Convex Optimization in Support Vector Machine

Lenny Fishler & Tin Tun Naing

April 2018

1 Introduction

1.1 A Brief Introduction to Machine Learning

Machine Learning is a very neat application of Mathematics, Statistics and Computer Science which help data professionals infer patterns about the data which are not trivially found with a naked eye. There are various approaches one can use, depending on the problem at hand as well as the size of the data set. Generally, there are two types of machine learning: Supervised Machine Learning and Unsupervised Machine Learning:

- **Supervised Machine Learning** is generally used for classification problems where the classes are known (Is this email spam or not spam? Is this handwritten digit a '7' or a '2'?). In this type of learning, the algorithm ingests data points, or training examples in which there is input and a 'label'. For example, there would be an image of a number and then the label for that number. (image of handwritten 2 with label '2', or text of an email, with label 'spam'). Some Neural Networks and SVMs are examples of Supervised ML Techniques.
- **Unsupervised Machine Learning** techniques come handy when there is a lot of data which is unlabeled. In other words, there is data, but little is known about this data, and we typically use computers to 'cluster' this data into appropriate categories, thereby figuring out patterns which were not known previously. K-means clustering is a good example of an unsupervised clustering algorithm.

1.2 Background of SVMs

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane, and falls in supervised machine learing category. In other words, given labeled training data, the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

SVMs use data transformations, also called kernels to nicely separate and transform the data so it is easily separable by a hyperplane, before transforming it into the original space and finding a good classification boundary. After getting the boundary, we can find coefficients of the optimal hyperplane that separates the two data sets. Ideally, by using the optimal hyperplane, given unlabelled data sets, we can classify them into groups.

1.3 Convex Optimization in Machine Learning

In optimization and machine learning, we are always trying to minimize the cost function, or error. For example, in linear regression, the cost function is least square equation. The main reason convex functions are useful in machine learning is that the existence of unique global minimum. For example,

in Gradient Descent algorithm, which concerns with finding the minimum of a function, the process will be less computationally intensive if the function is (strictly) convex.

2 Mathematical Concepts for SVM

2.1 Classification

Consider the data set: $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $x_i \in \mathbb{R}^p$ and $y_i \in \{1, -1\}$. We wish to find a function $f : \mathbb{R}^p \to \mathbb{R}$ such that

$$\begin{cases} f(x_i) > 0 & \text{if } y_i = 1 \\ f(x_i) < 0 & \text{if } y_i = -1 \end{cases} \tag{1}$$

Then using function f, given a random data set of x's, we can predict respective y's. In Classification, it is said that the level set of f at 0, $\{x|f(x) = 0\}$, separates/classifies/discriminates a data set into two sets.

2.2 Linear Discrimination

In (1), f can be any function. In classification, one can restrict f to be a certain type of function, such as linear, quadratic. If we restrict f to be an affine function, then geometrically, we are looking for a hyperplane such that $H = \{x|\langle a, x\rangle + b = 0\}$ and we can rewrite (1) as follows:

$$\begin{cases} \langle a, x_i\rangle + b > 0 & \text{if } y_i = 1 \\ \langle a, x_i\rangle + b < 0 & \text{if } y_i = -1 \end{cases} \tag{2}$$

where $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$. Now, for some $\delta > 0$, (2) is equivalent to

$$\begin{cases} \langle a, x_i\rangle + b \geq \delta & \text{if } y_i = 1 \\ \langle a, x_i\rangle + b \leq -\delta & \text{if } y_i = -1 \end{cases} \tag{3}$$

It is trivial that any a and b that satisfy (3) also satisfy (2). For the other direction, by multiplying (2) with some positive constant c large enough, then $(a', b') := (ca, cb)$.

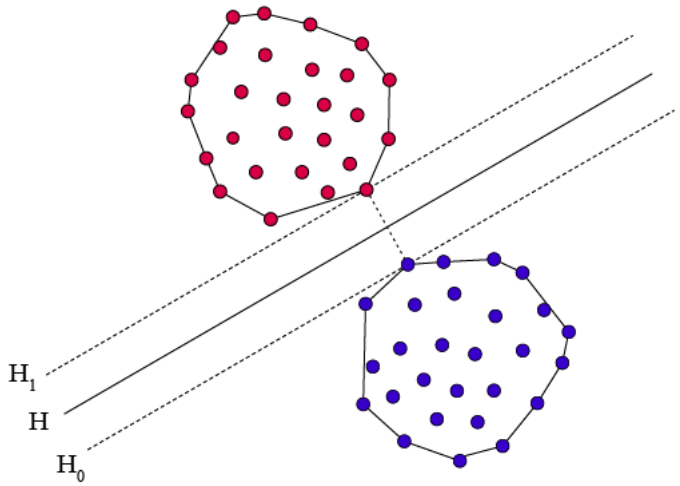

Figure 1: Hyperplanes separating two data sets

2.3 Separability

Farkas' Lemma: Let $A \in \mathbb{R}^{n\times m}$ and $U \in \mathbb{R}^n$. Then exactly one of the following statement is true:

1. There exists $X \in \mathbb{R}^m$ such that $AX \leq U$
2. There exists $Y \in \mathbb{R}^n$ such that $A^TY = 0$ and $Y \geq 0, Y^TU < 0$.

In other words, a vector is either in a closed convex cone or there exists a hyper plane separating that vector from the convex cone.

Theorem 1: Given two data sets, the convex hull of the two data sets intersects iff the two data sets are not linearly separable.
Proof: First let's prove $\Rightarrow$. Consider two data sets, $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$. Let z be in the intersection of the convex hulls of $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$. Then we can write z as the convex combination,

$$z = \sum_n \lambda_n x_n = \sum_n \mu_n y_n$$

where $\lambda_i, \mu_i > 0,\ \lambda_1 + \ldots + \lambda_n = \mu_1 + \ldots + \mu_n = 1$. Now if $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ are linearly separable, there exists a such that $\langle a, x_i\rangle + b > 0$ and $\langle a, y_i\rangle + b < 0$ for all x and y. However for z, we get

$$\langle a, z\rangle = \langle a, \sum_n \lambda_n x_n\rangle = \sum_n \lambda_n a^T x_n > \sum_n \lambda_n(-b) = -b\sum_n \lambda_n = -b$$
$$\langle a, z\rangle = \langle a, \sum_n \mu_n y_n\rangle = \sum_n \mu_n a^T y_n < \sum_n \mu_n(-b) = -b\sum_n \mu_n = -b$$

which is a contradiction. Thus, if the convex hull of the two data sets intersects, the two data sets are not linearly separable.

Now let's prove $\Leftarrow$. Consider (3). Let $I = \{i|y_i = 1\}$ and $J = \{j|y_j = -1\}$. In the matrix form, linear inequalities of (3) is equivalent to $AX \geq U$, where A is $p \times (n+1)$ whose $k-$th row is $[x_k^T\ 1]$ if $k \in I$ and $[-x_k^T\ -1]$ if $k \in J$, X is the concatenation of $a \in \mathbb{R}^n$ and b, and $U \in \mathbb{R}^p$, whose entries are $\delta > 0$.

If there does not exists $-X \in \mathbb{R}^{n+1}$ such that $A(-X) \leq -U$ (i.e there is no hyperplane separating the two data sets), then there exists $Y \in \mathbb{R}^p$ such that $A^TY = 0$ and $Y \geq 0, Y^T(-U) < 0$. Let the k-th term of Y be λ_k if $k \in I$ and be μ_k if $k \in J$. Then using the conditions of Y from Farkas' Lemma, $A^TY = 0$, and $U > 0, Y \geq 0, Y^TU > 0$, we have

$$\sum_{k\in I} \lambda_k x_k - \sum_{k\in J} \mu_k x_k = 0 \tag{4}$$
$$\sum_{k\in I} \lambda_k = \sum_{k\in J} \mu_k \tag{5}$$
$$\sum_{k\in i} \lambda_k \delta + \sum_{k\in J} \mu_k \delta > 0 \tag{6}$$
$$\sum_{k\in i} \lambda_k + \sum_{k\in J} \mu_k > 0 \tag{7}$$

since the last result above implies that $\lambda_k, \mu_k > 0$, we can define

$$\lambda_k' = \frac{\lambda_k}{\sum_{k\in I} \lambda_k} \text{ and } \mu_k' = \frac{\mu_k}{\sum_{k\in J} \mu_k}$$

Then dividing (4) with $\sum_{k\in I}\lambda_k$ gives us

$$\frac{\sum_{k\in I}\lambda_k x_k}{\sum_{k\in I}\lambda_k} = \frac{\sum_{k\in J}\mu_k x_k}{\sum_{k\in J}\mu_k}$$
$$\sum_{k\in I}\lambda'_k x_k = \sum_{k\in J}\mu'_k x_k$$

Also note that

$$\sum_{k\in I}\lambda'_k = \sum_{k\in J}\frac{\lambda_k}{\sum_{k\in I}\lambda_k} = \frac{\sum_{k\in I}\lambda_k}{\sum_{k\in I}\lambda_k} = 1$$

Similarly, we have $\sum_{k\in J}\mu'_k = 1$. Thus we have, $z = \sum_{k\in J}\mu'_k x_k = \sum_{k\in I}\lambda'_k x_k$, for some $z \in \mathbb{R}^p$, which implies z is in the convex hulls of both data sets. Thus, if the two data sets are not linear separable, then their convex hulls intersect. □

2.4 Lagrangian Multiplier

An optimization problem is typically written as follows:

$$\begin{aligned} &\min_x f(x) \\ &\text{s.t} \quad g_i(x) \le 0 \qquad i = 1\dots m \\ &\qquad h_i(x) = 0 \qquad i = 1\dots p \end{aligned}$$

$f(x)$ is called the objective function or the cost function, $g_i(x)$ is called inequality constraint function and $h_i(x)$ is called equality constraint function. In other words, we are looking for x^* such that

$$x^* = \inf\{f(x)|g_i(x) \le 0,\ i = 1\dots m,\ h_j(x) = 0,\ j = 1\dots p\}$$

The Lagrangian associated with the optimization problem above is:

$$L(x,\lambda,\mu) = f(x) + \sum_{i=1}^{m}\lambda_i g_i(x) + \sum_{i=1}^{p}\mu_i h_i(x) \tag{8}$$

where λ_i and μ_i are Lagrange multipliers.

2.5 Dual Problem

The Larangian Dual function of (8) is defined as follows:

$$F(\lambda,\mu) = \inf_x L(x,\lambda,\mu) \tag{9}$$

Then the dual optimization problem is:

$$\begin{aligned} &\max\ F(\lambda,\mu) \\ &\text{s.t}\ \lambda \ge 0 \end{aligned}$$

3 Finding the optimal hyperplane

In figure 1, the region between the two hyper plane is called the margin, $M = \{x|\langle a, -\delta \leq x\rangle + b \leq \delta\}$. The optimal hyperplane that separates the two data sets must lie exactly between the two data sets. Thus, by maximizing the distance between two hyper planes which lie in the margin, M, and separate the two data sets, we can find the optimal hyperplane.

Let hyperplanes $H_1 = \{x|\langle a, x\rangle + b = \delta\}$ and $H_2 = \{x|\langle a, x\rangle + b = -\delta\}$ (See figure 1). And let x_1 and x_2 be points on hyperplanes $H1$ and $H2$, respectively. Consider the line L that passes through x_1 in the direction of normal vector a. The equation of L will be $x_1 + at, t \in \mathbb{R}$. The intersection of L and H_2 is:

$$\begin{aligned}\langle a, x_2\rangle + b = a^T x_2 + b = a^T(x_1 + at) + b = -\delta \\ t = \frac{-\delta - b - a^T x_1}{a^T a} = \frac{-\delta - b - \delta - b}{||a||^2} = \frac{-2\delta}{||a||^2}\end{aligned}$$

Thus, the distance between H_1 and H_2 is:

$$|x_2 - x_1| = \Big|x_1 - \frac{2\delta}{||a||^2}\ a - x_1\Big| = \frac{2\delta}{||a||^2}\ ||a|| = \frac{2\delta}{||a||}$$

For mathematical convenience, set $\delta = 1$. Then, the distance between two hyper plane is $\frac{2}{||a||}$. Since we want to maximize the distance subject to the two hyper planes, we have an optimization problem:

$$\begin{aligned}\max\ & \frac{2}{||a||} \\ \text{s.t}\ & \langle a, x_k\rangle + b \geq 1 \qquad k \in I \\ & \langle a, x_k\rangle + b \leq -1 \qquad k \in J\end{aligned}$$

Note that maximizing $\frac{2}{||a||}$ is equivalent to minimizing $||a||$. Thus, the dual problem of our original optimal problem is

$$\begin{aligned}\min\ & \frac{||a||^2}{4} \\ \text{s.t}\ & \langle a, x_k\rangle + b \geq 1 \qquad k \in I \\ & \langle a, x_k\rangle + b \leq -1 \qquad k \in J\end{aligned}$$

and its Lagrangian function is:

$$L(a, b, \lambda, \mu) = \frac{||a||^2}{4} + \sum_{k\in I} \lambda_k\big(1 - \langle a, x_k\rangle - b\big) + \sum_{k\in J} \mu_k\big(1 + \langle a, x_k\rangle + b\big) \tag{10}$$

Note that, in (10), L is convex since it is just a summation of convex functions and it is also quadratic. Thus, it is guaranteed to have a unique global minimum and minimum is achieved when $a = 2\big(\sum_{k\in I} \lambda_k x_k - \sum_{k\in J} \mu_k x_k\big)$ and $\sum_{k\in I} \lambda_k = \sum_{k\in J} \mu_k$. Thus, for Lagragian function, we have:

$$F = \inf_{a,b} L(a, b, \lambda, \mu) = -\left|\left| \sum_{k\in I} \lambda_k x_k - \sum_{k\in j} \mu_k x_k \right|\right|^2 + \sum_{k\in I} \lambda_k + \sum_{k\in J} \mu_k \tag{11}$$

Let $s = \sum_{k\in I} \lambda_k = \sum_{k\in J} \mu_k$, and $\lambda'_k = \frac{\lambda_k}{s}, \mu'_k = \frac{\mu_k}{s}$, then the optimization problem of (11), which is the dual optimization problem of (10), is

$$\max_{\lambda,\mu} \; -\left|\left|\sum_{k\in I} \lambda_k x_k - \sum_{k\in J} \mu_k x_k\right|\right|^2 + \sum_{k\in I} \lambda_k + \sum_{k\in J} \mu_k \tag{12}$$

$$= -s^2\left|\left|\sum_{k\in I} \frac{\lambda_k}{s} x_k - \sum_{k\in J} \frac{\mu_k}{s} x_k\right|\right|^2 + 2s \tag{13}$$

$$\max_{s,\lambda',\mu'} \; -s^2\left|\left|\sum_{k\in I} \lambda'_k x_k - \sum_{k\in J} \mu_k x'_k\right|\right|^2 + 2s \tag{14}$$

$$\text{s.t} \; \sum_{k\in I} \lambda'_k = \sum_{k\in J} \mu'_k = 1 \tag{15}$$

$$\lambda \geq 0, \mu \geq 0 \tag{16}$$

In (14), s is achieved its maximum, when $s^{-1} = \left|\left|\sum_{k\in I} \lambda'_k x_k - \sum_{k\in J} \mu_k x'_k\right|\right|^2$. Thus, maximizing s is equivalent to minimizing $\left|\left|\sum_{k\in I} \lambda'_k x_k - \sum_{k\in j} \mu_k x'_k\right|\right|$. So we have,

$$\min_{\lambda',\mu'} \left|\left|\sum_{k\in I} \lambda'_k x_k - \sum_{k\in j} \mu_k x'_k\right|\right| \tag{17}$$

$$\text{s.t} \; \sum_{k\in I} \lambda'_k = \sum_{k\in J} \mu'_k = 1 \tag{18}$$

$$\lambda', \mu' \geq 0 \tag{19}$$

which is equivalent to our original optimal problem of maximizing the distance between the two hyperplane. In other words, (17) is minimizing the distance between one point in conv $\left(x_k\right)_{k\in I}$ and one point in conv $\left(x_k\right)_{k\in J}$.

Remark: In case that the data sets are not linearly separable, one can use Kernel Method (Kernel Trick) to add another dimension to the data sets. For example, if we are given two dimensionals data sets that cannot be linearly separable, we can transform those data sets into three dimensionals space, where it may be able to be linearly separatable.

References

[1] Rockafellar, R. Tyrell. *Convex Analysis*. 1970.

[2] Kowalczyk, Alexandre. *Support Vector Machines Succinctly*. Syncfusion, Inc. 2017.

[3] Singer, Yaron. *Advanced Optimization*. 2016.
`https://people.seas.harvard.edu/~yaron/AM221-S16/lecture_notes/AM221_lecture13.pdf`

[4] Patel, Savan. "SVM (Support Vector Machine) —Theory." Medium.com May 3, 2017.
`https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72`