Support Vector Machines

Fangfang Song

1 Introduction

Informally, given a training set(assume it's linearly separable), we manage to find a decision boundary that allows us to separate the positive examples and negative examples. A good choice will be put the decision boundary at the middle of the gap/street between the positive and negative examples, and we want to maximize the width of the street. To make a prediction on a example$x^{(i)}$, assume $\vec{w}$ is the normal vector of the decision boundary, we can project $\vec{x^{(i)}}$ on to $\vec{w}$, which is $\vec{x^{(i)}} \cdot \vec{w}$, if pass the decision boundary, it will be positive, else, it will be negative. Mathematically:

if $\vec{x^{(i)}} \cdot \vec{w} > c$ or$\vec{x^{(i)}} \cdot \vec{w} + b > 0$, positive example
if $\vec{x^{(i)}} \cdot \vec{w} < c$ or $\vec{x^{(i)}} \cdot \vec{w} + b > 0$, negative example
Furthermore, we will add some constrain and let the street has a width, we will say,
if $\vec{x^{(i)}} \cdot \vec{w} + b > 1$, positive example
if $\vec{x^{(i)}} \cdot \vec{w} + b < -1$, negative example

2 Derivation of SVM

For mathematical convenience, we introduce a function function$y^{(i)}(\vec{x^{(i)}} \cdot \vec{w} + b)$, where $y \in \{1, -1\}$, so the decision rule becomes : $y^{(i)}(\vec{x^{(i)}} \cdot \vec{w} + b) \geq 1$ And for the examples on the gutter:

$$\mathrm{y}^{(i)}(\vec{x^{(i)}} \cdot \vec{w} + b) = 1 \ (1)$$

Then lets find the width of the street:

width = (x$^+ - x^-) \cdot \frac{\vec{w}}{\|w\|}$ (2)

Combing with equation (1): $width = \frac{2}{\|w\|}$ we want to maximize the the width, which is equivalent to minimize $\|w\|$, which is equivalent to minimize$\frac{1}{2}\|w|^2$ So the optimization problem are:

$$\begin{array}{cc} min_{\gamma,w,b} & \frac{1}{2}||w||^2 \\ s.t. & y^{(i)}(w^T x^{(i)} + b) \geq 1, i = 1, ...m \end{array}$$

More concretely ,we introduce the concept of functional margin and geometric margin. Given a training example$(x^{(i)}, y^{(i)})$, we define the functional margin of the decision boundary (w, b) with respect to the training example

$$\hat{\gamma}^{(i)} = y^{(i)}(w^T x(i) + b) \tag{3}$$

The function margin of (w,b) with respect to S is defined to be the smallest of the functional margin of the individual training examples.

Functional margin can be thought as a testing function that will tell you whether a particular point is properly classified or not: the result would be positive for properly classified points. To maximize the margin you need more that just the sign, you need to have a notion of magnitude, the functional margin would give you a number but without a reference you can't tell if the point is actually far away or close to the decision plane. **Also, by scaling w and b, we can make the functional margin arbitrarily large without really changing anything meaningfully**. Thus we introduce the idea of **geometric margin**. The geometric margin is telling you not only if the point is properly classified(geometric margin has an sign) or not, but the magnitude of that distance in term of units of $||w||$. The geometric margin is the distance from the training example to the separating hyper plane, it is the functional margin scaled by $\|w\|$. **And it is robust to the scaling of w and b.** Mathematically,the geometric margin of a separating hyper-plane/decision boundary with respect to a training example$(x^{(i)}, y^{(i)})$ to be

$$\gamma^{(i)} = y^{(i)}((\frac{w}{||w||}^T x^{(i)} + \frac{b}{||w||})$$

$$\text{geometric margin} = \frac{\text{functional margin}}{||w||}$$

We want to find a separating hyper-plane which achieves the maximum geometric margin. Thus we pose the following optimization problem:

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

We can transform the problem into a nicer one by getting rid of the nasty "$||w|| = 1$" constraint:

$$\begin{array}{cc} max_{\hat{\gamma},w,b} & \hat{\gamma} \\ s.t. & y^{(i)}(w^T x^{(i)} + b \geq \hat{\gamma}, i = 1, ...m) \end{array}$$

Since geometric margin are functional margin measured in units of $\vec{w}$, also we can add an arbitrary scaling constraint on w and b without changing anything, we introduce the scaling constraint that the functional margin of w,b with respect to the training set must be 1:

$$\hat{\gamma} = 1$$

Now, the optimization problem becomes:

$$\begin{array}{cc} min_{\gamma,w,b} & \frac{1}{2}\|w\|^2 \\ s.t. & y^{(i)}(w^T x^{(i)} + b) \geq 1, i = 1, ...m \end{array}$$

2.1 Lagrange duality

Consider the following **primal** optimization problem:

$$\begin{array}{cc} min_w & f(w) \\ s.t. & g_i(w) \leq 0, i = 1, ...k \\ & h_i(w) = 0, i = 1, ...l \end{array}$$

Define the **Generalized Lagrangian**

$$L(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

The **primal** quantity:

$$\theta_p(w) = max_{\alpha,\beta:\alpha_i \geq 0} L(w, \alpha, \beta)$$

If w violates any of the primal constraints, then we can verify that $\theta_p(w) = \infty$
Hence,

$$\theta_p(w) = \begin{cases} f(w) & \text{if w satisfies primal constraints} \\ \infty & \text{otherwise} \end{cases}$$

Define the minimization problem

$$min_w \theta_p(w) = min_w max_{\alpha,\beta:\alpha_i \geq 0} L(w, \alpha, \beta)$$

And, this is the same problem as our original primal problem.

Now, lets consider the **dual optimization problem**, We define

$$\theta_D(\alpha, \beta) = min_w L(w, \alpha, \beta)$$

The **Dual optimization problem** is $max_\alpha min_{w,b} L(w, b, \alpha)$

$$max_{\alpha,\beta:alpha_i \geq 0} \theta_D(\alpha, \beta) = max_{\alpha,\beta:alpha_i \geq 0} min_\omega L(\omega, \alpha, \beta)$$

$$min_w max_{\alpha,\beta:\alpha_i \geq 0} L(w,\alpha,\beta) \geq max_{\alpha,\beta:alpha_i \geq 0} min_\omega L(\omega,\alpha,\beta)$$

Under certain conditions, we will have the value of the primal problem = the value of the dual problem. The conditions are:

1. f and g_i are convex(Hessian ≥ 0)h$_i$ are affine($h_i(w) = \alpha_i^T w + b$)

3. g_i are strictly feasible; this means that there exists some w so that $g_i(w) < 0$ for all i.

Under these conditions, there must exist w^*, α^*, β^* so that w^* is the solution to the primal problem, α^*, β^* are the solution to the dual problem, and w^*, α^*, β^* satisfy the KKT conditions:

$$\begin{array}{rcl} \frac{\partial}{\partial w_i} L(w^*,\alpha^*,\beta^*) & = & 0, \quad i = 1, ...n \\ \frac{\partial}{\partial \beta_i} L(w^*,\alpha^*,\beta^*) & = & 0, \quad i = 1, 2...l \\ \alpha^* g_i(w) & = & 0, \quad i = 1, ...k \\ g_i(w) & \leq & 0, \quad i = 1, ...k \\ \alpha^* & \geq & 0, \quad i = 1, ...k \end{array}$$

3 Optimal margin classfier

To find the optimal margin classifier, the primal optimization problem is:

$$\begin{array}{ll} min_{\gamma,w,b} & \frac{1}{2}||w||^2 \\ s.t. & y^{(i)}(w^T x^{(i)} + b \geq 1, i = 1, 2, ...m) \end{array}$$

Construct the Lagrangian for our optimization problem:

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i [y^{(i)}(w^T x^{(i)} + b) - 1] \tag{4}$$

The **dual** form of this problem is $max_\alpha min_{w,b} L(w,b,\alpha)$.

To minimize $L(w,b,\alpha)$, we set the derivatives of L with respect to w and b to zero. We have :

$$\nabla_w L(w,b,\alpha) = w - \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} = 0$$

This implies:

$$w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \tag{5}$$

$$\nabla_b L(w,b,\alpha) = \sum_{i=1}^{m} \alpha_i y^{(i)} = 0 \tag{6}$$

Now ,lets's take equation(5) and plug that back into the Lagrangian:

$$L(w,b,\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)=0}$$

Thus, the dual optimization problem is:

$$\begin{array}{ll} max_\alpha & W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)=0} \\ s.t. & \alpha_i \geq 0, i = 1, ...m \\ & \sum\limits_{i=1}^{m} \alpha_i y^{(i)} \geq 0 \end{array}$$

By using the KKT conditions, we find that α_i will all be zero except the **support vectors**(the points with the smallest margins)

To find α, we use **SMO** algorithm, once we have α, we know that $w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$. To get b, we consider the primal problems, and get $b = -\frac{max_{i:y^i=-1} w^{*} T x^{(i)} + min_{i:y^i=1} w^{*} T x^{(i)}}{2}$.Intuitively, what this equation does is to find the worst positive and negative examples, put the hyper-plane in the middle.

To make a prediction, we will calculate $w^T x + b$ and predict $y = 1$ if and only if this quantity is bigger than zero.

$$w^T x + b = (\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)})^T x + b = \sum_{i=1}^{m} \alpha_i y^{(i)} < x^{(i)} \cdot x > +b$$

We only need to find the inner product between x and the support vectors.

4 SMO algorithm

Repeat till convergence

1. Select some pair of to update next α_i and α_j
2. Optimize $W(\alpha)$ wiht respect to α_i and α_j, while holding all the other α_k fixed.

5 Kernel

When we have a training set that is not linearly separable, we may want to change perspective and mapping the data to a high dimensional space, which will increase the likelihood that the data is separable and we can use the SVM algorithm. Thus we define the corresponding **Kernel** to be

$$K(x,z) =< \phi(x) \cdot \phi(z) >$$

The magic thing about kernel is that we can find the Kernel without knowing the exact mapping function.

One of the popular Kernel are $K(x,z) = (x^T z + b)^d$, and $K(x,z) = exp(-\frac{\|x-z\|^2}{2\sigma^2})$

6 Regulation and the non-separable case

To make the algorithm work for non-linearly separable datasets as well as be less sensitive to outliers, we allow examples to have (functional) margin less than 1, and if an example whole functional margin is $1 - \zeta_i$, we would pay a cost of the objective function being increased by $C\zeta_i$. The parameter C controls the relative weighting between the twin goal of make the $\|w\|^2$ large and of ensuring the most examples have functional margin at least 1.

we reformulate our optimization by using l_1 regulation:

$$\begin{array}{cc} min_{\zeta,w,b} & \frac{1}{2}\|w|^2 + C\sum_{i=1}^m \zeta_i \\ s.t. & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \zeta_i, i = 1, ...m \\ & \zeta_i \geq 0, i = 1, ...m \end{array}$$

We form the Lagrangian:

$$L(w,b,\zeta,\alpha,\gamma) = \frac{1}{2}w^T w + C\sum_{i=1}^m \zeta_i - \sum_{i=1}^m \alpha_i(y^{(i)}(w^T x + b) - 1 + \zeta_i) - \sum_{i=1}^m \gamma_i \zeta_i$$

To get the dual form of the optimization problem, we set the derivative with respect to w and b to zero, substituting them back in , and simplifying, we obtain

$$\begin{array}{cc} max_\alpha & W(\alpha) = \sum\limits_{i=1}^m \alpha_i - \frac{1}{2}\sum\limits_{i,j=1}^m y^{(i)}y^{(j)}\alpha_i\alpha_j < x^{(i)}, x^{<j>} > \\ s.t. & 0 \leq \alpha_i \leq C, i = 1, ...m \\ & \sum\limits_{i=1}^m \alpha_i y^{(i)} = 0 \end{array}$$