Support Vector Machines

CMSC 422

SOHEIL FEIZI

sfeizi@cs.umd.edu

Slides adapted from MARINE CARPUAT

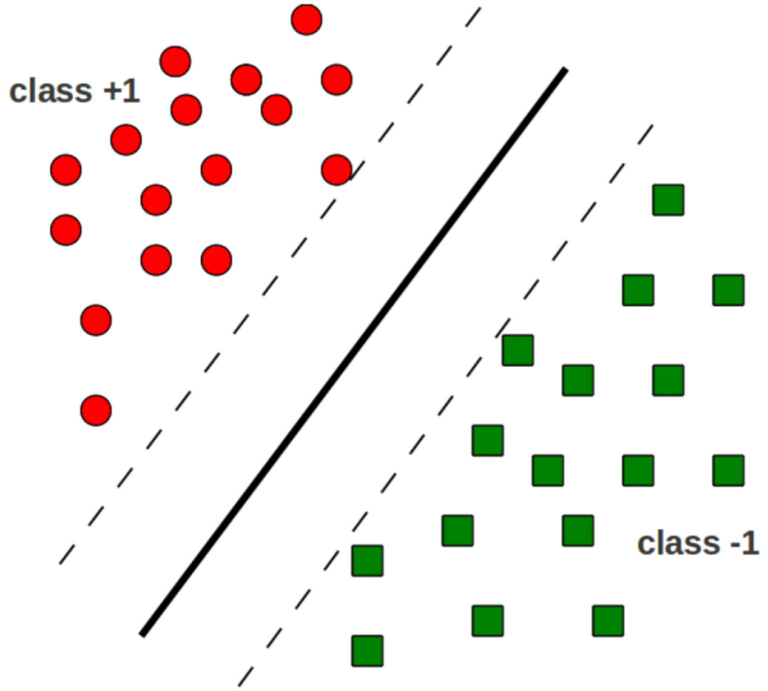
Back to linear classification

- So far: we've seen that kernels can help capture non-linear patterns in data while keeping the advantages of a linear classifier

- Support Vector Machines
  - A hyperplane-based classification algorithm
  - Highly influential
  - Backed by solid theoretical grounding (Vapnik & Cortes, 1995)
  - Easy to kernelize

The Maximum Margin Principle

- Find the hyperplane with maximum separation margin on the training data



Margin of a data set D

$$margin(\mathbf{D}, \boldsymbol{w}, b) = \begin{cases} \min_{(\boldsymbol{x},y)\in\mathbf{D}} y(\boldsymbol{w}\cdot\boldsymbol{x}+b) & \text{if } \boldsymbol{w} \text{ separates } \mathbf{D} \\ -\infty & \text{otherwise} \end{cases} \tag{3.8}$$

Distance between the hyperplane (w,b) and the nearest point in D

$$margin(\mathbf{D}) = \sup_{\boldsymbol{w},b} margin(\mathbf{D}, \boldsymbol{w}, b) \tag{3.9}$$

Largest attainable margin on D

Support Vector Machine (SVM)

A hyperplane based linear classifier defined by $\mathbf{w}$ and b

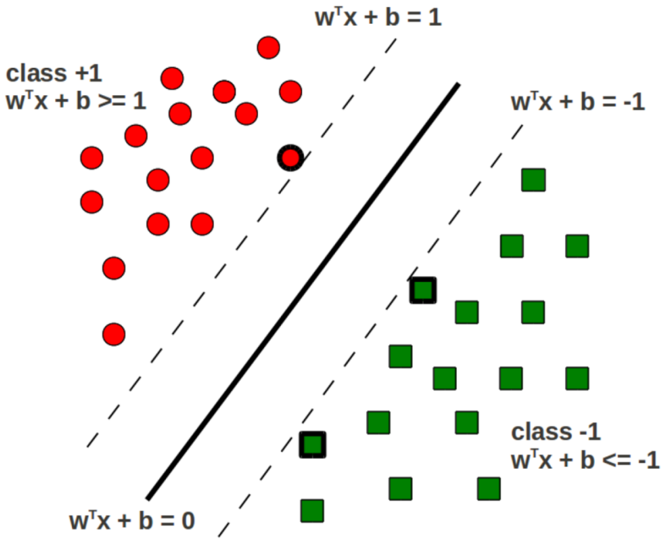
Prediction rule: $y = sign(\mathbf{w}^T\mathbf{x} + b)$

Given: Training data $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$

Goal: Learn $\mathbf{w}$ and b that achieve the maximum margin

Characterizing the margin

Let’s assume the entire training data is correctly classified by (**w**,b) that achieve the maximum margin

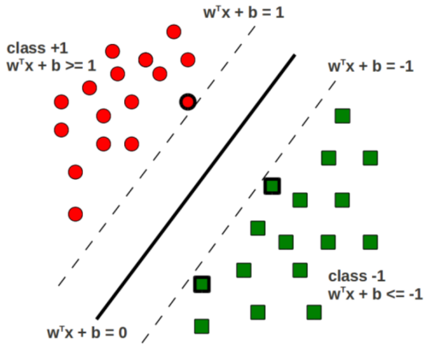


- Assume the hyperplane is such that
  - $\mathbf{w}^T\mathbf{x}_n + b \geq 1$ for $y_n = +1$
  - $\mathbf{w}^T\mathbf{x}_n + b \leq -1$ for $y_n = -1$
  - Equivalently, $y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1$
    $\Rightarrow \min_{1\leq n\leq N} |\mathbf{w}^T\mathbf{x}_n + b| = 1$
  - The hyperplane’s margin:

$$\gamma = \min_{1\leq n\leq N} \frac{|\mathbf{w}^T\mathbf{x}_n+b|}{||\mathbf{w}||} = \frac{1}{||\mathbf{w}||}$$

The Optimization Problem

We want to maximize the margin $\gamma = \frac{1}{||\mathbf{w}||}$



Maximizing the margin $\gamma =$ minimizing $||\mathbf{w}||$ (the norm)

Our optimization problem would be:

$$\boxed{\begin{array}{ll} \text{Minimize} \quad f(\mathbf{w}, b) = \dfrac{||\mathbf{w}||^2}{2} & \\ \text{subject to} \quad y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1, & n = 1, \ldots, N \end{array}}$$

Large Margin = Good Generalization

- Intuitively, large margins mean good generalization
  - Large margin => small ||**w**||
  - small ||**w**|| => regularized/simple solutions
- (Learning theory gives a more formal justification)

Solving the SVM Optimization Problem

Our optimization problem is:

$$\boxed{\begin{array}{ll} \text{Minimize} \;\; f(\mathbf{w}, b) = \dfrac{||\mathbf{w}||^2}{2} & \\ \text{subject to} \;\; 1 \leq y_n(\mathbf{w}^T\mathbf{x}_n + b), & n = 1, \ldots, N \end{array}}$$

Introducing Lagrange Multipliers α_n $(n = \{1, \ldots, N\})$, one for each constraint, leads to the **Lagrangian**:

$$\boxed{\begin{array}{l} \text{Minimize} \;\; L(\mathbf{w}, b, \alpha) = \dfrac{||\mathbf{w}||^2}{2} + \displaystyle\sum_{n=1}^{N} \alpha_n\{1 - y_n(\mathbf{w}^T\mathbf{x}_n + b)\} \\ \text{subject to} \;\; \alpha_n \geq 0; \quad n = 1, \ldots, N \end{array}}$$

Solving the SVM Optimization Problem

Take (partial) derivatives of L_P w.r.t. $\mathbf{w}$, b and set them to zero

$$\frac{\partial L_P}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n, \quad \frac{\partial L_P}{\partial b} = 0 \Rightarrow \sum_{n=1}^{N} \alpha_n y_n = 0$$

Substituting these in the Primal Lagrangian L_P gives the Dual Lagrangian

$$\boxed{\begin{array}{l} \text{Maximize} \quad L_D(\mathbf{w}, b, \alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{m,n=1}^{N} \alpha_m \alpha_n y_m y_n (\mathbf{x}_m^T \mathbf{x}_n) \\ \text{subject to} \quad \sum_{n=1}^{N} \alpha_n y_n = 0, \quad \alpha_n \geq 0; \quad n = 1, \dots, N \end{array}}$$

Solving the SVM Optimization Problem

Take (partial) derivatives of L_P w.r.t. $\mathbf{w}$, b and set them to zero

A Quadratic Program for which many off-the-shelf solvers exist

$$= \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n, \qquad \frac{\partial L_P}{\partial b} = 0 \Rightarrow \sum_{n=1}^{N} \alpha_n y_n = 0$$

Substituting thes the Primal Lagrangian L_P gives the Dual Lagrangian

Maximize $\quad L_D(\mathbf{w}, b, \alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{m,n=1}^{N} \alpha_m \alpha_n y_m y_n (\mathbf{x}_m^T \mathbf{x}_n)$

subject to $\quad \sum_{n=1}^{N} \alpha_n y_n = 0, \quad \alpha_n \geq 0; \quad n = 1, \ldots, N$

SVM: the solution!

Once we have the α_n's, $\mathbf{w}$ and b can be computed as:

$$\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n$$

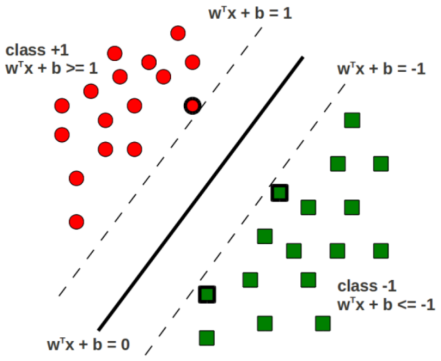
$$b = -\frac{1}{2}\left(\min_{n:y_n=+1} \mathbf{w}^T\mathbf{x}_n + \max_{n:y_n=-1} \mathbf{w}^T\mathbf{x}_n\right)$$

Note: Most α_n's in the solution are zero (sparse solution)

- Reason: Karush-Kuhn-Tucker (KKT) conditions
- For the optimal α_n's

$$\alpha_n\{1 - y_n(\mathbf{w}^T\mathbf{x}_n + b)\} = 0$$

- α_n is non-zero only if $\mathbf{x}_n$ lies on one of the two margin boundaries, i.e., for which $y_n(\mathbf{w}^T\mathbf{x}_n + b) = 1$
- These examples are called support vectors
- Support vectors "support" the margin boundaries

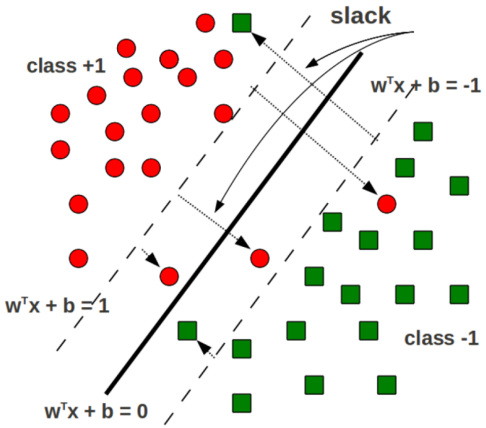


SVM in the non-separable case

- no hyperplane can separate the classes perfectly

- We still want to find the max margin hyperplane, but
  - We will allow some training examples to be **misclassified**
  - We will allow some training examples to fall **within** the margin region

SVM in the non-separable case



Recall: For the separable case (training loss $= 0$), the constraints were:

$$y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 \quad \forall n$$

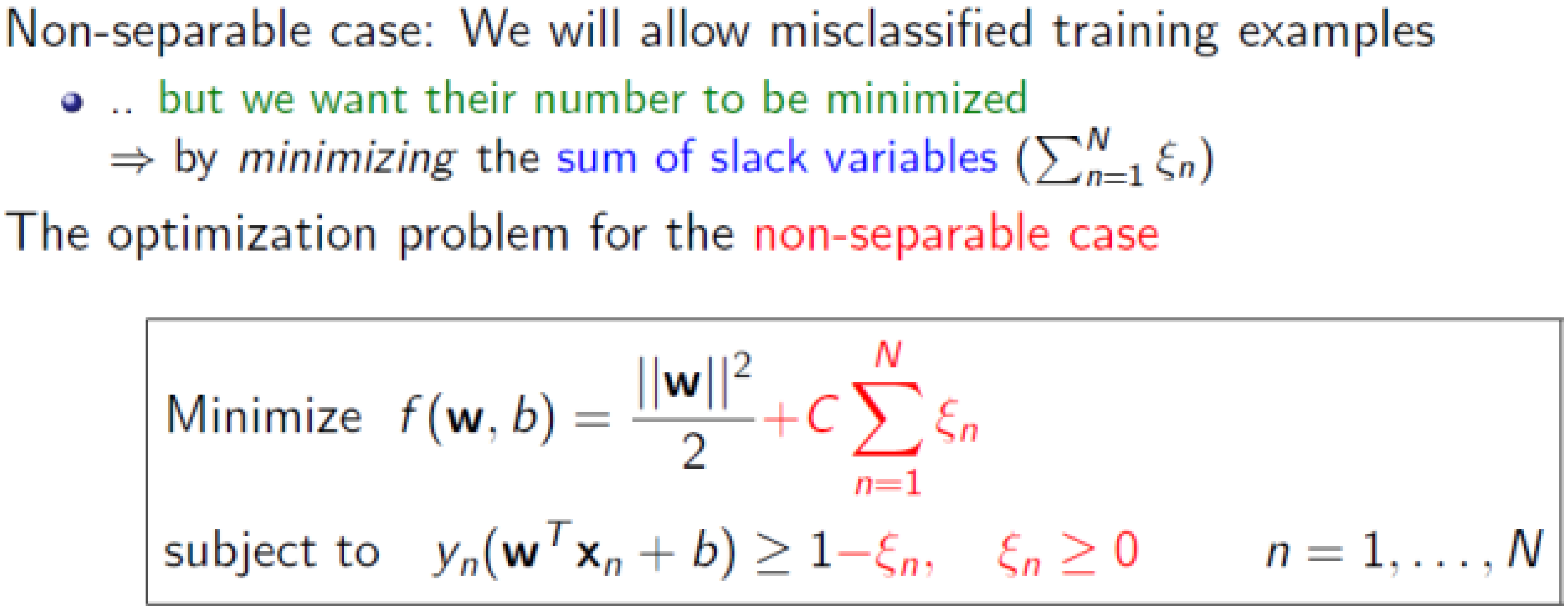
For the non-separable case, we relax the above constraints as:

$$\boxed{y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 - \xi_n \quad \forall n}$$

ξ_n is called slack variable (distance $\mathbf{x}_n$ goes past the margin boundary)

$\xi_n \geq 0, \forall n$, misclassification when $\xi_n > 1$

SVM Optimization Problem



Non-separable case: We will allow misclassified training examples

- .. but we want their number to be minimized

  $\Rightarrow$ by *minimizing* the sum of slack variables ($\sum_{n=1}^{N} \xi_n$)

The optimization problem for the non-separable case

$$\text{Minimize} \quad f(\mathbf{w}, b) = \frac{||\mathbf{w}||^2}{2} + C \sum_{n=1}^{N} \xi_n$$

$$\text{subject to} \quad y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1 - \xi_n, \quad \xi_n \geq 0 \qquad n = 1, \ldots, N$$

C hyperparameter dictates which term dominates the minimization

- Small C => prefer large margins and allows more misclassified examples
- Large C => prefer small number of misclassified examples, but at the expense of a small margin

Soft SVM

- Same optimization as :

$$\min_{\mathbf{w},b} \frac{\|\mathbf{w}\|^2}{2} + C \underbrace{\sum_{n=1}^{N} \max\left\{1 - y_n(\mathbf{w}^t\mathbf{x}_n), 0\right\}}_{\text{Hinge loss!}}$$

- Why?

- Have you seen this loss function before?