# Support Vector Machine

Bingyu Wang

August 26, 2014

## 1 What's SVM

Give general idea about SVM and introduce the goal of this notes, what kind of problems and knowledge will be covered by this node.

Define one single SVM model for two labels classification, which label $y \in \{-1, 1\}$ , the classifier will use parameters $w, b$, and write the classifier as

$$h_{w,b}(x) = g(w^T x + b)$$

## 2 Margins

### 2.1 Why Margins

Why we choose margins to start our SVM topic? This section will give the intuitions about margins and about the "confidence" of our predictions.

### 2.2 Functional and Geometric Margins

Define the functional margins, which can represent a confident and a correct prediction. The larger functional margins, the classifier better. However, by scaling $w, b$, we can make the functional margin arbitrarily large without really changing anything meaningful.

Then introduce the geometric margins and give it definition as:

$$\gamma = \min_{i=1,\dots,m} \gamma^{(i)} \quad \text{where} \quad \gamma^{(i)} = y^{(i)}((\frac{w}{||w||})^T x^{(i)} + \frac{b}{||w||})$$

## 3 The Optimal Margin Classifier

Optimization problem:

$$\begin{aligned} &\max_{\gamma,w,b} \gamma \\ &\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \gamma, \quad i = 1, \dots, m \\ &\qquad ||w|| = 1 \end{aligned}$$

Simplify this problem to:

$$\min_{\gamma,w,b} \frac{1}{2}||w||^2 \quad \text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, m \tag{1}$$

We can use Quadratic Programming(QP) to solve this problem. But we will try to use Lagrange duality to solve this problem, which may allow us use kernels and is also more efficient.

# 4 Lagrange Duality

## 4.1 Lagrange

Lagrange multipliers to solve the problem of the following form:

$$\begin{aligned} &\min_w f(w) \\ &\text{s.t.} \quad h_i(w) = 0, \quad i = 1, \ldots, l \end{aligned}$$

## 4.2 Primal Problem

## 4.3 Dual Problem

## 4.4 Karush-Kuhn-Tucker

Once the Primal problem and Dual problem equal to each other, the parameters will meet the KKT conditions. We just introduce the five conditions as following:

$$\frac{\partial}{\partial w_i}\mathcal{L}(w^*, \alpha^*, \beta^*) = 0, \quad i = 1, \ldots, n \tag{2}$$

$$\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*, \alpha^*, \beta^*) = 0, \quad i = 1, \ldots, l \tag{3}$$

$$\alpha_i^* g_i(w^*) = 0, \quad i = 1, \ldots, k \tag{4}$$

$$g_i(w^*) \leq 0, \quad i = 1, \ldots, k \tag{5}$$

$$\alpha^* \geq 0, \quad i = 1, \ldots, k \tag{6}$$

# 5 Optimal Margin Classifiers

Go back to our problem defined in (1), which is primal problem. Then use the above knowledge to derive the problem and then get the dual optimization problem**(I will introduce the derivation processing details here)**.

$$\begin{aligned} &\max_\alpha W(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2}\sum_{i,j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j < x^{(i)}, x^{(j)} > . \\ &\text{s.t.} \quad \alpha_i \geq 0, \quad , i = 1, \ldots, m \\ &\qquad \sum_{i=1}^m \alpha_i y^{(i)} = 0. \end{aligned}$$

Now we can easily to solve this problem. But we have two new problems:
1) What if the data cannot be separated in low-dimensions? We need **Kernel**, which maps features to higher-dimensions to be easier separated.
2) What if there existing some noise data points between the support vector? We need Regularization, re-define the problem (1) by using $l_1$ **regularization**.

# 6 Kernels

This section will introduce the Kernels. Why we need kernels, and how kernel can solve the non-separable data.

# 7 Regularization

This section will add $l_1$ regularization to our original model, and to do better about noise and non-separable data. (**Here the derivation processing is similar to previous derivation steps. So here I won't give the details about the derivation. But I can post the details online and giving them a link, if they are interested in the derivation.**)

# 8 SMO Algorithm

After getting the final dual optimization problem, I will introduce the SMO algorithms to solve this problem. And give the derive processing and algorithms.