Introduction:

Constructing optimally seperating hyperplane between two classes for seperation. We review the mathematical basis of SVM and how to construct it. We will also discuss the application of SVM in the context of classification. We will also discuss the application of SVM in the context of regression and generalizing to the nonseperable case where the classes may not be seperable by a linear boundary.

Explanation:

Our training data consists of two classes of points of N pairs of $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$ with $x_i \in \mathbb{R}$ and $y_i \in \{-1, 1\}$. Defined a hyperplane by $\{x : f(x) = x^T + \beta + \beta_0 = 0\}$. Where β is an unit vector and β_1 is teh intercept. A classification rule is induced by $f(x)$ is

$$G(x) = sign(x^T\beta + \beta_0)$$

Now we can find $x^T\beta + \beta_0$ and $y_i f(x_i) > 0, \forall i$. Let M denote $\frac{1}{\|\beta\|}$ as the margin between the two exteme lines of $G(x)$ The Optimization Problem is:

$$\text{maximize}_\beta M,$$

$$\text{Subject to } y_i f(x_i) \geq M, \forall i \in \{1, 2, ..., N\}$$

Linear Support Vector Machines:

The three hyper planes we get are:

$$x : f(x) = x^T\beta + \beta_0 = -1$$

$$x : f(x) = x^T\beta + \beta_0 = 0$$

$$x : f(x) = x^T\beta + \beta_0 = 1$$

The distance between the hyperplanes is(under the assumption that $\|\beta\| \neq 1$): Let $\beta = w$ and $\beta_0 = b$,we get the distance as :

$$D = \frac{1 - b}{\|w\|} - \frac{-1 - b}{\|w\|}$$

$$D = \frac{2}{\|w\|}$$

where D is twice the margin. So margin can be written as $\frac{1}{\|w\|}$. The SVM objective is boiled to minimizing it:

$$\text{maximize} \frac{1}{\|w\|} \text{ or minimize} \|w\|$$

As, l2 optimization is often more stable than l1 optmization

$$\text{minimize} \frac{\|w\|^2}{2} \text{ such that } y_i(wx_i) + b \geq 0, \forall i \in \{1, 2, 3, ..., N\}$$

Non Linear Support Vector Machines:

Suppose the classes is now overlapping in feature space.The way to deal with this is to maximize M, but allow some points to be on the wrong side of the margin. Define the slack variables as $\zeta = \{\zeta_1, \zeta_2, ..., \zeta_N\}$. The two natural ways to modify the constraint are

$y_i(x_i^T \beta + \beta_0) \geq M - \zeta, \forall i \in \{1, 2, 3, ..., N\}$, or $y_i(x_i^T \beta + \beta_0) \geq M(1 - \zeta)$

Computations:

The problem is quadratic with linear inequality constraints. We can solve this problem using the quadratic programming algorithm. It is conventionally re-express in the form

$$\text{minimize}_{\beta, \beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N} \zeta_i$$

$$\text{subject to } \zeta_i \geq 0, y_i(wx_i) + b \geq 1 - \zeta_i, \forall i \in \{1, 2, 3, ..., N\}$$

Now the Langrangian (primal) function is

$$L_p = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N} \zeta_i - \sum_{i=1}^{N} \alpha_i[y_i(x^T \beta + \beta_0) - (1 - \zeta_i)] - \sum_{i=1}^{N} \mu_i \zeta_i$$

which we minimize w.r.t ζ_i,β and β_0. We get,

$$\beta = \sum_{i=1}^{N} \alpha_i y_i x_i$$

$$0 = \sum_{i=1}^{N} \alpha_i y_i$$

After substituting it into the langrangian(Dual), we get

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i x_j$$

we see that the solution of $f(x)$ is given by:

$$f(x) = h(x)^T \beta + \beta_0 = \sum_{i=1}^{N} \alpha_i y_i k[h(x) h(x_i)] + \beta_0$$

So, both the requires the inner product of the kernel function with the data points.In fact we need not specify the transformation $h(x)$ explicitly, but requires only the knowledge of the kernel function.

There are popular choices of kernel functions. These are:

- $\ker(x, x') = (1 + k[x, x'])^d$
- $\ker(x, x') = e^{-\gamma \|x - x'\|^2}$