Machine Learning (Lab support)

Support vector machine (SVM)

Abdelkrime Aries
Laboratoire de la Communication dans les Systèmes Informatiques (LCSI)
École nationale Supérieure d'Informatique (ESI, ex. INI), Algiers, Algeria

Academic year: 2024-2025

Machine Learning (Lab support)

SVM: Introduction

- We have already seen ...
  - logistic regression looks for a linear separation between two classes
  - it draws a hyperplane between them
  - there could be many possible hyperplanes
- But ...
  - how about drawing a hyperplane which has the same distance from the two classes

Machine Learning (Lab support)

SVM: Plan

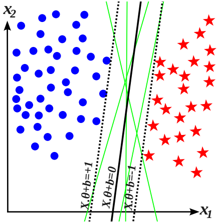
1. **Problem definition**
   - Hard-margin
   - Soft-margin
2. **Primal form**
   - Cost function
   - Class estimation
   - Optimization algorithms
3. **Dual form**
   - Cost function
   - Class estimation
   - Optimization algorithms

Section 1

Problem definition

SVM

Problem definition



- in logistic regression, the hyperplane's equation is $\sigma(z = b + \sum_{j=1}^{N} w_j x_j) = 0.5$
- we want it to be $z = b + \sum_{j=1}^{N} w_j x_j = 0$
- in this case, $y \in \{-1, +1\}$
- thus, $z^{(i)} \geq 1 \Rightarrow \hat{y}^{(i)} = 1$ and $z^{(i)} \leq -1 \Rightarrow \hat{y}^{(i)} = -1$
- the space between -1 and $+1$ is a margin which equals $\frac{2}{||w||}$
- the idea is to maximize this margin, thus minimizing $\frac{||w||^2}{2}$

SVM: Problem definition

Hard-margin

- $z^{(i)} \geq 1 \Rightarrow \hat{y}^{(i)} = 1$ and $z^{(i)} \leq -1 \Rightarrow \hat{y}^{(i)} = -1$
- So, $\hat{y}^{(i)} = sign(z^{(i)})$ where $z^{(i)} \in]-\infty, +\infty[$
- we want, $y^{(i)} = \hat{y}^{(i)}$ where $y^{(i)} \in \{-1, +1\}$
- in this case, $y^{(i)}\hat{y}^{(i)} = 1$, thus $y^{(i)}z^{(i)} \geq 1$
- the optimization problem will be formulated as:

$$\min_{w} \frac{\|w\|^2}{2}$$

$$\text{subject to } y^{(i)}z^{(i)} \geq 1,\ \forall i \in 1 \cdots M$$

- in this case, no sample must be inside the margin; even the ones in the correct side

SVM: Problem definition

Soft-margin

- when $y^{(i)}z^{(i)} < 1$ there are two possible interpretations:
  - $y^{(i)}z^{(i)} < 0$ the sample is on the wrong side of the decision vector
  - $y^{(i)}z^{(i)} \geq 0$ it is on the correct side, but it is inside the margin
- to allow this second case, ***hinge loss*** is used:

$$\zeta^{(i)} = \max(0, 1 - y^{(i)}z^{(i)})$$

- then, the goal will be to minimize this loss function:

$$\frac{||w||^2}{2} + C \sum_{i=1}^{M} \zeta^{(i)}$$

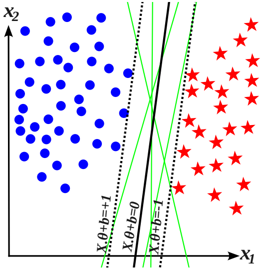
- C is trad-off between increasing the margin size and having samples on the correct side

Section 2

Primal form

SVM
Primal form

- Draw a decision line between two classes y
- Maximize the margin between the two classes.
- Use linear combination over features x like LR.
- Based on the coordinates x, a sample belongs to a given class if it is located on its side



SVM: Primal form

Cost function

- The loss function can be formulated as

$$J_w = \frac{\|w\|^2}{2} + C \sum_{i=1}^{M} \zeta^{(i)} \text{ where } \zeta^{(i)} = \max(0, 1 - y^{(i)} z^{(i)})$$

- when C is ...
  - big, having samples on the correct side is preferred over having a big margin
  - small, having having a big margin is preferred over samples on the correct side

$$\frac{\partial J_w}{\partial w_j} = w_j + C \sum_{i=1}^{M} \frac{\partial \zeta^{(i)}}{\partial w_j} \text{ where } \frac{\partial \zeta^{(i)}}{\partial w_j} = \begin{cases} 0 & \text{if } y^{(i)} z^{(i)} \geq 1 \\ -x_j^{(i)} y^{(i)} & \text{otherwise} \end{cases}$$

SVM: Primal form

Class estimation

- Once the model trained, $z^{(i)} = 0$ will be the decision hyperplane
- we define the sign function as $sign(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{otherwise} \end{cases}$
- $\hat{y} = sign(z)$
- $z = b + \sum_{j=1}^{N} w_j x_j$

SVM: Primal form

Optimization algorithms

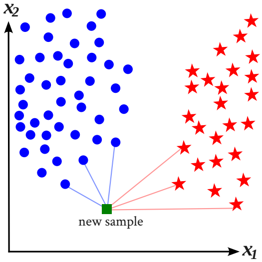
- Since the loss function is differentiable, ***Gradient descent*** can be used

Section 3

Dual form

SVM
Dual form

- Use similarity of the new sample with training samples.
  - If it is more similar to positive samples, then its class is positive
  - If it is more similar to negative samples, then its class is negative
- The similarities does not have the same weight
  - If the new sample is similar to a training sample which is far from the other class, then its class is more likely to be similar
  - If the new sample is similar to a training sample which is near the other class, then its class is less likely to be similar



SVM: Dual form

Cost function(1)

- By deconstructing the hinge loss, the optimization problem will be :

$$\min_{w,\zeta^{(i)}} \frac{||w||^2}{2} + C\sum_{i=1}^{M}\zeta^{(i)}$$

subject to $y^{(i)}z^{(i)} \geq 1 - \zeta^{(i)},\ \zeta^{(i)} \geq 0,\ \forall i \in 1\cdots M$

- where $\zeta^{(i)} = \max(0, 1 - y^{(i)}z^{(i)})$

SVM: Dual form

Cost function(2)

- By solving for the Lagrangian dual:

$$\max_{\lambda_i} \sum_{i=1}^{M} \lambda_i - \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} \lambda_i \lambda_j y^{(i)} y^{(j)} x^{(i)} x^{(j)}$$

$$\text{subject to } \sum_{i=1}^{M} \lambda_i y^{(i)} = 0,\ 0 \leq \lambda_i \leq C,\ \forall i \in 1 \cdots M$$

- $x^{(i)} x^{(j)}$ can be seen as a similarity measure called dot product
- We can use other similarity measures $K(x^{(i)}, x^{(j)})$
- $K(x^{(i)}, x^{(j)})$ is called ***kernel***

SVM: Dual form

Class estimation

$$\hat{y}_t = sign(b + \sum_{i=1}^{M} \lambda_i y^{(i)} K(x^{(i)}, x_t))$$

- $\hat{y}_t$ is the estimated class of the given test sample x_t
- $x^{(i)}$ are training samples
- $K(a, b)$ is a kernel function
  - Linear kernel $K(A, B) = A \cdot B^T$
  - RBF kernel $K(A, B) = \exp(-\frac{||A-B||^2}{2\sigma})$

SVM: Dual form

Optimization algorithms

- The problem can be solved using Quadratic programming
- One optimization algorithm is ***Sequential minimal optimization* [Platt, 1998]**

Section 4

Bibliography

Bibliography

Platt, J. (1998).
Sequential minimal optimization: A fast algorithm for training support vector machines.
Technical Report MSR-TR-98-14, Microsoft.

The margin has been reached
stop scrolling