Machine Learning, Spring 2019
EN 2105
MWF 9-9:50am

Chao Lan
EN 4084A
clan@uwyo.edu

Linear Support Vector Machine (LSVM)

Introduction

Linear support vector machine (LSVM) is designed for binary classification. It finds a linear decision boundary that maximally separates data of the two classes, denoted by $Y = \{-1, +1\}$.

A linear decision boundary is a linear hyper-plane G

$$G := \{x;\ x^T\beta + \beta_0 = 0\}, \tag{1}$$

where β is a normal of G (defined as any vector perpendicular to G) and β_0 is a bias term.

[*Discussion*] How to interpret G and β in the two-dimensional and three-dimensional spaces?

[*Exercise*] Verify that every point on G satisfies $x^T\beta + \beta_0 = 0$.

[*Exercise*] Prove that β is perpendicular to G (i.e. it is perpendicular to every vector in G).

The distance between a point x and G is the smallest distance between x and any point on G. It can be shown that the signed distance between x and G is

$$d(x, G) = \frac{x^T\beta + \beta_0}{||\beta||}. \tag{2}$$

By 'signed', we mean $d(x, G) > 0$ if x is on the side of G where β is pointing at, and $d(x, G) < 0$ if x is on the other side of G. Note that if $||\beta|| = 1$, the signed distance is $d(x, G) = x^T\beta + \beta_0$.

[*Exercise*] Prove the signed distance has the form of (2).

We can use G as a linear decision boundary and classify x to class $y = +1$ if $d(x, G) > 0$ and to class $y = -1$ otherwise. We say two classes are linearly separable if they can be separated by some linear decision boundary.

Optimal Separating Hyperplane and Max-Margin Classifier

In this section, we will assume the two classes are linearly separable. Without loss of generality, we also assume G will classify x_i to class $y = +1$ if $d(x_i, G) > 0$ and to class $y = -1$ otherwise.

[*Discussion*] If two classes can be linearly separable by multiple G's, which one is more desirable?

The optimal separating hyperplane G is the one with the maximum distance to training in-

stances. It finds such a G by solving the following optimization problem

$$\begin{aligned} &\max_{\beta, \beta_0, ||\beta||=1} M \\ &\text{s.t. } y_i\,(x_i^T\beta + \beta_0) \geq M,\ i = 1, \ldots, n. \end{aligned} \tag{3}$$

Since $x_i^T\beta + \beta_0$ is the signed distance if $||\beta|| = 1$, it is easy to verify that $y_i(x_i^T\beta + \beta_0)$ is the unsigned distance if x_i is correctly classified. Assuming all x_i are correctly classified, then (3) maximizes the distance lower bound M (which may indirectly maximize the distance).

[*Discussion*] What if x_i is misclassified by a candidate solution G?

Problem (3) can be simplified for easier optimization and analysis in three steps:

(1) remove $||\beta|| = 1$ by replacing $y_i(x_i^T\beta + \beta_0) \geq M$ with $y_i(x_i^T\beta + \beta_0) \geq M||\beta||$.

(2) replace $y_i(x_i^T\beta + \beta_0) \geq M||\beta||$ with $y_i(x_i^T\beta + \beta_0) \geq C$ with fixed $C = M||\beta||$.

(3) fix $C = 1$.

Then we have a new optimization problem

$$\begin{aligned} &\min_{\beta, \beta_0} \frac{1}{2}||\beta||^2 \\ &\text{s.t. } y_i\,(x_i^T\beta + \beta_0) \geq 1,\ i = 1, \ldots, n. \end{aligned} \tag{4}$$

Step (2) will restrict the magnitude of β but not its direction. Since G is only determined by the direction of β, step (2) will not change the optimal G.

[*Discussion*] Why would step (1) not change the optimal solution?

In (3), we can interpret $1/||\beta||$ as the thickness of a slab centered at G. All (correctly classified) x_i are excluded from the slab. Minimizing $||\beta||$ is equivalent to maximizing the slab. Therefore, the classifier learned by (4) is called the max-margin classifier.

Duality and Support Vector Classifier

To solve (4), we can first apply the Lagrange Multiplier and convert it into an unconstrained optimization problem. The new objective function is

$$J(\beta, \beta_0) = \frac{1}{2}||\beta||^2 - \sum\nolimits_{i=1}^{n} \alpha_i[y_i(x_i^T\beta + \beta_0) - 1], \tag{5}$$

where $\alpha_i \geq 0$ for $i = 1, ..., n$.

Since J is a quadratic function. We can apply the critical point method and have

$$\frac{\partial}{\partial\beta} J(\beta, \beta_0) = 0 \qquad \Rightarrow \qquad \beta = \sum\nolimits_{i=1}^{n} \alpha_i y_i x_i. \tag{6}$$

$$\frac{\partial}{\partial\beta_0} J(\beta, \beta_0) = 0 \qquad \Rightarrow \qquad \sum\nolimits_{i=1}^{n} \alpha_i y_i = 0. \tag{7}$$

[*Exercise*] Derive (6) and (7).

Plugging (6,7) back to (5), we eliminate β, β_0 and have a new objective function of α's

$$J_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \, \alpha_j \, y_i \, y_j \, x_i^T \, x_j. \tag{8}$$

[*Exercise*] Derive (8).

$J_D(\alpha)$ sets a lower bound of $J(\beta, \beta_0)$, and thus should be maximized[1] to find the optimal solution, under constraints (7), $\alpha_i \geq 0$ and the Karush-Kuhn-Tucker (KKT) condition

$$\alpha_i [y_i (x_i^T \beta + \beta_0) - 1] = 0, \quad i = 1, \ldots, n. \tag{9}$$

This new optimization problem is the Wolfe dual of the original optimization problem. It is a simpler convex optimization task and suggests an important insight that the optimal β is a linear combination of training instances based on (6).

The KKT condition (9) gives another important insight:

- if $\alpha_i > 0$, then $y_i (x_i^T \beta + \beta_0) - 1 = 0$. So x_i is correctly classified and lies on the margin.

- if $y_i (x_i^T \beta + \beta_0) - 1 \neq 0$, which means x_i is not on the margin, then $\alpha_i = 0$.

This suggests only instances on the margin can have $\alpha_i > 0$ and thus be used to express β_*. Such an instance is a support vector, and the classifier is support vector classifier (SVC). It is also called hard-margin linear SVM. It is hard as it does not tolerate misclassified instances (assuming data are linearly separable). It is linear as its assumes a linear separating hyperplane.

[*Discussion*] How if data are not linearly separable?

Soft-Margin Linear SVM

Soft-margin linear SVM does not assume data are linearly separable. It learns a linear decision boundary that tolerates misbehaved instances, including those that are misclassified or lie inside the slab (Fig 1, right). Mathematically, it does so by introducing a slack variable ϵ_i for instance x_i and changing the hard-margin constraint $y_i (x_i^T \beta + \beta_0) \geq 1$ to the soft-margin constraint

$$y_i (x_i^T \beta + \beta_0) \geq 1 - \epsilon_i, \quad \epsilon_i \geq 0 \tag{10}$$

so that the distance between x_i and G is now lower bounded by $\frac{1}{||\beta||}(1 - \epsilon_i)$ instead of $\frac{1}{||\beta||}$.

- if $\epsilon_i > 1$ then $y_i (x_i^T \beta + \beta_0)$ can be negative, which means x_i can be misclassified

- if $1 > \epsilon_i > 0$ then $y_i (x_i^T \beta + \beta_0)$ can be smaller than the thickness of the slab $1/||\beta||$, which means x_i can lie inside the slab

Although soft-margin linear SVM tolerates misbehaved instances, it aims to minimize the behaviors by minimizing $\sum_{i=1}^{n} \epsilon_i$. Adding this and (10) back to (4), we obtain the optimization

[1] We will explained why it should be maximized when introducing of KKT conditions.

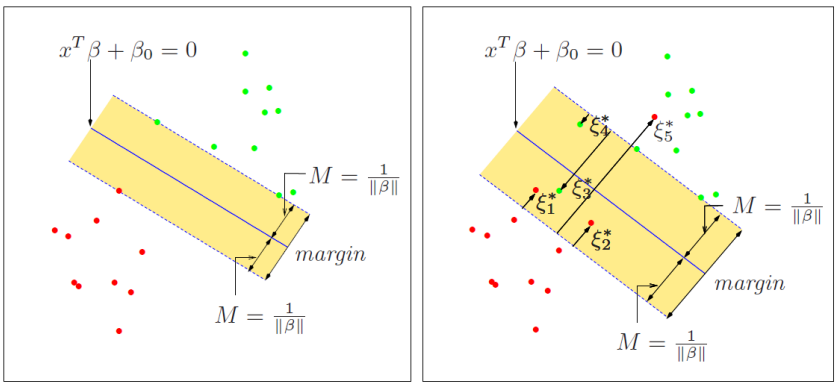

Figure 1: Separable Case (left) and Non-Separable Case (right)

problem of soft-margin linear SVM:

$$\begin{aligned} \min_{\beta,\beta_0,\epsilon} \ & \frac{1}{2}||\beta||^2 + C\sum_{i=1}^n \epsilon_i \\ \text{s.t. } & y_i(x_i^T\beta + \beta_0) \geq 1 - \epsilon_i, \\ & \epsilon_i \geq 0 \\ & i = 1, 2, \ldots, n, \end{aligned} \tag{11}$$

where C is a hyper-parameter controlling the degree of tolerance.

[*Discussion*] What happens if $C = \infty$?

Problem (11) can be solved in a similar way as solving hard-margin linear SVM. First, we derive a dual problem using the Lagrange multiplier; the Lagrange function is

$$L(\beta, \beta_0, \epsilon, \mu) = \frac{1}{2}||\beta||^2 + C\sum_{i=1}^n \epsilon_i - \sum_{i=1}^n \alpha_i[y_i(x_i^T\beta + \beta_0) - (1 - \epsilon_i)] - \sum_{i=1}^n \mu_i\epsilon_i, \tag{12}$$

where $\alpha_i, \mu_i \geq 0$. Setting the derivatives w.r.t. β, β_0 and ϵ_i to zero, respectively, we have

$$\beta = \sum_{i=1}^n \alpha_i y_i x_i \tag{13}$$

$$0 = \sum_{i=1}^n \alpha_i y_i \tag{14}$$

$$\alpha_i = C - \mu_i, \tag{15}$$

[*Exercise*] Derive (15).

Plugging these back to (12) gives the Wolfe dual objective function

$$L_D(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n \alpha_i\alpha_j y_i y_j x_i^T x_j. \tag{16}$$

[*Exercise*] Derive (16).

Similar to hard-margin LSVM, we can maximize $L_D(\alpha)$ to find the optimal solution, under the constraints that $0 \leq \alpha_i \leq C$ from (15), $\sum_{i=1}^n \alpha_i y_i = 0$ from (14), and the following three constraints derived from the KKT conditions:

$$\alpha_i[y_i(x_i^T\beta + \beta_0) - (1 - \epsilon_i)] = 0, \tag{17}$$

$$\mu_i \epsilon_i = 0, \tag{18}$$

$$y_i(x_i^T \beta + \beta_0) - (1 - \epsilon_i) \geq 0, \forall i. \tag{19}$$

This is a simpler convex optimization problem.

(13) suggests the optimal solution is (again) a linear combination of training instances. Based on (17), α_i is non-zero only if

$$y_i(x_i^T \beta + \beta_0) - (1 - \epsilon_i) = 0. \tag{20}$$

- if $\epsilon_i = 0$, then $y_i(x_i^T \beta + \beta_0) = 1$ which means x_i is correctly classified and lies on the margin.

- if $\epsilon_i > 0$, then $y_i(x_i^T \beta + \beta_0) < 1$. This can be divided into two cases:

  - if $\epsilon_i < 1$, then $0 < y_i(x_i^T \beta + \beta_0) < 1$ so x_i is correctly classified but inside the slab

  - if $\epsilon_i > 1$, then $y_i(x_i^T \beta + \beta_0) < 0$ which means x_i is mis-classified

  - in addition, (15) and (18) imply $\alpha_i = C$.

In summary, only three types of instances will be used to express the optimal decision boundary: (1) instances that are correctly classified and lie on the margin, (2) instances that are correctly classified and lie inside the margin, (3) instances that are misclassified. We call these instances support vectors, and the classification method soft-margin linear SVM. In addition, support vectors (2) and (3) will have $\alpha_i = C$.

KKT Conditions: An Introduction and its Application in LSVM

The KarushKuhnTucker (KKT) conditions guarantee that the solution of a dual problem is the solution of the original problem.

We first review some concepts and properties. Consider a standard optimization problem

$$\begin{aligned} \min \ & J(x) \\ \text{s.t.} \ & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_j(x) = 0, \quad j = 1, \ldots, p. \end{aligned} \tag{21}$$

Any x satisfying all constraints (but not necessarily minimize $J(x)$) is a feasible point. Applying the Lagrange multiplier method, we construct the Lagrange function

$$L(x, \lambda, \upsilon) = J(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} \upsilon_j h_i(x), \tag{22}$$

where $\lambda_i, \upsilon_j \geq 0$ are the Lagrange multipliers – or, dual variables. An important observation is that, for any $\lambda \geq 0$[2] and υ, any feasible point x satisfies

$$L(x, \lambda, \upsilon) \leq J(x). \tag{23}$$

[*Exercise*] Prove (23).

[2]We write $\lambda \geq 0$ to indicate that every $\lambda_j \geq 0$.

Let $p_* = \min_x J(x)$. Define the Lagrange dual function as

$$g(\lambda, \upsilon) = \min_x L(x, \lambda, \upsilon). \tag{24}$$

An important implication of (23) is that

$$g(\lambda, \upsilon) \leq p_*. \tag{25}$$

[*Exercise*] Prove (25).

Based on (25), to find p_* we want to maximize $g(\lambda, \upsilon)$, i.e.,

$$\max_{\lambda, \upsilon} \; g(\lambda, \upsilon), \quad s.t. \quad \lambda \geq 0. \tag{26}$$

This justifies why LSVM maximizes the Lagrange dual objective function.

[*Exercise*] Derive the maximization problem of LSVM based on the above framework.

In the above analysis, (21) is the primal problem and (26) is its dual problem. The dual problem is always convex, disregarding if the primal problem is convex or not. Any (λ, υ) satisfying $\lambda \geq 0$ and $g(\lambda, \upsilon) > -\infty$ is a dual feasible. A solution to (26), denoted by (λ_*, υ_*), is a dual optimal. The optimal duality gap is $p_* - g(\lambda_*, \upsilon_*)$. If the gap is zero, we say strong duality holds.

Now we introduce the KKT conditions. Assume strong duality holds. Let x_* be a primal optimal and λ_*, υ_* be a dual optimal. There is

$$\begin{aligned} J(x_*) &= g(\lambda_*, \upsilon_*) \\ &= \min_x L(x, \lambda_*, \upsilon_*) \\ &= \min_x J(x) + \sum\nolimits_{i=1}^m \lambda_{*i} f_i(x) + \sum\nolimits_{j=1}^p \upsilon_{*j} h_i(x) \\ &\leq J(x_*) + \sum\nolimits_{i=1}^m \lambda_{*i} f_i(x_*) + \sum\nolimits_{j=1}^p \upsilon_{*j} h_i(x_*) \\ &\leq J(x_*). \end{aligned} \tag{27}$$

[*Exercise*] Verify (27).

In (27), since LHS = RHS, the two inequalities must both be equality. This implies

- x_* minimizes $L(x, \lambda_*, \upsilon_*)$, which implies it is a critical point of L, i.e.,

$$J'(x_*) + \sum\nolimits_{i=1}^m \lambda_i f_i'(x_*) + \sum\nolimits_{j=1}^p \upsilon_j h_i'(x_*) = 0. \tag{28}$$

- $\sum_{i=1}^m \lambda_i f_i(x_*) = 0$, which implies the complementary slackness condition

$$\lambda_i f_i(x_*) = 0, \quad i = 1, \ldots, m. \tag{29}$$

[*Exercise*] Verify (29).

Putting all together, the primal optimal point x_* should satisfy

$$f_i(x_*) \leq 0 \tag{30}$$

$$h_i(x_*) = 0 \tag{31}$$

$$\lambda_i f_i(x_*) = 0 \tag{32}$$

$$J'(x_*) + \sum\nolimits_{i=1}^m \lambda_i f_i'(x_*) + \sum\nolimits_{j=1}^p \upsilon_j h_i'(x_*) = 0 \tag{33}$$

These are called the KKT conditions.

Now we see how the KKT conditions are applied to LSVM.

For hard-margin LSVM, (6) (7) are derived from (33), and (9) is from (32).

For soft-margin LSVM, (14) is from (33), (17, 18) are from (32), and the rest are from (30).