Support Vector Machines: Maximum Margin Classifiers

Machine Learning and Pattern Recognition:
September 23, 2010

Piotr Mirowski
Based on slides by Sumit Chopra, Fu-Jie Huang and Mehryar Mohri
Outline

- What is behind Support Vector Machines?
  - Constrained optimization
  - Lagrange constraints
  - “Dual” solution
- Support Vector Machines in detail
  - Kernel trick
  - LibSVM demo
Binary Classification Problem

- **Given:** Training data generated according to the distribution $D$

  \[
  (x_1, y_1), \ldots, (x_m, y_m) \in \mathbb{R}^n \times \{-1, 1\}
  \]

- **Problem:** Find a classifier (a function) $h(x): \mathbb{R}^n \rightarrow \{-1, 1\}$ such that it generalizes well on the test set obtained from the same distribution $D$

- **Solution:**
  - **Linear Approach:** linear classifiers (e.g. logistic regression, Perceptron)
  - **Non Linear Approach:** non-linear classifiers (e.g. Neural Networks, SVM)
Assume that the training data is linearly separable
Assume that the training data is linearly separable

\[ \mathbf{w} \cdot \mathbf{x} + b = 0 \]
Linearly Separable Data

Assume that the training data is linearly separable

\[ \vec{w} \cdot \vec{x} + b = 0 \]

Abscissa on axis parallel to \( \vec{w} \)

Abscissa of origin 0 is \( b \)

\[ \vec{w} \cdot \vec{x}_b + b = \tilde{y}_b \]

\[ \vec{w} \cdot \vec{x}_r + b = \tilde{y}_a \]

\[ \vec{w} \cdot \vec{O} + b = b \]
Assume that the training data is linearly separable

\[ \vec{w} \cdot \vec{x} + b = 0 \]

absissa on axis parallel to \( \vec{w} \)

absissa of origin \( 0 \) is \( b \)

Then the classifier is:

\[ h(x) = \vec{w} \cdot \vec{x} + b \quad \text{where} \quad w \in \mathbb{R}^n, b \in \mathbb{R} \]

Inference:

\[ \text{sign}(h(x)) \in \{-1, 1\} \]
Linearly Separable Data

- Assume that the training data is linearly separable

Margin \( \rho = \frac{1}{\|\vec{w}\|} \) (in the \( \{O, \vec{x}_1, \vec{x}_2\} \) space)

Maximize margin \( \rho \) (or \( 2\rho \)) so that:

For the closest points: \( h(x) = \vec{w} \cdot \vec{x} + b \in \{-1, 1\} \)
A Constrained Optimization Problem

\[ \min_w \frac{1}{2} \|w\|^2 \]

s.t.: \[ y_i(w \cdot x_i + b) \geq 1, \quad i = 1, \ldots, m \]

Equivalent to maximizing the margin \[ \rho = \frac{1}{\|w\|} \]

A convex optimization problem:
- Objective is convex
- Constraints are affine hence convex

Therefore, admits an unique optimum at \( w_0 \)
Optimization Problem

**Compare:**

\[
\min_w \frac{1}{2} \|w\|^2
\]

**s.t.:**

\[
y_i(w \cdot x_i + b) \geq 1, \quad i = 1, \ldots, m
\]

**With:**

\[
\min_w \left( \sum_{i=1}^{m} \left( -y_i(w \cdot x_i + b) + \frac{\lambda}{2} \|w\|^2 \right) \right)
\]

**objective**

**constraints**

**energy/errors**

**regularization**
Optimization: Some Theory

The problem:

\[ \min_{x} f_0(x) \]  

Objective function

\[ s.t.: \]

- \[ f_i(x) \leq 0, \quad i = 1, \ldots, m \]  
  Inequality constraints
- \[ h_i(x) = 0, \quad i = 1, \ldots, p \]  
  Equality constraints

Solution of problem:

- Global (unique) optimum – if the problem is convex
- Local optimum – if the problem is not convex

(Notation change: the parameters to optimize are noted \( x \))
Example: Standard Linear Program (LP)

\[
\min_{x} \ c^T x \\
\begin{align*}
\text{s.t.:} & \\
Ax &= b \\
x &\geq 0
\end{align*}
\]

Example: Least Squares Solution of Linear Equations (with L₂ norm regularization of the solution x)

i.e. Ridge Regression

\[
\min_{x} \ x^T x \\
\begin{align*}
\text{s.t.:} & \\
Ax &= b
\end{align*}
\]
Constrained / unconstrained optimization

Hierarchy of objective function:
smooth = infinitely derivable
convex = has a global optimum

\[ f_0 \]

- convex
- non-convex

- smooth
- non-smooth

SVM
NN
Introducing the concept of Lagrange function on a toy example
Toy Example: Equality Constraint

Example 1:

\[
\begin{align*}
&\text{min } x_1 + x_2 \quad \equiv f \\
&s.t.: \quad x_1^2 + x_2^2 - 2 = 0 \quad \equiv h_1
\end{align*}
\]

At Optimal Solution:

\[
\nabla f(x^o) = \lambda_1 \nabla h_1(x^o)
\]
Toy Example: Equality Constraint

- $x$ is not an optimal solution, if there exists $s \neq 0$ such that

$$h_1(x+s) = 0$$
$$f(x+s) < f(x)$$

- Using first order Taylor's expansion

$$h_1(x+s) = h_1(x) + \nabla h_1(x)^T s = \nabla h_1(x)^T s = 0 \quad (1)$$
$$f(x+s) - f(x) = \nabla f(x)^T s < 0 \quad (2)$$

- Such an $s$ can exist only when

$\nabla h_1(x)$ and $\nabla f(x)$ are not parallel
Thus we have

$$\nabla f(x^o) = \lambda_1 \nabla h_1(x^o)$$

The Lagrangian

$$L(x, \lambda_1) = f(x) - \lambda_1 h_1(x)$$

Thus at the solution

$$\nabla_x L(x^o, \lambda_1^o) = \nabla f(x^o) - \lambda_1^o \nabla h_1(x^o) = 0$$

This is just a necessary (not a sufficient) condition”

$x$ solution implies $\nabla h_1(x) \parallel \nabla f(x)$
Toy Example: Inequality Constraint

Example 2:

\[
\begin{align*}
\min & \quad x_1 + x_2 \\
\text{s.t.:} & \quad 2 - x_1^2 - x_2^2 \geq 0
\end{align*}
\]

\[
\begin{align*}
\nabla f &= \begin{pmatrix}
\frac{\partial f}{\partial x_1} \\
\frac{\partial f}{\partial x_2}
\end{pmatrix} \\
\nabla c_1 &= \begin{pmatrix}
\frac{\partial c_1}{\partial x_1} \\
\frac{\partial c_1}{\partial x_2}
\end{pmatrix}
\end{align*}
\]
**Toy Example:**

**Inequality Constraint**

$x$ is not an optimal solution, if there exists such that

$$c_1(x + s) \geq 0$$

$$f(x + s) < f(x)$$

Using first order Taylor's expansion

\[
c_1(x + s) = c_1(x) + \nabla c_1(x)^T s \geq 0 \quad (1)
\]

\[
f(x + s) - f(x) = \nabla f(x)^T s < 0 \quad (2)
\]
Toy Example: Inequality Constraint

**Case 1: Inactive constraint**
- Any sufficiently small $s$ as long as $\nabla f_1(x) \neq 0$
- Thus $s = -\alpha \nabla f(x)$ where $\alpha > 0$

**Case 2: Active constraint**
- $c_1(x) = 0$
  
  \[
  \nabla c_1(x)^T s \geq 0 \quad (1)
  \]
  \[
  \nabla f(x)^T s < 0 \quad (2)
  \]

In that case, $s = 0$ when:

$$\nabla f(x) = \lambda_1 \nabla c_1(x), \quad \text{where } \lambda_1 \geq 0$$
Thus we have the Lagrange function (as before)

\[ L(x, \lambda_1) = f(x) - \lambda_1 c_1(x) \]

The optimality conditions

\[ \nabla_x L(x^o, \lambda_1^o) = \nabla f(x^o) - \lambda_1^o \nabla c_1(x^o) = 0 \quad \text{for some} \quad \lambda_1^o \geq 0 \]

and

\[ \lambda_1^o c_1(x^o) = 0 \]

Complementarity condition

either \[ c_1(x^o) = 0 \] (active) or \[ \lambda_1^o = 0 \] (inactive)
Same Concepts in a More General Setting
Lagrange Function

The Problem

\[ \min_{x} f_{0}(x) \]

s.t.:
\[ \begin{align*}
  f_{i}(x) & \leq 0, & i = 1, \ldots, m \\
  h_{i}(x) & = 0, & i = 1, \ldots, p
\end{align*} \]

Standard tool for constrained optimization: the Lagrange Function

\[ L(x, \lambda, \nu) = f_{0}(x) + \sum_{i=1}^{m} \lambda_{i} f_{i}(x) + \sum_{i=1}^{p} \nu_{i} h_{i}(x) \]

dual variables or Lagrange multipliers
Lagrange Dual Function

Defined, for \( \lambda, \nu \) as the minimum value of the Lagrange function over \( x \)

\( m \) inequality constraints
\( p \) equality constraints

\[ g : \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R} \]

\[
g(\lambda, \nu) = \inf_{x \in D} L(x, \lambda, \nu) = \inf_{x \in D} \left( f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right)
\]
**Lagrange Dual Function**

- **Interpretation of Lagrange dual function:**
  - Writing the original problem as unconstrained problem but with hard indicators (penalties)

  \[
  \text{minimize } \begin{pmatrix} f_0(x) + \sum_{i=1}^{m} I_0(f_i(x)) + \sum_{i=1}^{p} I_1(h_i(x)) \end{pmatrix}
  \]

  where

  \[
  I_0(u) = \begin{cases} 
  0 & u \leq 0 \\
  \infty & u > 0
  \end{cases}
  \]

  \[
  I_1(u) = \begin{cases} 
  0 & u = 0 \\
  \infty & u \neq 0
  \end{cases}
  \]
Lagrange Dual Function

Interpretation of Lagrange dual function:
- The Lagrange multipliers in Lagrange dual function can be seen as “softer” version of indicator (penalty) functions.

\[
\begin{align*}
\text{minimize} \quad & \left( f_0(x) + \sum_{i=1}^{m} I_0(f_i(x)) + \sum_{i=1}^{p} I_1(h_i(x)) \right) \\
\text{inf} \quad & \left( f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right)
\end{align*}
\]
If \((x^o, \lambda^o, \nu^o)\) is a saddle point, i.e. if
\[
\forall x \in \mathbb{R}^n, \quad \forall \lambda \geq 0, \quad L(x^o, \lambda, \nu) \leq L(x^o, \lambda^o, \nu^o) \leq L(x, \lambda^o, \nu^o)
\]
... then \((x^o, \lambda^o, \nu^o)\) is a solution of the primal problem \(p^o\).
Lagrange Dual Problem

- Lagrange dual function gives a lower bound on the optimal value of the problem.
- We seek the “best” lower bound to minimize the objective:

\[
\begin{align*}
\text{maximize} & \quad g(\lambda, \nu) \\
\text{s.t.:} & \quad \lambda \geq 0
\end{align*}
\]

- The dual optimal value and solution:

\[d^o = g(\lambda^o, \nu^o)\]

- The Lagrange dual problem is convex even if the original problem is not.
Primal / Dual Problems

Primal problem:

\[
\begin{align*}
\text{min } & \quad f_0(x) \\
\text{s.t.:} & \quad f_i(x) \leq 0, \quad i = 1, \ldots, m \\
& \quad h_i(x) = 0, \quad i = 1, \ldots, p
\end{align*}
\]

Dual problem:

\[
\begin{align*}
\text{max } & \quad g(\lambda, \nu) \\
\text{s.t.:} & \quad \lambda \geq 0
\end{align*}
\]

\[
g(\lambda, \nu) = \inf_{x \in D} \left( f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right)
\]
Optimality Conditions: First Order

Karush-Kuhn-Tucker (KKT) conditions
If the strong duality holds, then at optimality:

\[
\begin{align*}
  f_i(x^o) &\leq 0, \quad i = 1, \ldots, m \\
  h_i(x^o) &= 0, \quad i = 1, \ldots, p \\
  \lambda^o_i &\geq 0, \quad i = 1, \ldots, m \\
  \lambda^o_i f_i(x^o) &= 0, \quad i = 1, \ldots, m \\
  \nabla f_0(x^o) + \sum_{i=1}^{m} \lambda^o_i \nabla f_i(x^o) + \sum_{i=1}^{p} \nu^o_i \nabla h_i(x^o) &= 0
\end{align*}
\]

KKT conditions are
\- necessary in general (local optimum)
\- necessary and sufficient in case of convex problems (global optimum)