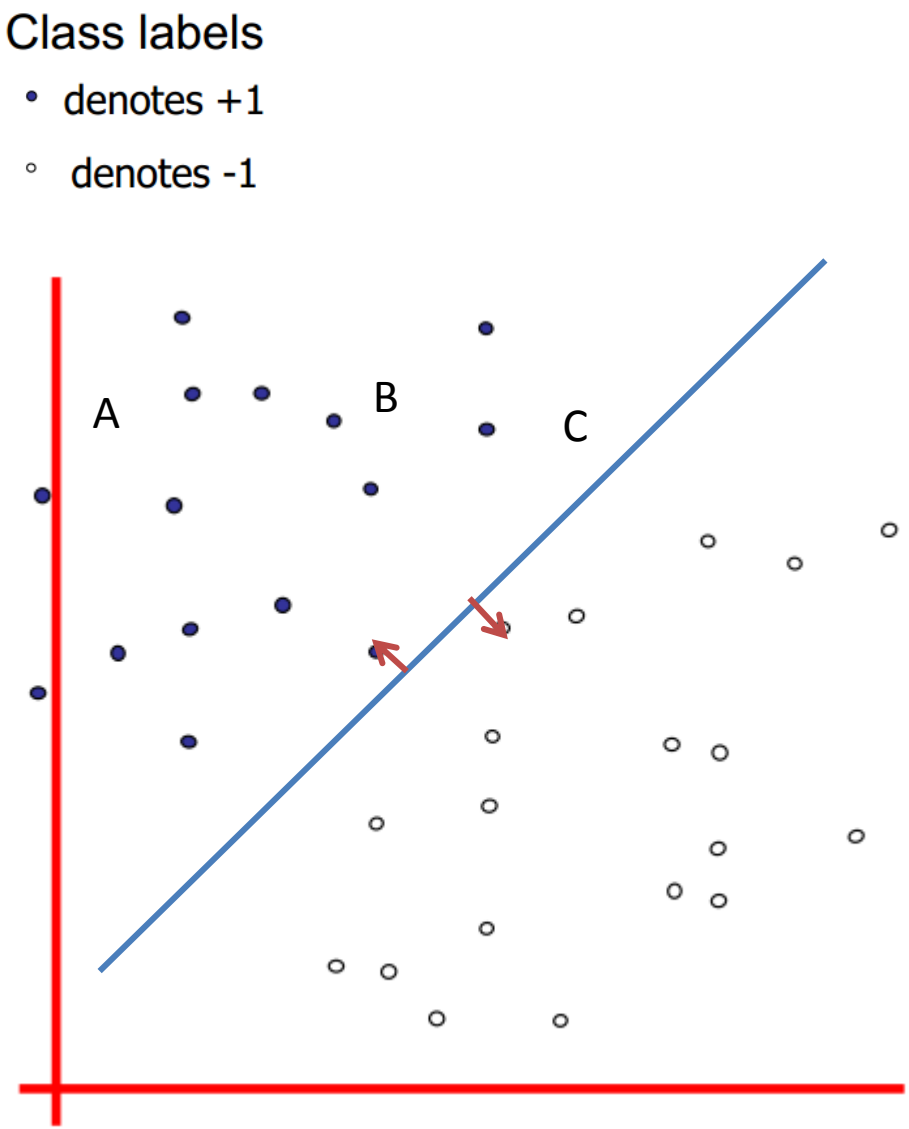
- For linearly separable data like this
- Draw a straight line
- It becomes our decision boundary

Intuition



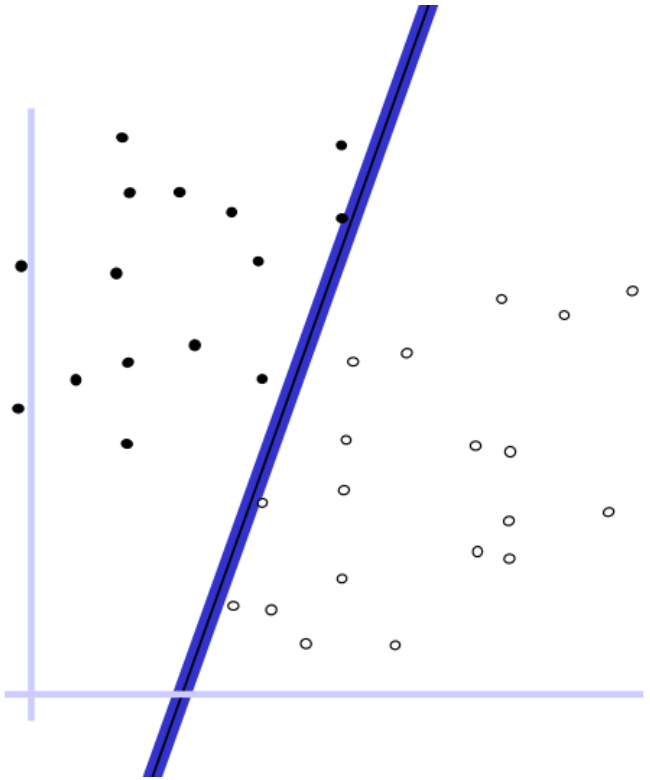
- There might be many lines
- Which one is the best?
- If a line could be expressed as a function of (w, b), where w controls the direction of the line and b controls its intercept
- What w and b for the best line?

Intuition

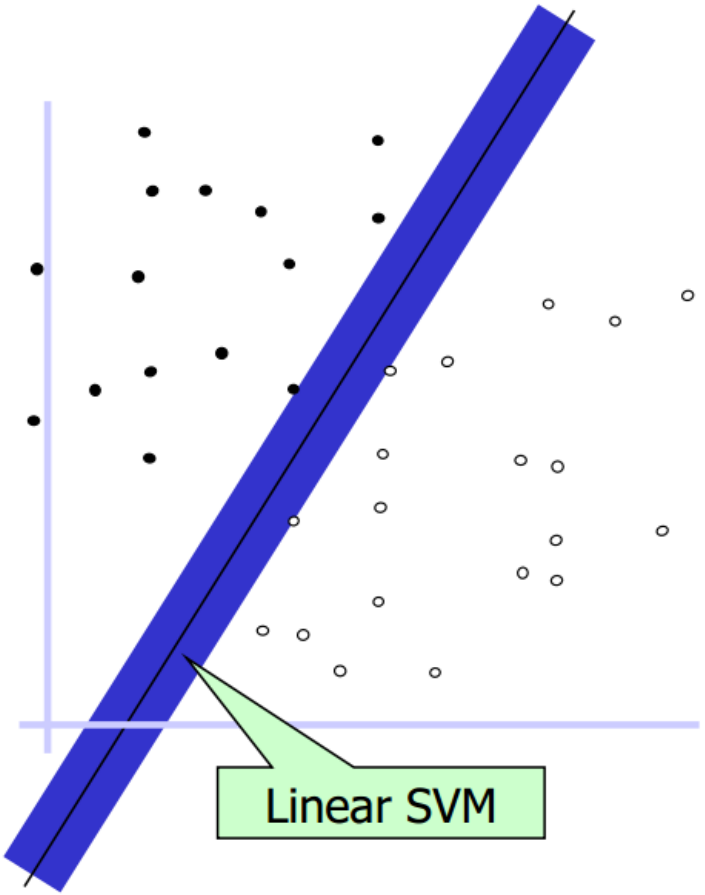


- If 3 new points would appear at positions A, B and C
- How confident we are to classify them into class +1?
- I would say:
  P(A=+1) > P(B=+1) > P(C=+1)
- Remember our decision boundary
- So further the point away from decision boundary, more confident we are in predication
- Let name the points that are closest to the boundary -- “support vectors”, and name the actual distance from them to the boundary -- “geographical margin”
- Then to make good predication, we just need the margin to be as big as possible

Intuition



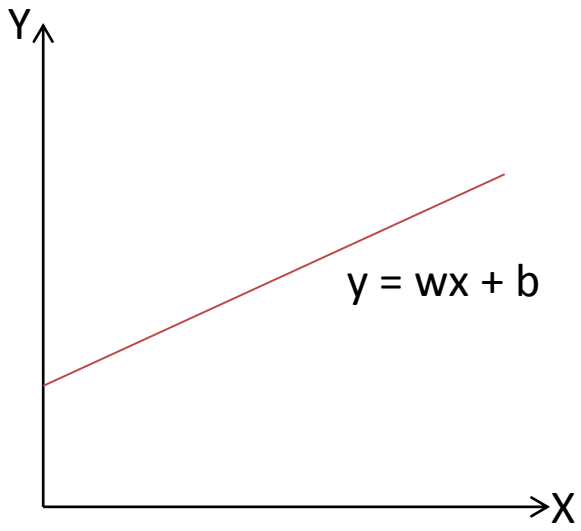
Is this line good?
Maybe, but not good enough!



This line is the best, for it has the widest margin among all lines. Finding its (w, b), our SVM problem solved.

Mathematics behind

So far, I hope you have established the concept of margin for SVM problems.

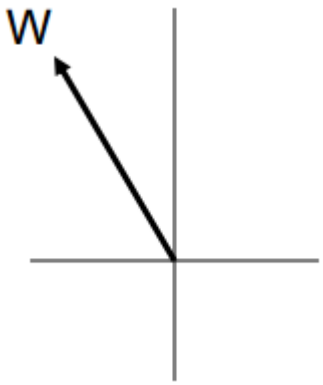


In 2-dimensional space, a line could be expressed as:

$$y = wx + b$$

In 3-dimensional space, this describes a plane, and in N-dimensional space, this describes a hyperplane.

Mathematics behind



- A vector is a vertical column.
- In d-dimensional space, a vector contains a list of d numbers . such as: W = (-1, 2, 4)', is a vector in 3-dimensional space.
- 3 numbers together define W's direction in this space,
- Norm ||W|| tells its length:

$$||W|| = \text{sqrt}(W'W)$$

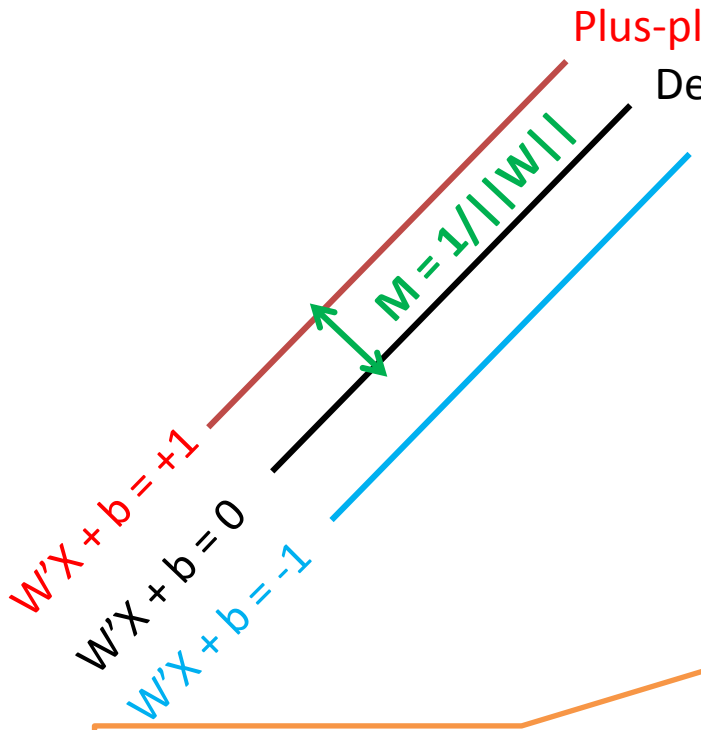
Inner produce:
$X'Y = \sum XiYi$

For rest of this tutorial, let's suppose we are in a 3-dimensional (or a higher dimensional) space, $X \in R^N$.

A plane (or hyperplace in N-dimensional space) goes through the original is: $W'X = 0$

More generally, we express it as: **$W'X + b = 0$**. Only when b = 0, it goes through the origin.

Mathematics behind



Suppose we have three parallel planes, and recall that we have data points of two classes:

Y = +1 and Y = -1,

A question is: how we can find (W, b) such that:

- all positive training points (X,Y=1) in the red zone
- all negative ones (X, Y=-1) in the blue zone,
- the margin M is maximized

In SVM, this margin is **1/||W||**, as it involves a lot of mathematical reasoning I will explain why it is so in next tutorial

Mathematically speaking, we want :

- W'Xi + b >= 1, if Yi=1
- W'Xi + b <= -1, if Yi=-1
- we can unify them: Yi(W'Xi+b) >= 1

SVM as constrained optimization

- Variables: W, b
- **Objective function**: maximize the margin M=1/||W||
- Equiv. to minimize ||W||, or $||W||^2$, or $\frac{1}{2}||W||^2$
- **Subject to** each training point on the correct side:

  $Yi(W'Xi+b) >= 1$

½ is only for mathematical convenience

Rewrite it, we have:

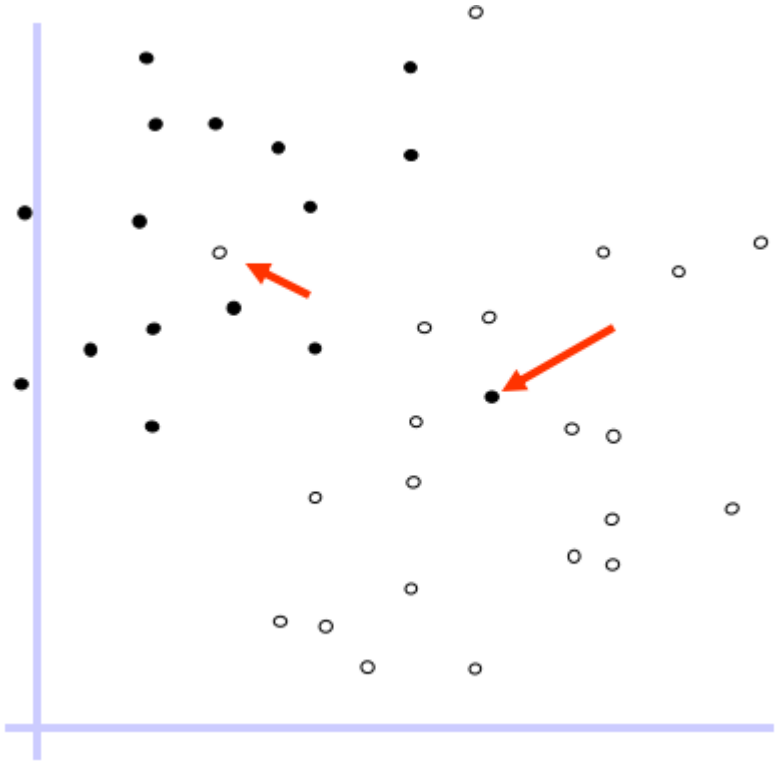
$$\min_{W,b} \frac{1}{2}||W||^2$$

Subject to $Yi\ (W'Xi + b) >= 1$, for all i

- It turns out to be a QP (Quadratic Program) problem, for the objective function is quadratic, convex and the constrains are linear. An efficient global solution exists.
- As conclusion: a SVM problem can be solved as a QP problem.

SVM for non-linearly separable case

Previously, we are using a linearly separable case, i.e., all points would be separated by a single line.
How about in a non-linearly separable case, such as this?



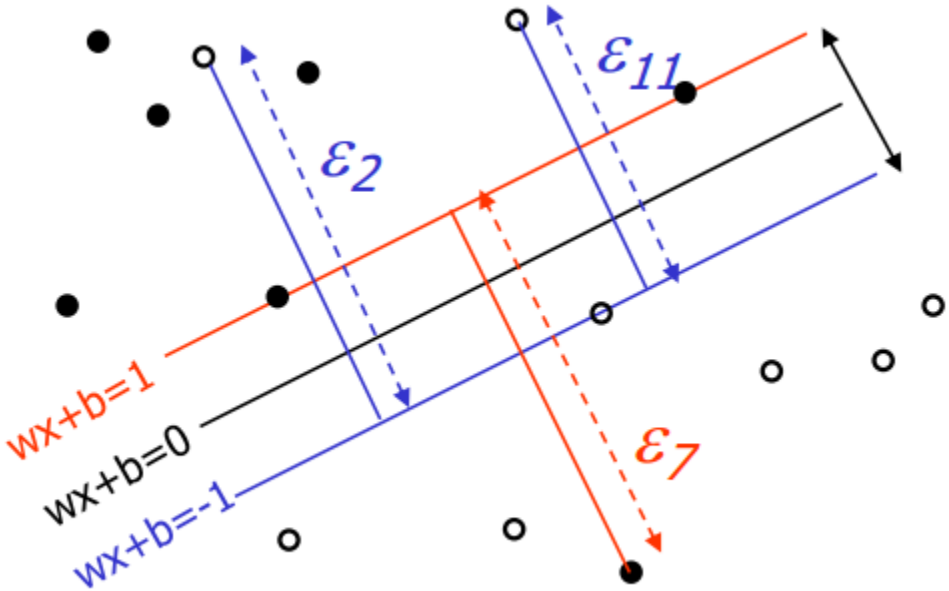
- Solution 1: relax the constraints, allowing a few “bad apples”.
- Solution 2: Map it into a higher dimension space. (Kernels)

Can we insist on Yi (W’Xi + b) >= 1, for all i?

Solution 1: slack variables

- For a given linear boundary W, b, we can compute how far off into the wrong side a bad point is. Let's call this difference "slack", expressed as ε.
- In the figure on right side, there are 3 special points. Two white points in black group, and one black one in white group.
- ε_2 indicates the distance from the white point to its "home line", so do ε_{11} and ε_7

In your lab this is represented by "y".



Q: To allow a few "bad apples", we should relax the constraints, but how to do this?

$$Yi\ (W'Xi + b) >= 1- \varepsilon_i$$
$$\varepsilon_i \geq 0$$

$$\begin{cases} \varepsilon_i \neq 0, i\ is\ in \text{ the wrong group} \\ \varepsilon_i = 0, i\ is\ in\ the\ good\ group \end{cases}$$

Solution 1: slack variables

Q: Any changes to the objective function?

$$\min_{W,b} \frac{1}{2}||W||^2 + C\sum \varepsilon_i$$

C is a trade-off parameter, as slack variables ε_i might be very small, C should be set large.

In your lab homework, I suggest you try different C, ranging from small number to big number to see the difference in the results.

As conclusion, the optimization problem for non-linear separable case in SVM is formulated as:

$$\min_{W,b} \frac{1}{2}||W||^2 + C\sum \varepsilon_i$$

Subject to

$$Yi\ (W'Xi + b) >= 1 - \varepsilon_i \text{ , } for\ all\ i$$

$$\varepsilon_i \geq 0 \text{ , } for\ all\ i$$

Back to your lab assignment

Formula A: we got from this tutorial

$\min_{W,b} \quad \frac{1}{2}||W||^2 + C\sum \varepsilon_i$

Subject to

Yi (W'Xi + b) >= 1- ε_i , *for all i*

$\varepsilon_i \geq 0$, *for all i*

I use this expression for the convenience for next tutorial

Formula B: given in the lab description

$$\min_{(w,\gamma,y)\in\mathbb{R}^{n+1+m}} \quad \nu e^T y + \frac{1}{2}||w||_2^2$$

$$\text{subject to} \quad D(Aw - e\gamma) + y \geq e,$$
$$y \geq 0,$$

The two formulas are quite the same things, the only difference lies in that formula B involves matrix:

- C on the left side matches *v* on the right side,
- Slack variables is expressed by "y" instead of ε_i in formula B,
- Intercept parameter b is changed to multiplication of e and $\boldsymbol{\gamma}$, where e is a vector of 1, and $\boldsymbol{\gamma}$ is a scalar.
- Yi in the left is actually D in the right, Xi is actually expressed as Aw in the right,
- W is replaced by w, but they mean the same.

Some comment

- If you feel it is difficult to understand the lab description ?
  - Don’t worry, I will explain it in another way in my next tutorial
- What you need to do for your homework is ?
  - Simple. Understand SVM (shown in my slides), and express the objective function and the constraint (in the lab description) in AMPL
- How to generate training data ?
  - The same as in Neural Network, but by using “gensvmdat”
- Some hints ?
  - Here is a simpler way. In the description, it says “e” is a vector of 1s, in AMPL you can simply use 1 instead; correspondingly, “D” can be expressed as a vector of -1 and 1 instead of matrix.
- Deadline of this homework ?
  - Within 1st week of 2nd tutorial of each lab session