Support Vector Machines (1/2)

Dr. Víctor Uc Cetina

Universidad Autónoma de Yucatán

Content

Confident predictions

- Consider logistic regression where the probability $p(y = 1|\mathrm{x}; \theta)$ is modeled by $h_\theta(\mathrm{x}) = g(\theta^\top \mathrm{x})$.

Confident predictions

- Consider logistic regression where the probability $p(y = 1|\mathrm{x}; \theta)$ is modeled by $h_\theta(\mathrm{x}) = g(\theta^\top \mathrm{x})$.
- We would then predict "1" on an input x if and only if $h_\theta(\mathrm{x}) \geq 0.5$, or equivalently if $\theta^\top \mathrm{x} \geq 0$.

Confident predictions

- Consider logistic regression where the probability $p(y = 1|\mathrm{x}; \theta)$ is modeled by $h_\theta(\mathrm{x}) = g(\theta^\top \mathrm{x})$.
- We would then predict "1" on an input x if and only if $h_\theta(\mathrm{x}) \geq 0.5$, or equivalently if $\theta^\top \mathrm{x} \geq 0$.
- Consider a positive training example ($y = 1$), the larger the $\theta^\top \mathrm{x}$ is, the larger also is $h_\theta(\mathrm{x}) = p(y = 1|\mathrm{x}; w, b)$, and thus also the higher our degree of confidence that the label is 1.

Confident predictions

- Consider logistic regression where the probability $p(y = 1|\mathrm{x}; \theta)$ is modeled by $h_\theta(\mathrm{x}) = g(\theta^\top \mathrm{x})$.
- We would then predict "1" on an input x if and only if $h_\theta(\mathrm{x}) \geq 0.5$, or equivalently if $\theta^\top \mathrm{x} \geq 0$.
- Consider a positive training example ($y = 1$), the larger the $\theta^\top \mathrm{x}$ is, the larger also is $h_\theta(\mathrm{x}) = p(y = 1|\mathrm{x}; w, b)$, and thus also the higher our degree of confidence that the label is 1.
- Thus, informally we can think of our prediction as being a very confident one that $y = 1$ if $\theta^\top \mathrm{x} \gg 0$.

Confident predictions

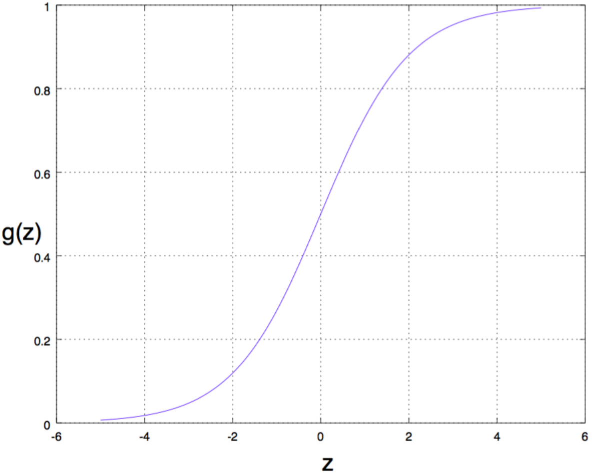


Figure: Sigmoid function (where $z = \theta^\top x$).

Confident predictions

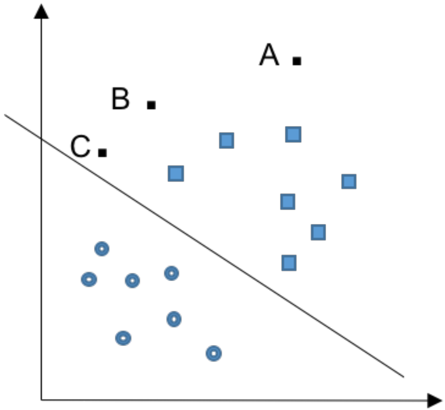


Figure: Separating hyperplane, this is the line given by the equation $\theta^\top x = 0$.

Confident predictions

- Similarly, we think of logistic regression as making a very confident prediction of $y = 0$, if $\theta^\top \mathrm{x} \ll 0$.

Confident predictions

- Similarly, we think of logistic regression as making a very confident prediction of $y = 0$, if $\theta^\top \mathrm{x} \ll 0$.
- Given a training set, again informally it seems that we'd have found a good fit to the training data if we can find θ so that:

  $\theta^\top \mathrm{x} \gg 0$ whenever $y^{(i)} = 1$

  and

  $\theta^\top \mathrm{x} \ll 0$ whenever $y^{(i)} = 0$

  since this would reflect a very confident (and correct) set of classifications for all the training examples.

Functional Margin

- We will consider a linear classifier for a binary classification problem with labels y and features x.

Functional Margin

- We will consider a linear classifier for a binary classification problem with labels y and features x.
- We will use $y \in \{-1, 1\}$ and instead of the vector θ, we will use parameters w, b.

Functional Margin

- We will consider a linear classifier for a binary classification problem with labels y and features x.
- We will use $y \in \{-1, 1\}$ and instead of the vector θ, we will use parameters w, b.
- So our classifier can be written as $h_{w,b}(\mathrm{x}) = g(w^\top \mathrm{x} + b)$.

Functional Margin

- We will consider a linear classifier for a binary classification problem with labels y and features x.
- We will use $y \in \{-1, 1\}$ and instead of the vector θ, we will use parameters w, b.
- So our classifier can be written as $h_{w,b}(\mathrm{x}) = g(w^\top \mathrm{x} + b)$.
- Here, $g(z) = 1$ if $z \geq 0$, and $g(z) = -1$ otherwise.

Functional Margin

- Given a training example $(x^{(i)}, y^{(i)})$, we define the functional margin of (w, b) with respect to that training example as

  $\hat{\gamma}^{(i)} = y^{(i)}(w^\top \mathrm{x}^{(i)} + b).$

Functional Margin

- Given a training example $(x^{(i)}, y^{(i)})$, we define the functional margin of (w, b) with respect to that training example as

  $\hat{\gamma}^{(i)} = y^{(i)}(w^\top \mathrm{x}^{(i)} + b).$

- If $y^{(i)} = 1$, then the functional margin to be large, then we need $w^\top \mathrm{x} + b$ to be a large positive number.

Functional Margin

- Given a training example $(x^{(i)}, y^{(i)})$, we define the functional margin of (w, b) with respect to that training example as

  $\hat{\gamma}^{(i)} = y^{(i)}(w^\top \mathrm{x}^{(i)} + b).$

- If $y^{(i)} = 1$, then the functional margin to be large, then we need $w^\top \mathrm{x} + b$ to be a large positive number.
- Conversely, if $y^{(i)} = -1$, then for the functional margin to be large, then we need $w^\top \mathrm{x} + b$ to be a large negative number.

Functional Margin

- Given a training example $(x^{(i)}, y^{(i)})$, we define the functional margin of (w, b) with respect to that training example as

  $\hat{\gamma}^{(i)} = y^{(i)}(w^\top \mathrm{x}^{(i)} + b).$

- If $y^{(i)} = 1$, then the functional margin to be large, then we need $w^\top \mathrm{x} + b$ to be a large positive number.
- Conversely, if $y^{(i)} = -1$, then for the functional margin to be large, then we need $w^\top \mathrm{x} + b$ to be a large negative number.
- Moreover, if $y^{(i)}(w^\top \mathrm{x} + b) > 0$, then our prediction on this example is correct. Hence, a large functional margin represents a confident and a correct prediction.

Functional Margin

- For a linear classifier with the choice of $h_{w,b}(\mathrm{x}) = g(w^\top \mathrm{x} + b)$, note that if we replace w with $2w$ and b with $2b$, then $g(w^\top x + b) = g(2w^\top x + 2b)$ and this would not change $h_{w,b}(\mathrm{x})$.

Functional Margin

- For a linear classifier with the choice of $h_{w,b}(\mathrm{x}) = g(w^\top \mathrm{x} + b)$, note that if we replace w with $2w$ and b with $2b$, then $g(w^\top x + b) = g(2w^\top x + 2b)$ and this would not change $h_{w,b}(\mathrm{x})$.
- Hence $h_{w,b}(\mathrm{x})$ depends only on the sign and not on the magnitude of $w^\top \mathrm{x} + b$.

Functional Margin

- For a linear classifier with the choice of $h_{w,b}(\mathrm{x}) = g(w^\top \mathrm{x} + b)$, note that if we replace w with $2w$ and b with $2b$, then $g(w^\top x + b) = g(2w^\top x + 2b)$ and this would not change $h_{w,b}(\mathrm{x})$.
- Hence $h_{w,b}(\mathrm{x})$ depends only on the sign and not on the magnitude of $w^\top \mathrm{x} + b$.
- However, replacing (w, b) with $(2w, 2b)$ also results in multiplying our functional margin by a factor of 2.

Functional Margin

- For a linear classifier with the choice of $h_{w,b}(\mathrm{x}) = g(w^\top \mathrm{x} + b)$, note that if we replace w with $2w$ and b with $2b$, then $g(w^\top x + b) = g(2w^\top x + 2b)$ and this would not change $h_{w,b}(\mathrm{x})$.
- Hence $h_{w,b}(\mathrm{x})$ depends only on the sign and not on the magnitude of $w^\top \mathrm{x} + b$.
- However, replacing (w, b) with $(2w, 2b)$ also results in multiplying our functional margin by a factor of 2.
- Thus, it seems that by exploiting our freedom to scale w and b, we can make the functional margin arbitrarily large without really changing anything meaningful.

Functional Margin

- Intuitively, it might therefore make sense to impose some sort of normalization condition such as that $||w||_2 = 1$.

Functional Margin

- Intuitively, it might therefore make sense to impose some sort of normalization condition such as that $||w||_2 = 1$.
- We might replace (w, b) with $(w/||w||_2, b/||w||_2)$ and instead consider the functional margin of $(w/||w||_2, b/||w||_2)$.

Functional Margin

- Intuitively, it might therefore make sense to impose some sort of normalization condition such as that $||w||_2 = 1$.
- We might replace (w, b) with $(w/||w||_2, b/||w||_2)$ and instead consider the functional margin of $(w/||w||_2, b/||w||_2)$.
- Given a training set $S = \{(\mathrm{x}^{(i)}, y^{(i)}); i = 1, \ldots, m\}$, we define the functional margin of (w, b) with respect to S as the smallest of the functional margins of the individual training examples, denoted by $\hat{\gamma}$:

$$\hat{\gamma} = \min_{i=1,\ldots,m} \hat{\gamma}^{(i)}.$$

Geometric Margin

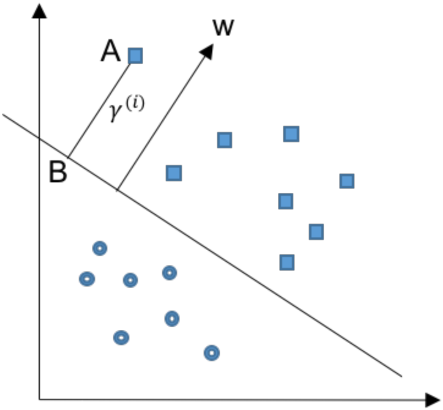
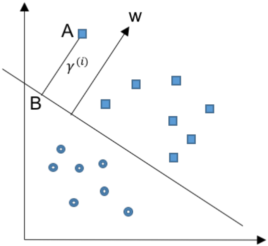


Figure: The decision boundary corresponding to (w, b) is shown, along with the vector w. Note that w is orthogonal (at 90°) to the separating hyperplane.

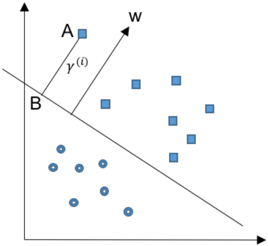
Geometric Margin

- How can we find $\gamma^{(i)}$?



Geometric Margin

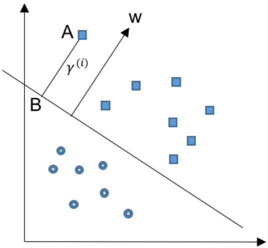
- How can we find $\gamma^{(i)}$?



- Since A represents example $x^{(i)}$, point B is given by $x^{(i)} - \gamma^{(i)} \cdot \frac{w}{||w||}$.

Geometric Margin

- How can we find $\gamma^{(i)}$?



- Since A represents example $\mathrm{x}^{(i)}$, point B is given by $\mathrm{x}^{(i)} - \gamma^{(i)} \cdot \frac{w}{||w||}$.
- But this point lies on the decision boundary and satisfies $w^\top \mathrm{x} + b = 0$.

Geometric Margin

- Hence

$$w^\top \left(x^{(i)} - \gamma^{(i)} \cdot \frac{w}{||w||} \right) + b = 0.$$

Geometric Margin

- Hence

$$w^\top \left(\mathrm{x}^{(i)} - \gamma^{(i)} \cdot \frac{w}{||w||} \right) + b = 0.$$

- Solving for $\gamma^{(i)}$

$$\gamma^{(i)} = \frac{w^\top \mathrm{x}^{(i)} + b}{||w||} = \left(\frac{w}{||w||} \right)^\top \mathrm{x}^{(i)} + \frac{b}{||w||}.$$

Geometric Margin

- Hence

$$w^\top \left(\mathrm{x}^{(i)} - \gamma^{(i)} \cdot \frac{w}{||w||} \right) + b = 0.$$

- Solving for $\gamma^{(i)}$

$$\gamma^{(i)} = \frac{w^\top \mathrm{x}^{(i)} + b}{||w||} = \left(\frac{w}{||w||} \right)^\top \mathrm{x}^{(i)} + \frac{b}{||w||}.$$

- More generally we define the geometric margin for (w, b) with respect to a training example $(\mathrm{x}^{(i)}, y^{(i)})$ to be

$$\gamma^{(i)} = y^{(i)} \left(\left(\frac{w}{||w||} \right)^\top \mathrm{x}^{(i)} + \frac{b}{||w||} \right).$$

Geometric Margin

- Given a training set $S = \{(x^{(i)}, y^{(i)}); i = 1, \ldots, m\}$, we also define the geometric margin of (w, b) with respect to S to be the smallest of the geometric margins on the individual training examples:

$$\gamma = \min_{i=1,\ldots,m} \gamma^{(i)}.$$

Functional and Geometric Margins

- Functional margin

$$\hat{\gamma}^{(i)} = y^{(i)}(w^\top \mathrm{x}^{(i)} + b).$$

Functional and Geometric Margins

- Functional margin

$$\hat{\gamma}^{(i)} = y^{(i)}(w^\top x^{(i)} + b).$$

- Geometric margin

$$\gamma^{(i)} = y^{(i)}\left(\left(\frac{w}{||w||}\right)^\top x^{(i)} + \frac{b}{||w||}\right).$$

Functional and Geometric Margins

- Functional margin

  $$\hat{\gamma}^{(i)} = y^{(i)}(w^\top \mathrm{x}^{(i)} + b).$$

- Geometric margin

  $$\gamma^{(i)} = y^{(i)}\left(\left(\frac{w}{||w||}\right)^\top \mathrm{x}^{(i)} + \frac{b}{||w||}\right).$$

- Note that if $||w|| = 1$, then the functional margin equals the geometric margin. This thus gives us a way of relating these two different notions of margin.

Optimal Margin Classifier

- Given a training set, it seems natural to try to find a decision boundary that maximizes the geometric margin, since this would reflect a very confident set of predictions on the training set and a good "fit" to the training data.

Optimal Margin Classifier

- Given a training set, it seems natural to try to find a decision boundary that maximizes the geometric margin, since this would reflect a very confident set of predictions on the training set and a good "fit" to the training data.
- Specifically, this will result in a classifier that separates the positive and the negative training examples with a "gap" (geometric margin).

Optimal Margin Classifier

- Given a training set, it seems natural to try to find a decision boundary that maximizes the geometric margin, since this would reflect a very confident set of predictions on the training set and a good "fit" to the training data.
- Specifically, this will result in a classifier that separates the positive and the negative training examples with a "gap" (geometric margin).
- We will assume that we are given a training set that is linearly separable; i.e., that it is possible to separate the positive and negative examples using some separating hyperplane. How can we find the one that achieves the maximum geometric margin?

Optimal Margin Classifier

- We can pose the following optimization problem:

$$\begin{aligned} \max_{\gamma,w,b} \quad & \gamma \\ \text{s.t.} \quad & y^{(i)}(w^\top x^{(i)} + b) \geq \gamma, i = 1, \ldots, m \\ & ||w|| = 1. \end{aligned}$$

Optimal Margin Classifier

- We can pose the following optimization problem:

$$\max_{\gamma,w,b} \quad \gamma$$

$$\text{s.t.} \quad y^{(i)}(w^\top x^{(i)} + b) \geq \gamma, i = 1, \ldots, m$$

$$||w|| = 1.$$

- We want to maximize γ, subject to each training example having functional margin at least γ.

Optimal Margin Classifier

- We can pose the following optimization problem:

$$\begin{aligned} \max_{\gamma,w,b} \quad & \gamma \\ \text{s.t.} \quad & y^{(i)}(w^\top x^{(i)} + b) \geq \gamma, i = 1, \ldots, m \\ & ||w|| = 1. \end{aligned}$$

- We want to maximize γ, subject to each training example having functional margin at least γ.
- The $||w|| = 1$ constraint ensures that the functional margin equals to the geometric margin, so we are also guaranteed that all the geometric margins are at least γ.

Optimal Margin Classifier

- We can pose the following optimization problem:

$$\begin{aligned} \max_{\gamma,w,b} \quad & \gamma \\ \text{s.t.} \quad & y^{(i)}(w^\top \mathrm{x}^{(i)} + b) \geq \gamma, i = 1, \ldots, m \\ & ||w|| = 1. \end{aligned}$$

- We want to maximize γ, subject to each training example having functional margin at least γ.
- The $||w|| = 1$ constraint ensures that the functional margin equals to the geometric margin, so we are also guaranteed that all the geometric margins are at least γ.
- Thus, solving this problem will result in (w, b) with the largest possible geometric margin with respect to the training set.

Optimal Margin Classifier

- The $||w|| = 1$ constraint is a non-convex one, and this problem certainly isn't in any format that we can plug into standard optimization software to solve.

Optimal Margin Classifier

- The $||w|| = 1$ constraint is a non-convex one, and this problem certainly isn't in any format that we can plug into standard optimization software to solve.
- Lets transform the problem into a nicer one. Consider:

$$\max_{\gamma,w,b} \quad \frac{\hat{\gamma}}{||w||}$$

$$\text{s.t.} \quad y^{(i)}(w^\top \mathrm{x}^{(i)} + b) \geq \hat{\gamma}, i = 1, \ldots, m$$

Optimal Margin Classifier

- The $||w|| = 1$ constraint is a non-convex one, and this problem certainly isn't in any format that we can plug into standard optimization software to solve.
- Lets transform the problem into a nicer one. Consider:

$$\max_{\gamma,w,b} \quad \frac{\hat{\gamma}}{||w||}$$
$$\text{s.t.} \quad y^{(i)}(w^\top x^{(i)} + b) \geq \hat{\gamma}, i = 1, \ldots, m$$

- We're going to maximize $\hat{\gamma}/||w||$, subject to the functional margins all being at least $\hat{\gamma}$. Since the geometric and functional margins are related by $\gamma = \hat{\gamma}/||w||$, this will give us the answer we want.

Optimal Margin Classifier

- The $||w|| = 1$ constraint is a non-convex one, and this problem certainly isn't in any format that we can plug into standard optimization software to solve.
- Lets transform the problem into a nicer one. Consider:

$$\begin{aligned} \max_{\gamma,w,b} \quad & \frac{\hat{\gamma}}{||w||} \\ \text{s.t.} \quad & y^{(i)}(w^\top \mathrm{x}^{(i)} + b) \geq \hat{\gamma}, i = 1, \ldots, m \end{aligned}$$

- We're going to maximize $\hat{\gamma}/||w||$, subject to the functional margins all being at least $\hat{\gamma}$. Since the geometric and functional margins are related by $\gamma = \hat{\gamma}/||w||$, this will give us the answer we want.
- Moreover, we've gotten rid of the constraint $||w|| = 1$ that we didn't like.

Optimal Margin Classifier

- Since maximizing $\hat{\gamma}/||w|| = 1/||w||$ is the same as minimizing $||w||^2$, we can solve the following optimization problem:

$$\begin{aligned} \min_{\gamma,w,b} \quad & \frac{1}{2}||w||^2 \\ \text{s.t.} \quad & y^{(i)}(w^\top x^{(i)} + b) \geq 1, i = 1, \dots, m \end{aligned}$$

Optimal Margin Classifier

- Since maximizing $\hat{\gamma}/||w|| = 1/||w||$ is the same as minimizing $||w||^2$, we can solve the following optimization problem:

$$\min_{\gamma,w,b} \quad \frac{1}{2}||w||^2$$

$$\text{s.t.} \quad y^{(i)}(w^\top \mathrm{x}^{(i)} + b) \geq 1, i = 1, \ldots, m$$

- We've now transformed the problem into a form that can be efficiently solved. The above is an optimization problem with a convex quadratic objective and only linear constraints. Its solution gives us the optimal margin classifier.

Optimal Margin Classifier

- Since maximizing $\hat{\gamma}/||w|| = 1/||w||$ is the same as minimizing $||w||^2$, we can solve the following optimization problem:

$$\min_{\gamma,w,b} \quad \frac{1}{2}||w||^2$$
$$\text{s.t.} \quad y^{(i)}(w^\top x^{(i)} + b) \geq 1, i = 1, \ldots, m$$

- We've now transformed the problem into a form that can be efficiently solved. The above is an optimization problem with a convex quadratic objective and only linear constraints. Its solution gives us the optimal margin classifier.
- This optimization problem can be solved using commercial quadratic programming (QP) code.

Lagrange Multipliers

- Consider an optimization problem of the following form:

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & h_i(w) = 0, i = 1, \ldots, l. \end{aligned}$$

Lagrange Multipliers

- Consider an optimization problem of the following form:

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & h_i(w) = 0, i = 1, \ldots, l. \end{aligned}$$

- The method of Lagrange multipliers can be used to solve it. So we define the Lagrangian as

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

Lagrange Multipliers

- Consider an optimization problem of the following form:

$$\min_w \quad f(w)$$
$$\text{s.t.} \quad h_i(w) = 0, i = 1, \ldots, l.$$

- The method of Lagrange multipliers can be used to solve it. So we define the Lagrangian as

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

- The β_i's called the Lagrange multipliers.

Lagrange Multipliers

- So, given the Lagrangian

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

Lagrange Multipliers

- So, given the Lagrangian

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

- We would then find and set the $\mathcal{L}$'s partial derivatives to zero.

$$\frac{\partial \mathcal{L}}{\partial w_i} = 0$$

$$\frac{\partial \mathcal{L}}{\partial \beta_i} = 0$$

and solve for w and β.

Lagrange Multipliers

- Now, consider the following optimization problem:

$$\min_w \quad f(w)$$

$$\text{s.t.} \quad g_i \leq 0, i = 1, \ldots, k$$

$$h_i(w) = 0, i = 1, \ldots, l.$$

Lagrange Multipliers

- Now, consider the following optimization problem:

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

- The Lagrangian for this optimization problem can be defined as

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

Lagrange Multipliers

- So, given the Lagrangian

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

Lagrange Multipliers

- So, given the Lagrangian

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

- We would then find and set the $\mathcal{L}$'s partial derivatives to zero.

$\frac{\partial \mathcal{L}}{\partial w_i} = 0$

$\frac{\partial \mathcal{L}}{\partial \beta_i} = 0$

and solve for w and β.

Karush-Kuhn-Tucker conditions

There must exist w^*, α^*, β^* so that w^* is the solution to the optimization problem, and w^*, α^*, β^* satisfy the Karush-Kuhn-Tucker (KKT) conditions, which are as follows:

Karush-Kuhn-Tucker conditions

There must exist w^*, α^*, β^* so that w^* is the solution to the optimization problem, and w^*, α^*, β^* satisfy the Karush-Kuhn-Tucker (KKT) conditions, which are as follows:

$$\frac{\partial}{\partial w_i}\mathcal{L}(w^*, \alpha^*, \beta^*) \quad = \quad 0, i = 1, \ldots, n$$

Karush-Kuhn-Tucker conditions

There must exist w^*, α^*, β^* so that w^* is the solution to the optimization problem, and w^*, α^*, β^* satisfy the Karush-Kuhn-Tucker (KKT) conditions, which are as follows:

$$
\begin{aligned}
\frac{\partial}{\partial w_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, i = 1, \ldots, n \\
\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, i = 1, \ldots, l
\end{aligned}
$$

Karush-Kuhn-Tucker conditions

There must exist w^*, α^*, β^* so that w^* is the solution to the optimization problem, and w^*, α^*, β^* satisfy the Karush-Kuhn-Tucker (KKT) conditions, which are as follows:

$$
\begin{aligned}
\frac{\partial}{\partial w_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, i = 1, \ldots, n \\
\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, i = 1, \ldots, l \\
\alpha_i^* g_i(w^*) &= 0, i = 1, \ldots, k
\end{aligned}
$$

Karush-Kuhn-Tucker conditions

There must exist w^*, α^*, β^* so that w^* is the solution to the optimization problem, and w^*, α^*, β^* satisfy the Karush-Kuhn-Tucker (KKT) conditions, which are as follows:

$$\begin{aligned}
\frac{\partial}{\partial w_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, i = 1, \ldots, n \\
\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, i = 1, \ldots, l \\
\alpha_i^* g_i(w^*) &= 0, i = 1, \ldots, k \\
g_i(w^*) &\leq 0, i = 1, \ldots, k
\end{aligned}$$

Karush-Kuhn-Tucker conditions

There must exist w^*, α^*, β^* so that w^* is the solution to the optimization problem, and w^*, α^*, β^* satisfy the Karush-Kuhn-Tucker (KKT) conditions, which are as follows:

$$
\begin{aligned}
\frac{\partial}{\partial w_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, i = 1, \ldots, n \\
\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, i = 1, \ldots, l \\
\alpha_i^* g_i(w^*) &= 0, i = 1, \ldots, k \\
g_i(w^*) &\leq 0, i = 1, \ldots, k \\
\alpha^* &\geq 0, i = 1, \ldots, k
\end{aligned}
$$

Optimal Margin using Lagrange Method

Previously we posed the following optimization problem for finding the optimal margin classifier:

$$\min_{w,b} \quad \frac{1}{2}||w||^2$$
$$\text{s.t.} \quad y^{(i)}(w^\top \mathrm{x}^{(i)} + b) \geq 1, i = 1, \ldots, m$$

Optimal Margin using Lagrange Method

Previously we posed the following optimization problem for finding the optimal margin classifier:

$$\min_{w,b} \quad \frac{1}{2}||w||^2$$
$$\text{s.t.} \quad y^{(i)}(w^\top x^{(i)} + b) \geq 1, i = 1, \ldots, m$$

We can write the constraints as

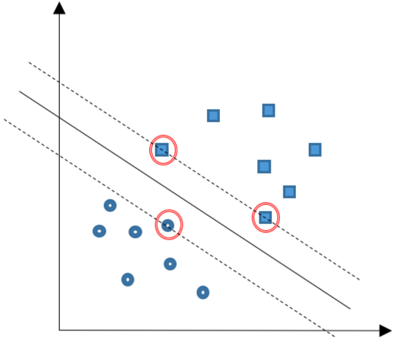
$$g_i(w) = -y^{(i)}(w^\top x^{(i)} + b) + 1 \leq 0.$$

Optimal Margin using Lagrange Method

From the KKT condition $\alpha_i g_i(w) = 0$ we have that $\alpha_i > 0$ only for the training examples that have a functional margin equal to one: the ones corresponding to constraints that hold with equality $g_i(w) = 0$.

Optimal Margin using Lagrange Method

From the KKT condition $\alpha_i g_i(w) = 0$ we have that $\alpha_i > 0$ only for the training examples that have a functional margin equal to one: the ones corresponding to constraints that hold with equality $g_i(w) = 0$.



Optimal Margin using Lagrange Method

Then, we can construct the Lagrangian as:

$$\mathcal{L}(\mathrm{w}, b, \alpha) = \frac{1}{2}||\mathrm{w}||^2 - \sum_{i=1}^{m} \alpha_i \Big[y^{(i)}(\mathrm{w}^\top \mathrm{x}^{(i)} + b) - 1 \Big].$$

Optimal Margin using Lagrange Method

Then, we can construct the Lagrangian as:

$$\mathcal{L}(\mathrm{w}, b, \alpha) = \frac{1}{2}||\mathrm{w}||^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)}(\mathrm{w}^\top \mathrm{x}^{(i)} + b) - 1 \right].$$

To solve this Lagrangian first we set the derivatives of $\mathcal{L}$ with respect to w and b to zero:

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

Optimal Margin using Lagrange Method

Then, we can construct the Lagrangian as:

$$\mathcal{L}(\mathrm{w}, b, \alpha) = \frac{1}{2}||\mathrm{w}||^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)}(\mathrm{w}^\top \mathrm{x}^{(i)} + b) - 1 \right].$$

To solve this Lagrangian first we set the derivatives of $\mathcal{L}$ with respect to w and b to zero:

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

which implies

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

Optimal Margin using Lagrange Method

As for the derivative of $\mathcal{L}$

$$\mathcal{L}(\mathrm{w}, b, \alpha) = \frac{1}{2}||\mathrm{w}||^2 - \sum_{i=1}^{m} \alpha_i \Big[y^{(i)}(\mathrm{w}^\top \mathrm{x}^{(i)} + b) - 1 \Big].$$

Optimal Margin using Lagrange Method

As for the derivative of $\mathcal{L}$

$$\mathcal{L}(\mathrm{w}, b, \alpha) = \frac{1}{2}||\mathrm{w}||^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)} (\mathrm{w}^\top \mathrm{x}^{(i)} + b) - 1 \right].$$

with respect to b we obtain:

$$\frac{\partial}{\partial b} \mathcal{L}(\mathrm{w}, b, \alpha) = \sum_{i=1}^{m} \alpha_i y^{(i)} = 0.$$

Optimal Margin using Lagrange Method

Now, replacing

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

in

$$\mathcal{L}(\mathrm{w}, b, \alpha) = \frac{1}{2}||\mathrm{w}||^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)}(\mathrm{w}^\top \mathrm{x}^{(i)} + b) - 1 \right].$$

Optimal Margin using Lagrange Method

Now, replacing

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

in

$$\mathcal{L}(\mathrm{w}, b, \alpha) = \frac{1}{2}||\mathrm{w}||^2 - \sum_{i=1}^{m} \alpha_i \Big[y^{(i)} (\mathrm{w}^\top \mathrm{x}^{(i)} + b) - 1 \Big].$$

and simplifying it, we get

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

Optimal Margin using Lagrange Method

And considering that

$$\frac{\partial}{\partial b}\mathcal{L}(\mathrm{w}, b, \alpha) = \sum_{i=1}^{m} \alpha_i y^{(i)} = 0.$$

the last term in

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

must be zero. Therefore we have:

Optimal Margin using Lagrange Method

And considering that

$$\frac{\partial}{\partial b}\mathcal{L}(\mathrm{w}, b, \alpha) = \sum_{i=1}^{m} \alpha_i y^{(i)} = 0.$$

the last term in

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

must be zero. Therefore we have:

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

Optimal Margin using Lagrange Method

So, by minimizing $\mathcal{L}(\mathrm{w}, b, \alpha)$ with respect to w and b we have:

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

Optimal Margin using Lagrange Method

So, by minimizing $\mathcal{L}(\mathrm{w}, b, \alpha)$ with respect to w and b we have:

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

which is now a just function of α:

$$\mathcal{W}(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle \mathrm{x}^{(i)}, \mathrm{x}^{(j)} \rangle.$$

Optimal Margin using Lagrange Method

So, by minimizing $\mathcal{L}(\mathrm{w}, b, \alpha)$ with respect to w and b we have:

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

which is now a just function of α:

$$\mathcal{W}(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \left\langle \mathrm{x}^{(i)}, \mathrm{x}^{(j)} \right\rangle.$$

where

$$\left\langle \mathrm{x}^{(i)}, \mathrm{x}^{(j)} \right\rangle$$

is a dot product.

Optimal Margin using Lagrange Method

Now we need to maximize $\mathcal{W}(\alpha)$ subject to some constraints:

$$\max_\alpha \mathcal{W}(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle \mathrm{x}^{(i)}, \mathrm{x}^{(j)} \rangle.$$

$$\text{s.t.} \quad \alpha_i \geq 0, \quad i = 1, \ldots, m$$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

Optimal Margin using Lagrange Method

Once, we have found the optimal w, b, α values, we can predict a new input x to be 1 if

$$\mathrm{w}^\top \mathrm{x} + b \geq 0.$$

Optimal Margin using Lagrange Method

Once, we have found the optimal w, b, α values, we can predict a new input x to be 1 if

$$\mathrm{w}^\top \mathrm{x} + b \geq 0.$$

However, knowing that $\mathrm{w} = \sum_{i=1}^{m} \alpha_i y^{(i)} \mathrm{x}^{(i)}$ we can make

$$\mathrm{w}^\top \mathrm{x} + b \quad = \left(\sum_{i=1}^{m} \alpha_i y^{(i)} \mathrm{x}^{(i)} \right)^\top \mathrm{x} + b$$

Optimal Margin using Lagrange Method

Once, we have found the optimal w, b, α values, we can predict a new input x to be 1 if

$$\mathrm{w}^\top \mathrm{x} + b \geq 0.$$

However, knowing that $\mathrm{w} = \sum_{i=1}^{m} \alpha_i y^{(i)} \mathrm{x}^{(i)}$ we can make

$$\begin{aligned} \mathrm{w}^\top \mathrm{x} + b &= \left(\sum_{i=1}^{m} \alpha_i y^{(i)} \mathrm{x}^{(i)} \right)^\top \mathrm{x} + b \\ &= \sum_{i=1}^{m} \alpha_i y^{(i)} \langle \mathrm{x}^{(i)}, \mathrm{x} \rangle + b. \end{aligned}$$

Reference

- Andrew Ng. **Machine Learning Course Notes**. 2003.
- Christopher Bishop. **Pattern Recognition and Machine Learning**. Springer. 2006.

Thank You!

Dr. Víctor Uc Cetina
victoruccetina@gmail.com