ISML_3: Linear Classifier and (Linear) Support Vector Machine

Lingqiao Liu

Outlines

- Linear Classifier
- Support Vector Machine: Primal form
  - Separable case
  - Non-separable case
- Support Vector Machine: Dual form
  - KKT condition
  - Dual form of separable linear SVM
  - Dual form of non-separable linear SVM
- Solving optimization problem with CVX
- Summary

Requirements

- Basic concepts of linear classifier
- Basic ideas and concepts in SVMs
- Hard-margin SVMs:
  - formulation (but not derivation)
  - How to calculate margin and find support vectors
- Soft-margin SVMs:
  - Motivation
  - Formulation
  - Hinge loss formulation
- SVM dual
  - How to derive
  - Formulation
  - Relationship to primal problem solutions

Outlines

- Linear Classifier
- Support Vector Machine: Primal form
  - Separable case
  - Non-separable case
- Support Vector Machine: Dual form
  - KKT condition
  - Dual form of separable linear SVM
  - Dual form of non-separable linear SVM
- Solving optimization problem with CVX
- Summary

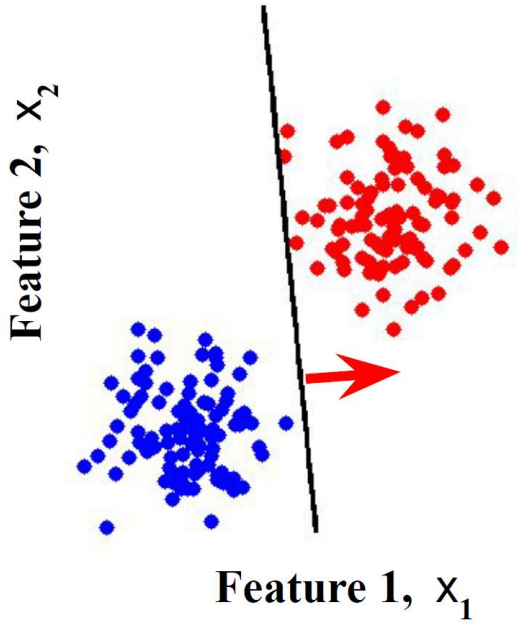
Linear classifier

- Suppose each data sample can be represented as a vector (called feature vector)
- In linear classifier, the classification can be done by calculating the linear combination of feature values

$$s_c = \sum_i w_i^c x_i + b_c = \mathbf{w_c}^\top \mathbf{x} + b_c$$

  - $\{\mathbf{w}_c, b_c\}$ are the parameters of the classifier
  - For binary classification, we only need a $\mathbf{w}, b$, the decision value is larger than 0, the prediction is class 1, otherwise class -1
  - For multi-class classification, we need to learn c sets of $\mathbf{w}, b$, one for each class. The class that gives the highest decision value will be classified as the predicted class

Geometrically

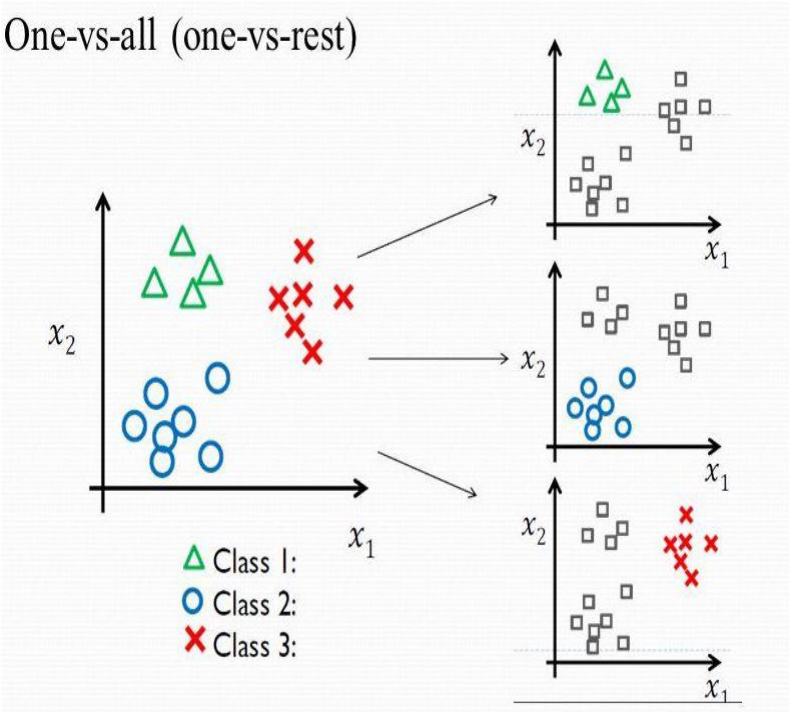


The linear boundary forms a Hyperplane in high-dimensional space

$$\mathbf{w}^\top \mathbf{x} + b = 0$$

Multiclass classification

- Multiclass classification can be converted into c binary classification problems by using the one-vs-rest scheme
  - For each problem, we use the c-th class as positive class and all the others as the negative class
  - We focus on binary classification today

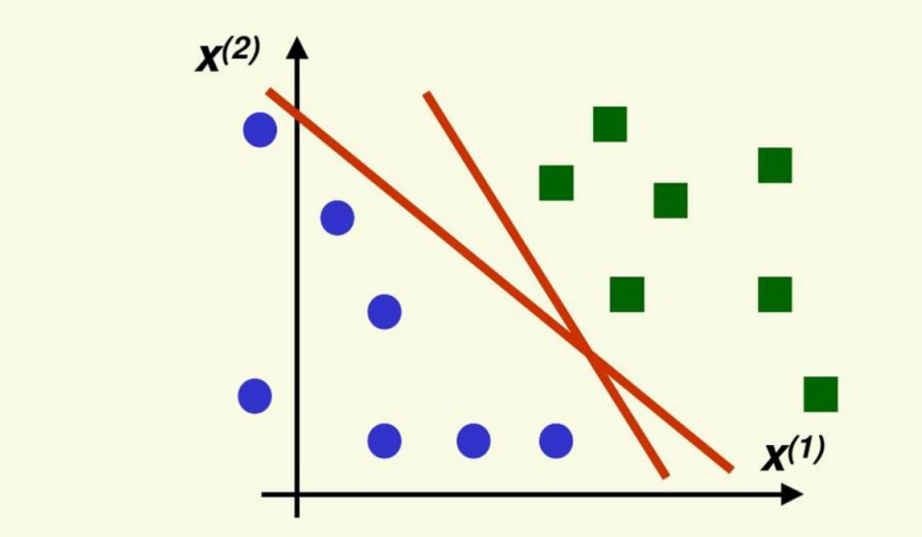


Outlines

- Linear Classifier
- Support Vector Machine: Primal form
  - Separable case
  - Non-separable case
- Support Vector Machine: Dual form
  - KKT condition
  - Dual form of separable linear SVM
  - Dual form of non-separable linear SVM
- Solving optimization problem with CVX
- Summary

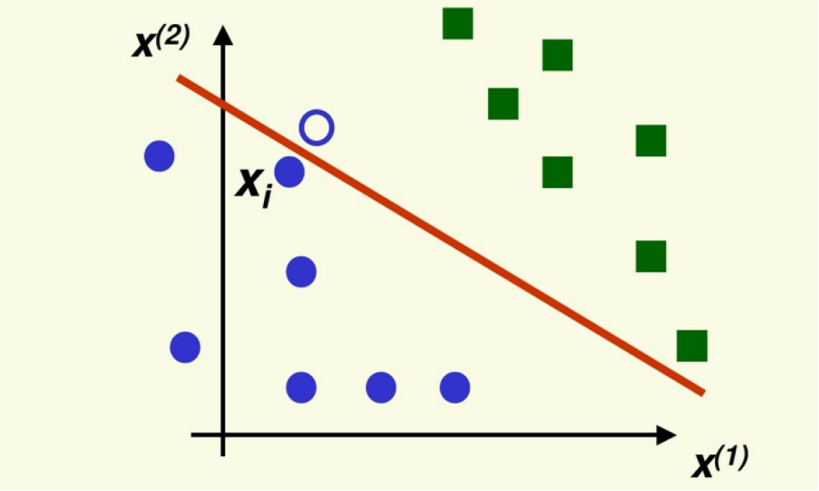
Motivation

- To separate two classes, which hyperplane we should choose



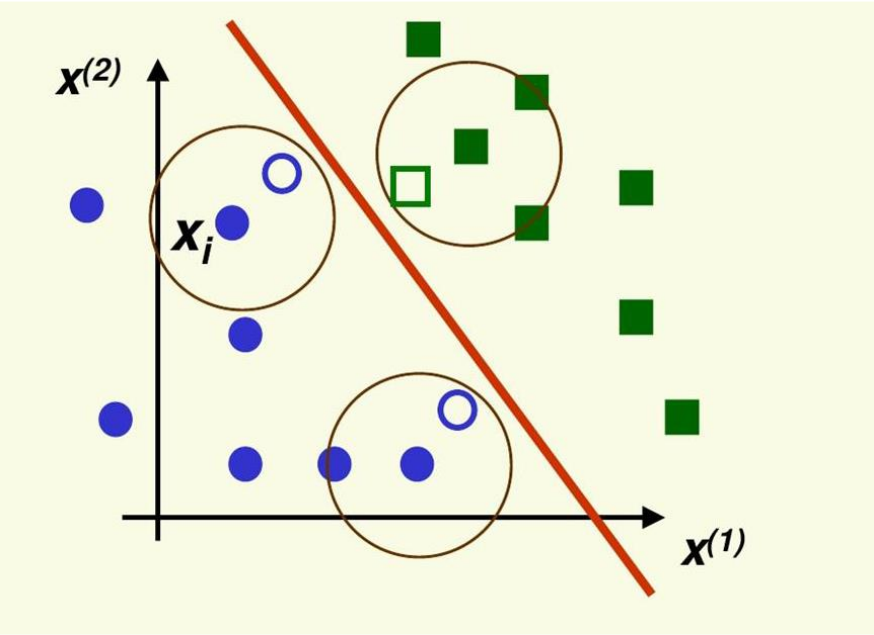
What makes a good separating hyperplane

- Training data is just a sampled subset of all possible data
- Suppose the hyperplane is close to one sample
- If we see new samples close to sample i, it is likely that it will be wrongly placed on the other side of hyperplane
  - Assumption: similar sample, similar label
- Lead to poor generalization



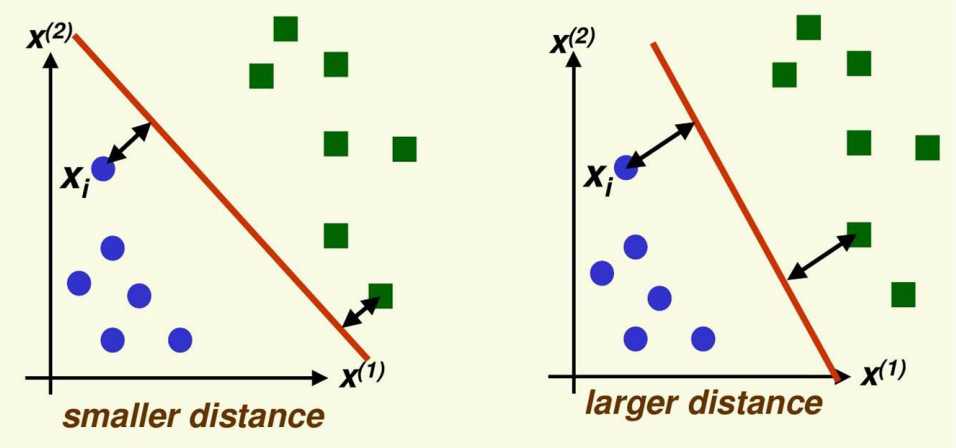
What makes a good separating hyperplane

- Hyperplane should be as far as possible from every sample
- In this way, new samples close to old samples will be classified correctly
- Good generalization



Support Vector Machine

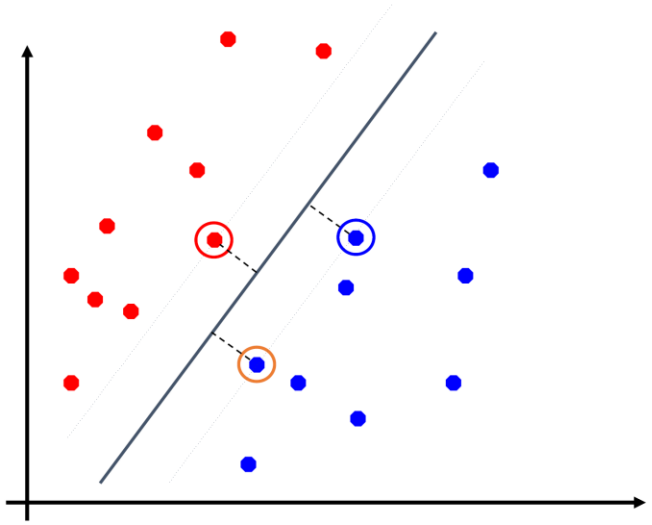
- Idea: maximize the distance to the closest example = margin



- This leads to an optimization problem

Properties of the optimal hyperplane

- Distance to the closest negative example = distance to the closest positive example
  - Why?
- Examples closest to the hyperplane are support vectors.
  - Removing non-support vector will not affect the optimal hyper-plane



Optimization problem for finding the optimal hyperplane

- The primal optimization problem for linear support vector machine (Hard-margin SVMs)

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2$$
$$s.t. \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \quad \forall i$$

- Decision rule

$$\hat{y} = 1 \quad if \quad \mathbf{w}^\top \mathbf{x} + b > 0$$
$$\hat{y} = -1 \quad if \quad \mathbf{w}^\top \mathbf{x} + b < 0$$

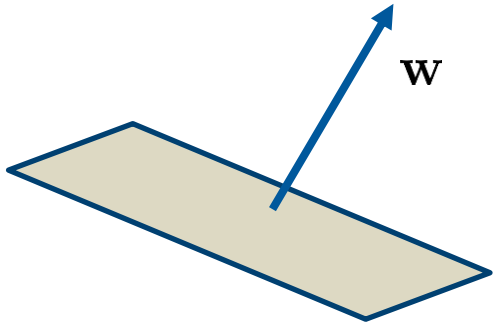
Proof: (1) Related Math Conclusions

- Cosine angle between two vectors

$$cosine(\mathbf{a}, \mathbf{b}) = \frac{<\mathbf{a},\mathbf{b}>}{\|\mathbf{a}\|_2\|\mathbf{b}\|_2} = \frac{\mathbf{a}^\top\mathbf{b}}{\|\mathbf{a}\|_2\|\mathbf{b}\|_2}$$

- Norm vector of a hyperplane

$$\mathbf{w}^\top\mathbf{x} + b = 0$$

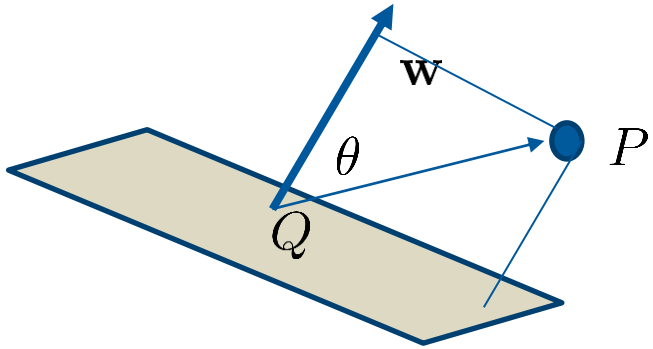


Proof: (1) Related Math Conclusions

- Distance between a point to a hyperplane

$$d = \frac{|\mathbf{w}^\top \mathbf{x}' + b|}{\|\mathbf{w}\|_2}$$

Proof:



$$\begin{aligned} d &= |PQ\cos(\theta)| \\ &= |\|(\mathbf{x}' - \mathbf{x})\|_2 \frac{\mathbf{w}^\top(\mathbf{x}' - \mathbf{x})}{\|\mathbf{w}\|_2 \|(\mathbf{x}' - \mathbf{x})\|_2}| \\ &= \frac{|\mathbf{w}^\top \mathbf{x}' - \mathbf{w}^\top \mathbf{x}|}{\|\mathbf{w}\|_2} = \frac{|\mathbf{w}^\top \mathbf{x}' - (\mathbf{w}^\top \mathbf{x} + b) + b|}{\|\mathbf{w}\|_2} = \frac{|\mathbf{w}^\top \mathbf{x}' + b|}{\|\mathbf{w}\|_2} \end{aligned}$$

Proof: (2) Original optimization problem

SVM objective

- (1) maximize the distance to the closest example
- (2) positive class and negative class on each side of the hyperplane

$$\max_{\mathbf{w},b,\eta} \eta$$

$$s.t. \quad \frac{|\mathbf{w}^\top \mathbf{x}_i + b|}{\|\mathbf{w}\|_2} \geq \eta$$

The minimal distance between hyper-plane

$$\mathbf{w}^\top \mathbf{x}_i + b \geq 0 \quad if \quad y_i > 0$$

$$\mathbf{w}^\top \mathbf{x}_i + b \leq 0 \quad if \quad y_i < 0$$

Correct classification

Proof: (2) transformed optimization problem

$$\max_{\mathbf{w},b,\eta} \eta$$

$$s.t. \quad \frac{|\mathbf{w}^\top \mathbf{x}_i + b|}{\|\mathbf{w}\|_2} \geq \eta$$

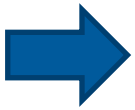
$$\mathbf{w}^\top \mathbf{x}_i + b \geq 0 \quad if \quad y_i > 0$$

$$\mathbf{w}^\top \mathbf{x}_i + b \leq 0 \quad if \quad y_i < 0$$

$$\max_{\mathbf{w},b,\eta} |\eta|$$

$$s.t. \quad \frac{y_i(\mathbf{w}^\top \mathbf{x}_i + b)}{\|\mathbf{w}\|_2} \geq |\eta|$$

$$\text{Assume } y_i = \{-1, 1\}$$



$$\min_{\mathbf{w},b,\eta} \frac{1}{\eta^2}$$

$$s.t. \quad \frac{y_i(\mathbf{w}^\top \mathbf{x}_i + b)}{\|\mathbf{w}\|_2} \geq |\eta|$$

Proof: (2) transformed optimization problem (continue)

$$\min_{\mathbf{w},b,\eta} \frac{1}{\eta^2}$$
$$s.t. \quad \frac{y_i(\mathbf{w}^\top \mathbf{x}_i + b)}{\|\mathbf{w}\|_2} \geq |\eta|$$

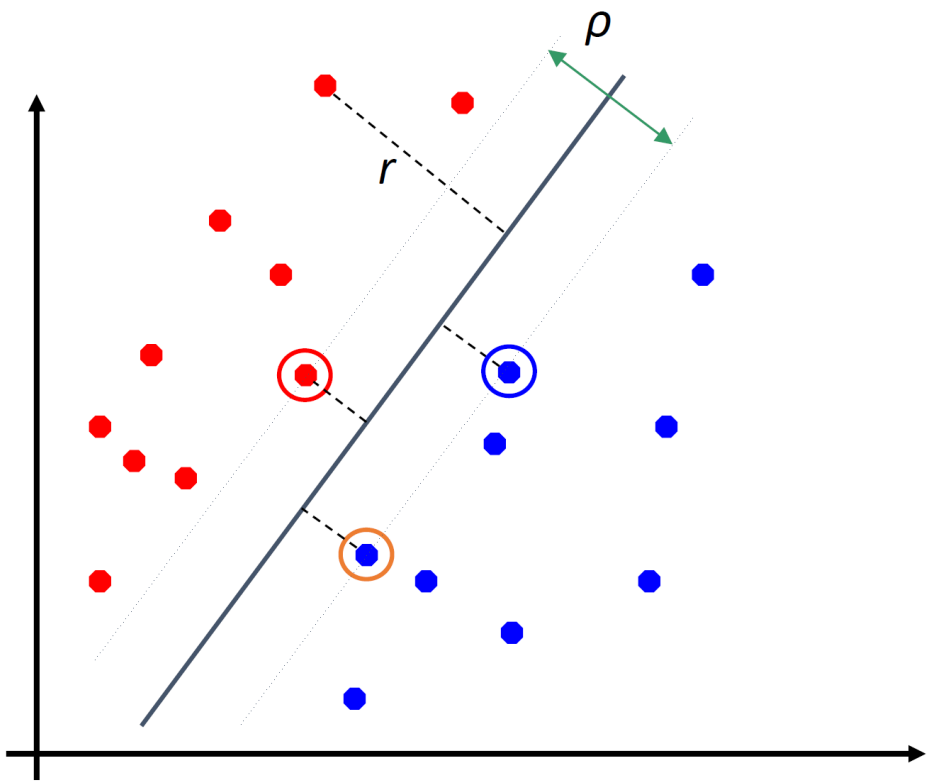
$$\min_{\mathbf{w},b,\eta} \frac{1}{\eta^2}$$
$$s.t. \quad y_i((\frac{\mathbf{w}}{\|\mathbf{w}\|_2|\eta|}^\top \mathbf{x}_i + \frac{b}{\|\mathbf{w}\|_2|\eta|}) \geq 1$$

define $\mathbf{w}' = \frac{\mathbf{w}}{\|\mathbf{w}\|_2|\eta|}$ and $b' = \frac{b}{\|\mathbf{w}\|_2|\eta|}$

$$\|\mathbf{w}'\|_2 = \frac{\|\mathbf{w}\|_2}{\|\mathbf{w}\|_2|\eta|} = \frac{1}{|\eta|}$$

$$\min_{\mathbf{w}',b'} \frac{1}{2}\|\mathbf{w}'\|_2^2$$
$$s.t. \quad y_i(\mathbf{w}'^\top \mathbf{x}_i + b') \geq 1$$

Take a closer look at the SVM problem



$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2$$

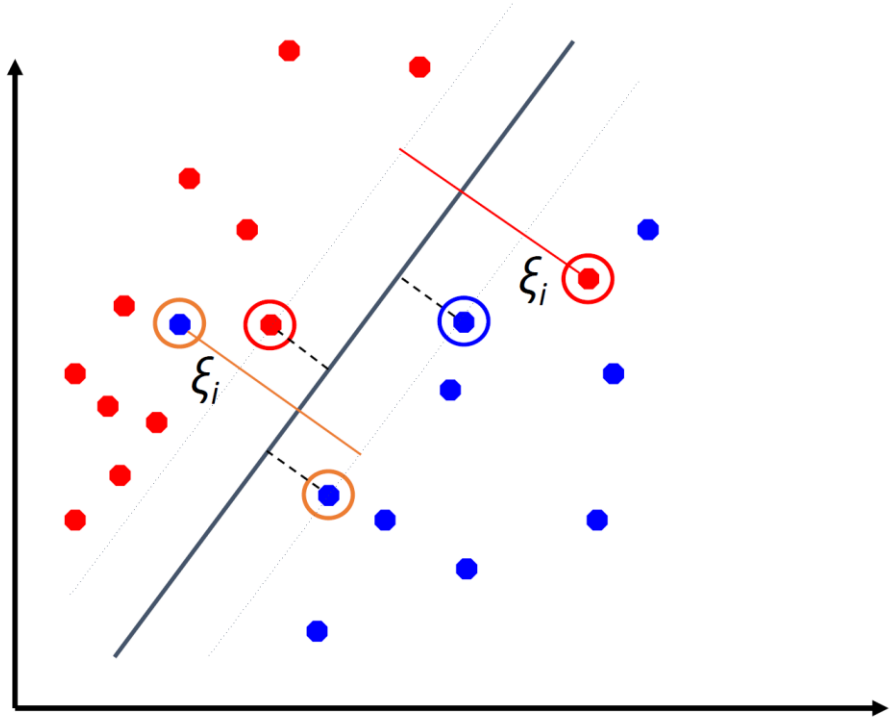
$$s.t. \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \quad \forall i$$

- Margin: $\rho = \frac{1}{\|\mathbf{w}\|_2}$
- All the support vectors are in

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) = 1$$

Non-separable case

What if the training set is not linearly separable?



$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$s.t. \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \quad \forall i$$

In that case, not all inequality constraints can be satisfied

Slack variables and soft-margin SVM

- Slack variables indicates how much a sample violates the inequality constraint

$$\min_{\mathbf{w},b,\{\xi_i\}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_i \xi_i$$

$$s.t. \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \forall i$$

$$\xi_i \geq 0$$

- And we want to minimize the total violation
- C is a hyper-parameter represents how much violation allowed.
  - Question: How to choose C?

Hinge loss view of soft-margin SVM

- The soft-margin SVM can also be viewed from the perspective of hinge-loss

$$\min_{\mathbf{w},b,\{\xi_i\}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_i \xi_i$$
$$s.t. \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \forall i$$
$$\xi_i \geq 0$$

$$\min_{\mathbf{w},b,\{\xi_i\}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_i \xi_i$$
$$s.t. \quad 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b) \leq \xi_i, \quad \forall i$$
$$0 \leq \xi_i$$

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_i \max(0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b))$$

Hinge loss view of soft-margin SVM

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_i \max(0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b))$$

- If $y_i(\mathbf{w}^\top \mathbf{x}_i + b) > 1$, which means the sample has been correctly classified with sufficient confidence (not just >0, but > 1), then the loss incurred for this sample is 0
- This means, if the "confidence" is large enough, we do not pursuit a larger "confidence".
- Instead, we focus on those samples that haven't been confidently classified.

Outlines

- Linear Classifier
- Support Vector Machine: Primal form
  - Separable case
  - Non-separable case
- Support Vector Machine: Dual form
  - KKT condition
  - Dual form of separable linear SVM
  - Dual form of non-separable linear SVM
- Solving optimization problem with CVX
- Summary

SVM dual problem

- Sometimes it is beneficial to convert the SVM optimization problems into another equivalent optimization problem called the dual formulation of SVM, or dual problem of SVM
- By applying certain transforms, we can derive a dual problem for any optimization problem.
  - An important type of dual problem is called Lagrangian dual problem
  - If the original problem is convex, then there are some equivalence between the optimal solutions and the original problem. In other words, a convex optimization problem could be solved in its original (called primary form) form or its dual form

Convex optimization

$$\min_{x \in \Omega} f(x)$$

$$\begin{aligned} \min_{x} \; & f(x) \\ s.t. \quad & g(x) \leq 0 \\ & h(x) = 0 \end{aligned}$$

An optimization is a convex optimization problem if $f(x)$ is a convex function and Ω is a convex set

Convex function: $$f(\theta x + (1-\theta)y) \geq \theta f(x) + (1-\theta) f(y)$$

Convex Set: $$If \quad x, y \in \Omega, \quad then \quad \theta x + (1-\theta) y \in \Omega$$

If f is a quadratic function and g is linear function, the problem is convex.

Lagrangian dual problem

- Original problem (Primal problem)

$$\min_{\mathbf{x}} f(\mathbf{x})$$
$$s.t \quad g_i(\mathbf{x}) \leq 0$$

Assume $f(x)$ and $g_i(x)$ are convex functions, e.g., linear or quadratic functions

Define

Lagrangian function

$$L(\mathbf{x}, \{\mu_i\}) = f(\mathbf{x}) + \sum_i \mu_i g_i(\mathbf{x})$$

- Lagrangian dual problem

$$\max_{\{\mu_i\}} \left(\min_{\mathbf{x}} L(\mathbf{x}, \{\mu_i\}) \right)$$
$$s.t \quad \mu_i \geq 0$$

$\mathbf{x}$: Primal variables

μ: Dual variables

Relationship between dual problem and primal problem

- Notation: using $\mathbf{x}^*$ and μ^* denote the optimal solution for the primal problem and dual problem, respectively
- Note that the inner problem is an unconstrained optimization problem, so we should have

$$\max_{\{\mu_i\}} \left(\min_{\mathbf{x}} L(\mathbf{x}, \{\mu_i\}) \right) \quad \nabla L(\mathbf{x}^*, \{\mu_i\}) = \nabla f(\mathbf{x}^*) + \sum_i \mu_i \nabla g_i(\mathbf{x}^*) = 0$$

$$s.t \quad \mu_i \geq 0$$

- Complementary slackness $\mu_i^* g_i(\mathbf{x}^*) = 0 \quad \forall\ i$
- More information (if you want to know "why") can be found in this tutorial

lagrangian_duality.pdf (stanford.edu)

Summary: Karush–Kuhn–Tucker (KKT) conditions

Relationship between the solution to the primal problem and dual problem can be specified by the KKT condition. The dual problem can be derived by applying KKT condition and make some transformations

1. Stationarity

$$\nabla L(\mathbf{x}^*, \{\mu_i\}) = \nabla f(\mathbf{x}^*) + \sum_i \mu_i \nabla g_i(\mathbf{x}^*) = 0$$

2. Primary feasibility

$$g_i(\mathbf{x}^*) \leq 0 \quad \forall\ i$$

3. Dual feasibility

$$\mu_i^* \geq 0 \quad \forall\ i$$

4. Complementary slackness

$$\mu_i^* g_i(\mathbf{x}^*) = 0 \quad \forall\ i$$

Derive dual problem for soft-margin SVMs

$$\min_{\mathbf{w},b,\{\xi_i\}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_i \xi_i$$

$$s.t. \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \forall i$$

$$\xi_i \geq 0$$

Step 1: take derivative over Lagrangian function

$$L(\mathbf{w}, b, \{\xi_i\}, \{\alpha_i\}, \{\beta_i\})$$

$$= \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_i \xi_i + \sum_i \alpha_i(1 - \xi_i - y_i(\mathbf{w}^\top \mathbf{x}_i + b)) - \sum_i \beta_i \xi_i$$

1. Stationarity $\nabla L(\mathbf{x}^*, \{\mu_i\}) = \nabla f(\mathbf{x}^*) + \sum_i \mu_i \nabla g_i(\mathbf{x}^*) = 0$

$$\frac{\partial L}{\partial \mathbf{w}}|_{\mathbf{w}=\mathbf{w}^*} = \mathbf{w}^* - \sum_i \alpha_i y_i \mathbf{x}_i = 0 \quad \Rightarrow \quad \mathbf{w}^* = \sum_i \alpha_i y_i \mathbf{x}_i$$

$$\frac{\partial L}{\partial b}|_{b=b^*} = -\sum_i \alpha_i y_i = 0 \quad \Rightarrow \quad \sum_i \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \xi_i}|_{\xi_i=\xi_i^*} = C - \alpha_i - \beta_i = 0 \quad \Rightarrow \quad \alpha_i = C - \beta_i$$

Derive dual problem for soft-margin SVMs

$$L(\mathbf{w}, b, \{\xi_i\}, \{\alpha_i\}, \{\beta_i\})$$
$$= \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_i \xi_i + \sum_i \alpha_i(1 - \xi_i - y_i(\mathbf{w}^\top \mathbf{x}_i + b)) - \sum_i \beta_i \xi_i$$

Step 2: represent primal variables with dual variables, substitute into the Lagrangian function

$$\mathbf{w}^* = \textstyle\sum_i \alpha_i y_i \mathbf{x}_i \qquad \sum_i \alpha_i y_i = 0 \qquad \alpha_i = C - \beta_i$$

$$\tfrac{1}{2}\|\mathbf{w}^*\|_2^2 = \tfrac{1}{2}(\textstyle\sum_{i=1}^N \alpha_i y_i \mathbf{x}_i)^\top (\sum_{i=1}^N \alpha_i y_i \mathbf{x}_i) = \tfrac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j < \mathbf{x}_i, \mathbf{x}_j >$$

$$\textstyle\sum_i \alpha_i(1-\xi_i-y_i(\mathbf{w}^\top \mathbf{x}_i+b)) = \sum_i \alpha_i - \sum_i \alpha_i \xi_i - \sum_i \alpha_i y_i (\sum_j \alpha_j y_j \mathbf{x}_j)^\top \mathbf{x}_i + b))$$

$$\textstyle\sum_i \alpha_i y_i (\sum_j \alpha_j y_j \mathbf{x}_j)^\top \mathbf{x}_i = \sum_i \sum_j \alpha_i \alpha_j y_i y_j < \mathbf{x}_j, \mathbf{x}_i >$$

Derive dual problem for soft-margin SVMs

$$\sum_i \alpha_i(1-\xi_i-y_i(\mathbf{w}^\top\mathbf{x}_i+b)) = \sum_i \alpha_i - \sum_i \alpha_i\xi_i - \sum_i \alpha_i y_i(\sum_j \alpha_j y_j \mathbf{x}_j)^\top \mathbf{x}_i + b))$$

$$= \sum_i \alpha_i - \sum_i \alpha_i\xi_i - \sum_i\sum_j \alpha_i\alpha_j y_i y_j < \mathbf{x}_i, \mathbf{x}_j > -b(\sum_i \alpha_i y_i)$$

$$= \sum_i \alpha_i - \sum_i \alpha_i\xi_i - \sum_i\sum_j \alpha_i\alpha_j y_i y_j < \mathbf{x}_i, \mathbf{x}_j >$$

$$\frac{\partial L}{\partial \mathbf{w}}|_{\mathbf{w}=\mathbf{w}^*} = \mathbf{w}^* - \sum_i \alpha_i y_i \mathbf{x}_i = 0 \quad \Rightarrow \quad \mathbf{w}^* = \sum_i \alpha_i y_i \mathbf{x}_i$$

$$\frac{\partial L}{\partial b}|_{b=b^*} = -\sum_i \alpha_i y_i = 0 \quad \Rightarrow \quad \sum_i \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \xi_i}|_{\xi_i=\xi_i^*} = C - \alpha_i - \beta_i = 0 \quad \Rightarrow \quad \alpha_i = C - \beta_i$$

Derive dual problem for soft-margin SVMs

$$L(\mathbf{w}, b, \{\xi_i\}, \{\alpha_i\}, \{\beta_i\})$$
$$= \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_i \xi_i + \sum_i \alpha_i(1 - \xi_i - y_i(\mathbf{w}^\top \mathbf{x}_i + b)) - \sum_i \beta_i \xi_i$$

$$\frac{1}{2}\|\mathbf{w}^*\|_2^2 = \frac{1}{2}(\sum_{i=1}^N \alpha_i y_i \mathbf{x}_i)^\top (\sum_{i=1}^N \alpha_i y_i \mathbf{x}_i) = \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j < \mathbf{x}_i, \mathbf{x}_j >$$

$$\sum_i \alpha_i(1 - \xi_i - y_i(\mathbf{w}^\top \mathbf{x}_i + b)) = \sum_i \alpha_i - \sum_i \alpha_i \xi_i - \sum_i \sum_j \alpha_i \alpha_j y_i y_j < \mathbf{x}_i, \mathbf{x}_j >$$

Derive dual problem for soft-margin SVMs

$$L(\mathbf{w}, b, \{\xi_i\}, \{\alpha_i\}, \{\beta_i\})$$

$$= \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_i \xi_i + \sum_i \alpha_i(1 - \xi_i - y_i(\mathbf{w}^\top \mathbf{x}_i + b)) - \sum_i \beta_i \xi_i$$

$$\frac{1}{2}\|\mathbf{w}^*\|_2^2 = \frac{1}{2}(\sum_{i=1}^N \alpha_i y_i \mathbf{x}_i)^\top (\sum_{i=1}^N \alpha_i y_i \mathbf{x}_i) = \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j < \mathbf{x}_i, \mathbf{x}_j >$$

$$\sum_i \alpha_i(1 - \xi_i - y_i(\mathbf{w}^\top \mathbf{x}_i + b)) = \sum_i \alpha_i - \sum_i \alpha_i \xi_i - \sum_i \sum_j \alpha_i \alpha_j y_i y_j < \mathbf{x}_i, \mathbf{x}_j >$$

$$\sum_i (C - \alpha_i - \beta_i)\xi_i = 0$$

$$\frac{\partial L}{\partial \mathbf{w}}\Big|_{\mathbf{w}=\mathbf{w}^*} = \mathbf{w}^* - \sum_i \alpha_i y_i \mathbf{x}_i = 0 \quad \Rightarrow \quad \mathbf{w}^* = \sum_i \alpha_i y_i \mathbf{x}_i$$

$$\frac{\partial L}{\partial b}\Big|_{b=b^*} = -\sum_i \alpha_i y_i = 0 \quad \Rightarrow \quad \sum_i \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \xi_i}\Big|_{\xi_i=\xi_i^*} = C - \alpha_i - \beta_i = 0 \quad \Rightarrow \quad \alpha_i = C - \beta_i$$

Derive dual problem for soft-margin SVMs

$$L(\mathbf{w}, b, \{\xi_i\}, \{\alpha_i\}, \{\beta_i\})$$
$$= \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_i \xi_i + \sum_i \alpha_i(1 - \xi_i - y_i(\mathbf{w}^\top \mathbf{x}_i + b)) - \sum_i \beta_i \xi_i$$

$$\frac{1}{2}\|\mathbf{w}^*\|_2^2 = \frac{1}{2}\left(\sum_{i=1}^N \alpha_i y_i \mathbf{x}_i\right)^\top \left(\sum_{i=1}^N \alpha_i y_i \mathbf{x}_i\right) = \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j < \mathbf{x}_i, \mathbf{x}_j >$$

$$\sum_i \alpha_i(1 - \xi_i - y_i(\mathbf{w}^\top \mathbf{x}_i + b)) = \sum_i \alpha_i - \sum_i \alpha_i \xi_i - \sum_i \sum_j \alpha_i \alpha_j y_i y_j < \mathbf{x}_i, \mathbf{x}_j >$$

$$L(\mathbf{w}, b, \{\xi_i\}, \{\alpha_i\}, \{\beta_i\}) = \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j < \mathbf{x}_i, \mathbf{x}_j >$$

Derive dual problem for soft-margin SVMs

Step 3: Check constraints by using dual feasibility

Any constraints? Dual variables must be no less than 0

$$\alpha_i \geq 0, \beta_i \geq 0 \xrightarrow{\alpha_i = C - \beta_i} \boxed{0 \leq \alpha_i \leq C}$$

$$\frac{\partial L}{\partial \mathbf{w}}|_{\mathbf{w}=\mathbf{w}^*} = \mathbf{w}^* - \sum_i \alpha_i y_i \mathbf{x}_i = 0 \Rightarrow \mathbf{w}^* = \sum_i \alpha_i y_i \mathbf{x}_i$$

$$\frac{\partial L}{\partial b}|_{b=b^*} = -\sum_i \alpha_i y_i = 0 \Rightarrow \sum_i \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \xi_i}|_{\xi_i=\xi_i^*} = C - \alpha_i - \beta_i = 0 \Rightarrow \alpha_i = C - \beta_i$$

$$\boxed{\sum_i \alpha_i y_i = 0}$$

SVM dual problem

- Dual problem

$$\max_{\alpha_i} \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j < \mathbf{x}_i, \mathbf{x}_j >$$
$$s.t. \quad 0 \leq \alpha_i \leq C$$
$$\sum_i \alpha_i y_i = 0$$

- Primal problem

$$\min_{\mathbf{w}, b, \{\xi_i\}} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_i \xi_i$$
$$s.t. \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \forall i$$
$$\xi_i \geq 0$$

Convert dual solutions to primal solutions

$$\mathbf{w}^* = \sum_i \alpha_i^* y_i \mathbf{x}_i$$

- How about b?
  - By using complementary slackness $\mu_i g_i(\mathbf{x}) = 0 \quad \forall\ i$

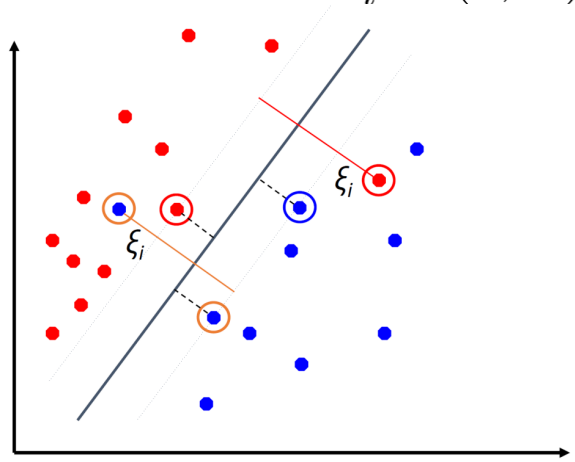
$$\alpha_i^* \left(1 - y_i(\mathbf{w}^{*\top}\mathbf{x}_i + b^*) - \xi_i^*\right) = 0$$
$$\beta_i^* \xi = (C - \alpha_i^*)\xi_i^* = 0$$

Convert dual solutions to primal solutions

- A closer look at complementary slackness condition

$$\alpha_i^* \left(1 - y_i(\mathbf{w}^{*\top}\mathbf{x}_i + b^*) - \xi_i^*\right) = 0$$
$$\beta_i^* \xi = (C - \alpha_i^*)\xi_i^* = 0$$

  - If $y_i(\mathbf{w}^{*\top}\mathbf{x}_i + b^*) \geq 1$ then $\xi_i^* = 0$, then $\alpha_i^* = 0$
  - If $y_i(\mathbf{w}^{*\top}\mathbf{x}_i + b^*) < 1$ then $\xi_i^* \neq 0$, then $\alpha_i^* = C$
  - How about $\alpha_i^* \in (0, C)$? then $\xi_i^* = 0$, then $y_i(\mathbf{w}^{*\top}\mathbf{x}_i + b^*) = 1$



$$\mathbf{w}^* = \sum_i \alpha_i^* y_i \mathbf{x}_i$$

Points satisfying the last two conditions are called support vectors

How to solve b

- Find any sample with its corresponding $\alpha_i^* \in (0, C)$, solving b by

$$y_i(\mathbf{w}^{*\top}\mathbf{x}_i + b^*) = 1$$

$$b^* = y_i - \mathbf{w}^{*\top}\mathbf{x}_i$$

SVM dual problem

- Primal problem
  - Number of variables is proportional to the number of feature dimensions
  - Number of constraints is proportional to the number of samples
- Dual problem
  - Number of variables is proportional to the number of samples
- Solving SVM from dual or primal?
  - Depends which one results in smaller optimization problem, i.e., number of variables to optimize
  - Efficient solutions have been developed by using the relationship between dual problem and primal problem

Outlines

- Linear Classifier
- Support Vector Machine: Primary form
  - Separable case
  - Non-separable case
- Support Vector Machine: Dual form
  - KKT condition
  - Dual form of separable linear SVM
  - Dual form of non-separable linear SVM
- Solving optimization problem with CVX
- Summary

Optimization solver

- Optimization problem is usually solved by existing solver
- For the optimization problem in SVMs
  - Using standard convex optimization solver
    - CVX Linear program — CVXPY 1.1.13 documentation
  - Using efficient solver specifically developed for SVMs
    - LIBSVM -- A Library for Support Vector Machines (ntu.edu.tw)
    - LIBLINEAR -- A Library for Large Linear Classification (ntu.edu.tw)

How to use CVX for solving optimization problem

- Scientific programming in Python
  - Commonly used packages: Numpy, Scipy
  - NumPy Tutorial (tutorialspoint.com)
  - Data are represented as Ndarray and various matrix operations are supported by that
- Using CVX
  - CVX provides a more user-friendly interface to define your optimization problem
  - Before CVX, you need to convert your problem into its canonical form, e,g. a quadratic programming problem

$$\min f(x) = q^T x + \frac{1}{2} x^T Q x$$
$$s.t. Ax = a$$
$$Bx \leq b$$
$$x \geq 0$$

$$\min_{\mathbf{w},b,\{\xi_i\}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_i \xi_i$$
$$s.t. \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \forall i$$
$$\xi_i \geq 0$$

Example

Linear program

A linear program is an optimization problem with a linear objective and affine inequality constraints. A common standard form is the following:

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \leq b. \end{array}$$

Here $A \in \mathcal{R}^{m \times n}$, $b \in \mathcal{R}^m$, and $c \in \mathcal{R}^n$ are problem data and $x \in \mathcal{R}^n$ is the optimization variable. The inequality constraint $Ax \leq b$ is elementwise.

Example

```
# Import packages.
import cvxpy as cp
import numpy as np

# Generate a random non-trivial linear program.
m = 15
n = 10
np.random.seed(1)
s0 = np.random.randn(m)
lamb0 = np.maximum(-s0, 0)
s0 = np.maximum(s0, 0)
x0 = np.random.randn(n)
A = np.random.randn(m, n)
b = A @ x0 + s0
c = -A.T @ lamb0

# Define and solve the CVXPY problem.
x = cp.Variable(n)
prob = cp.Problem(cp.Minimize(c.T@x),
                  [A @ x <= b])
prob.solve()

# Print result.
print("\nThe optimal value is", prob.value)
print("A solution x is")
print(x.value)
print("A dual solution is")
print(prob.constraints[0].dual_value)
```

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \leq b. \end{array}$$

Important skills

- Converting summation to matrix forms

- $$\sum_i a_i x_i = \mathbf{a}^\top \mathbf{x}$$

- $s.t.\ \ \mathbf{w}^\top \mathbf{x}_i \geq 0 \ \ \forall i$  $s.t.\ \ \mathbf{X}\mathbf{w} \geq \mathbf{0}$

$\mathbf{X} \in \mathbf{R}^{N \times d}$ with each row be $\mathbf{x}_i \in \mathbf{R}^d$. $\mathbf{0} \in \mathbf{R}^N$ is an all-zero vector.

Important skills

- $\sum_i \sum_j a_i a_j z_{ij} = \mathbf{a}^\top \mathbf{Z}\mathbf{a}$, where the i,jth element of $\mathbf{Z}$ is z_{ij}
- $\sum_{i=1}^{N} \sum_{j=1}^{N} a_i y_i a_j y_j < \mathbf{x}_i, \mathbf{x}_j >= (\alpha * \mathbf{y})^\top \mathbf{X}^\top \mathbf{X}(\alpha * \mathbf{y})$

  $*$ denotes elemenwise product.

  $\mathbf{X} \in \mathbf{R}^{d \times N}$, where each column of $\mathbf{X}$ is $\mathbf{x}_i$.

- Please refer to the documents and examples of CVX

Summary

- Concept of linear classifier
- Basic idea of linear SVMs: margin maximization
- Primal problem (hard-margin)

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2$$
$$s.t. \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \quad \forall i$$

- Related concepts
  - Support vectors
  - Margin

Summary

- Primal problem (soft-margin)

$$\min_{\mathbf{w},b,\{\xi_i\}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_i \xi_i$$
$$s.t. \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \forall i$$
$$\xi_i \geq 0$$

- Dual problem: concepts and derivation

$$\max_{\alpha_i} \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j < \mathbf{x}_i, \mathbf{x}_j >$$
$$s.t. \quad 0 \leq \alpha_i \leq C$$
$$\sum_i \alpha_i y_i = 0$$

Summary

- Dual problem
  - Relationship with primal problem
    - Support vector
    - Calculating w and b
- Using optimization toolbox
  - CVX
  - Skills for converting summation into matrix operations