SVM and Kernel Method Review Note

Lizi Chen

1 Introduction

Support Vector Machine is a supervised learning algorithm for data classification. The 'Support Vector' refers to the group of vectors that separate the data with largest 'distance'.

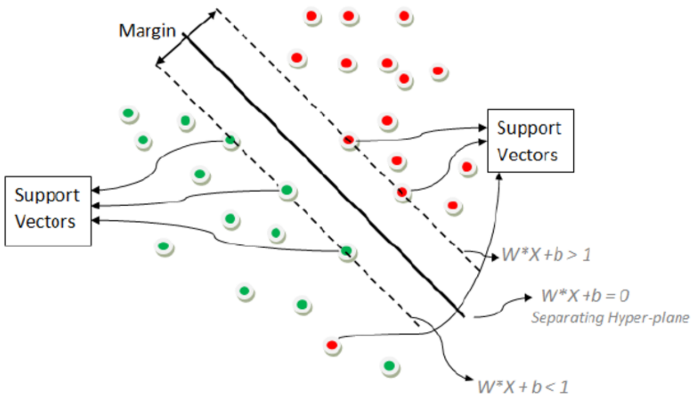

Figure 1: Example of Margin and Support Vectors

The target is to train a (binary, in the picture above) classifier that has the largest margin in between two classes of data. The margin is represented by the separating hyper-plane and the distance from the hyper-plane to support vector.

Mathematically,

$$\text{hyper-plane:}\quad w^T x + b = 0$$

$$\text{distance:}\quad d = 2 * \frac{|w^T x + b|}{||w||}$$

We want to find a pair of w^* and b^* that maximize d.

2 Convert to Constraint Optimization Question

Set the true class label of data Y be either 1 or -1. That is

$$y \in \{-1, 1\}$$

and let the predicted score be y^* or $f(x)$. Thus Margin is defined as

$$m = y \cdot f(x)$$

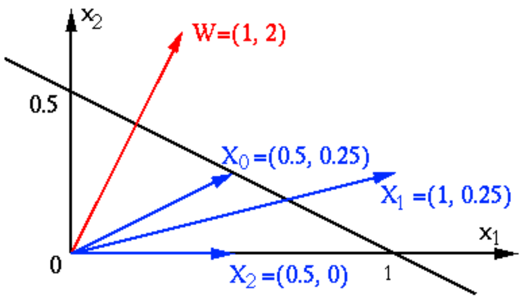

Figure 2: SVM 2D Example

2.1 Example:

The straight line in 2D space $\mathbf{x} = [x_1, x_2]^T$ described by the following equation:

$$f(x) = x^T w + b = [x_1, x_2] \begin{bmatrix} w_1 \\ 2_2 \end{bmatrix} + b = [x_1, x_2] \begin{bmatrix} 1 \\ 2 \end{bmatrix} - 1 = x_1 + 2x_2 - 1 = 0$$

Distance between the origin and the line $f(x)$ is:

$$\frac{|b|}{||w||} = \frac{1}{\sqrt{w_1^2 + w_2^2}} = \frac{1}{\sqrt{5}} = 0.447$$

For the three data points substitute their $x_1 - axis$ and $x_2 - axis$ values into $f(x)$:

$$f(x_0) = 0, \text{ , hence } x_0 \text{ is on the plane/line}$$

$$f(x_1) > 0, \text{ , hence } x_1 \text{ is above the straight line}$$

$$f(x_2) < 0, \text{ , hence } x_2 \text{ is below the straight line.}$$

Note: The following example does not include b: todo

Another example, when $y_0 = 1$ and the prediction $f(x_0) = 1$, the data point x_0 is a support vector. Similarly, when $y_1 = -1$ and $f(x_1) = -1$, $m_{x_1} = 1$, the data point x_1 is also a support vector but on the other class side.

When $m > 0$, data point is correctly classified. When $m < 0$; that's when $y > 0, f(x) < 0$ or $y < 0, f(x) > 0$, data point is incorrectly classified.

2.2 Loss Function

Therefore, **Loss Function** can be written as:

$$\text{loss}_{\text{Hinge Loss}} = max\{1 - m, 0\} \tag{1}$$

As said before; when $f(x_i) = 1$, data point x_i is a support vector. In the picture Figure. 1, x_i will be on the dotted line. Thus, we can re-write distance:

$$\text{distance:} \quad d = 2 * \frac{|w^T x_i + b|}{||w||} = 2 * \frac{1}{||w||}$$

Therefore, in order to find the maximum margin, we need to find the smallest $||w||$. Equivalently:

$$\text{minimize:} \quad \frac{1}{2}||w||^2$$

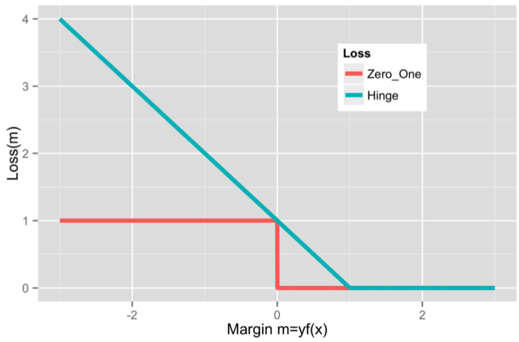

Figure 3: Hinge Loss for SVM

Also, with the constraint to minimize the Hinge Loss in Eq.1.

Formally, in Tikhonov style, the SVM prediction function is the solution to:

$$\text{minimize}_{w,b}\ \frac{1}{2}||w||^2 + c \cdot \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y_i[w^T x_i + b])$$

, where c is the regularization parameter, that usually put on the empirical risk part; $\frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y_i[w^T x_i + b])$, rather than the penalty part; $\text{minimize}_{w,b}\ \frac{1}{2}||w||^2$.

Equivalently, in Constraint form:

$$\text{minimize}\ \frac{1}{2}||w||^2 + c \cdot \frac{1}{n}\sum_{i=1}^{n} \xi_i$$

$$\text{s.t. } \xi_i \geq \max(0, 1 - y_i[w^T x_i + b])$$

3 Lagrangian Duality

For any primal form optimization problem; just like the previous inequality constrained optimization problem, there is a recipe for constructing a corresponding Lagrangian dual problem. Here we introduce the Lagrange multipliers and dual variables to convert initial constraint problem to a concave maximization problem.

reference: David Rosenberg 04d, 04b,

Pre-requisites:
来源：http://www.cnblogs.com/LeftNotEasy/archive/2011/05/02/basic-of-svm.html:

转化为对偶问题，并优化求解: 这个优化问题可以用拉格朗日乘子法去解，使用了 KKT 条件的理论，这里直接作出这个式子的拉格朗日目标函数：

$$L(w,b,a) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} a_i(y_i(w^T x_i + b) - 1) \tag{2}$$

求解这个式子的过程需要拉格朗日对偶性的相关知识（另外 pluskid 也有一篇文章专门讲这个问题），并且有一定的公式推导，如果不感兴趣，可以直接跳到后面用蓝色公式表示的结论，该部分推导主要参考自 plukids 的文章。

4 Kernel Function

David Rosenberg week 5a

Radial Basis Function

Gaussian Kernel

5 SVM vs. Logistic Regression

SVM and Logistic Regression work comparable in practise. But there are some tricks to remember. Let n be the number of features, m be the number of training examples.

- If n is large relative to m → use Logistic Regression, or SVM without kernel.
- If n is small, m is intermediate → use SVM with Gaussian kernel.
- If n is small, m is large, create/add more features → then use Logisic Regression or SVM without a kernel.

SVM and Logistic Regression also differ in Loss Function. SVM minimizes hinge loss while Logistic Regression minimizes logistic loss:

$$J_{svm}(w,b) = \frac{1}{2}||w||^2 + c \cdot \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y_i[w^T x_i + b]) \tag{3}$$

$$J_{LR}(\theta) = ||\theta||^2 - \frac{1}{m}\sum_{i=1}^{m}\Big[-\log(1 + e^{\theta x^i})\Big] \tag{4}$$

Also, for loss functions:

- Logistic loss diverges faster hinge loss. So, in general, it will be more sensitive to outliers.
- Logistic loss does not go to zero even if the point is classified sufficiently confidently. This might lead to minor degradation in accuracy.

6 References

First Course in machine learning

Machine Learning, coursera Andrew Ng

Support Vector Machine

Introduction to Statistical Learning, application with R

Foundation of Machine Learning

Chapter: Support vector machines and machine learning on documents
Introduction to Information Retrieval https://nlp.stanford.edu/IR-book/html/htmledition/support-vector-machines-and-machine-learning-on-documents-1.html