SVM for Statisticians

Youyi Fong

Fred Hutchinson Cancer Research Institute

November 13, 2011
Primal Problem and Penalized Loss Function

- Minimize $J$ over $b$, $\beta$ and $\xi$ under some constraints
  \[
  J = \frac{1}{2} \| \beta \|^2 + C \sum_{i=1}^{n} \xi_i 
  \]
  \[
y_i (b + x_i \beta) \geq 1 - \xi_i 
  \]
  \[
  \xi_i \geq 0
  \]

- Rewrite (2) as
  \[
  \xi_i \geq 1 - y_i (b + x_i \beta)
  \]

  Combining with (3), we have
  \[
  \xi_i \geq \max (1 - y_i (b + x_i \beta), 0) = \{1 - y_i (b + x_i \beta)\}_+
  \]

  When $J$ is minimized, $\xi_i$ should be equal to its lower bound simply because we are minimizing $J$ over $b$, $\beta$ and $\xi$. Thus effectively we are minimizing
  \[
  J = C \sum_{i=1}^{n} \{1 - y_i (b + x_i \beta)\}_+ + \| \beta \|^2
  \]

- The penalized loss form is convenient mathematically, but inconvenient for optimization because of the hinge loss function $\max (\cdot, 0)$.
- Primal formulation is convenient for optimization through the use of slack variables $\xi_i$. 
Because the constraint (2) involves multiple parameters, it is difficult to handle. We dualize with respect to constraints (2) and (3). The primary Lagrangian is

\[ L_p = \frac{1}{2} \| \beta \|^2 + C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \{ y_i (b + x_i \beta) - 1 + \xi_i \} - \sum_{i=1}^{n} \gamma_i \xi_i. \]  

(5)

\[ L_p \] has to be minimized with respect to \( \beta, b \) and \( \xi_i \), and maximized with respect to non-negative Lagrange multipliers \( \alpha_i \) and \( \gamma_i \). Taking derivatives with respect to the primal space variables, we get

\[ \beta = \sum_{i=1}^{n} \alpha_i y_i x_i \]  

(6)

\[ \sum_{i=1}^{n} \alpha_i y_i = 0 \]  

(7)

\[ C = \alpha_i + \gamma_i \]  

(8)
Plugging (6)-(8) back into (5), we get the dual variables Lagrangian

\[
L_d = \frac{1}{2} \|\beta\|^2 + \sum_{i=1}^{n} (C - \alpha_i - \gamma_i) \xi_i - \sum_{i=1}^{n} \alpha_i y_i \left( b + x_i^T \beta \right) + \sum_{i=1}^{n} \alpha_i
\]

\[
= \frac{1}{2} \beta^T \beta - \sum_{i=1}^{n} \alpha_i y_i x_i^T \beta + \sum_{i=1}^{n} \alpha_i \quad \text{by (7) and (8)}
\]

\[
= \left( \frac{1}{2} \beta^T - \sum_{i=1}^{n} \alpha_i y_i x_i^T \right) \beta + \sum_{i=1}^{n} \alpha_i
\]

\[
= -\frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j x_i^T x_j + \sum_{i=1}^{n} \alpha_i \quad \text{by (6)}
\]

We are to maximize \( L_d \) under two constraints that come from (7) and (8) and with the non-negativity of Lagrange multipliers

\[
\sum_{i=1}^{n} \alpha_i y_i = 0 \quad (9)
\]

\[
0 \leq \alpha_i \leq C \quad (10)
\]

Comparing this set of constraints to the contraints of the primal problem, (2), we see that they involve one variable at a time. This allows for simple decomposition algorithms to work
KKT Conditions

- Stationarity
- Primal and dual feasibility
- Complementary slackness

\[ \alpha_i \{ y_i (b + x_i \beta) - 1 + \xi_i \} = 0 \]
\[ (C - \alpha_i) \xi_i = 0 \]
Intercept-free Model

- If we don’t want $b$ to be in the model, as in AUC work, we can set $b = 0$. As a result, the constraint (7)/(9) do not apply.
- $b$ is also known as the bias term.
Weighted SVM

- Suppose we want to maximize
  \[ J = C \sum_{i=1}^{n} w_i \{1 - y_i (b + x_i \beta)\} + \|\beta\|^2, \]

- The primary Lagrangian becomes
  \[ L_p = \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^{n} w_i \xi_i - \sum_{i=1}^{n} \alpha_i \{y_i (b + x_i \beta) - 1 + \xi_i\} - \sum_{i=1}^{n} \gamma_i \xi_i \]

  Constraint (8) becomes
  \[ C \omega_i = \alpha_i + \gamma_i \]

- Dual variables Lagrangian becomes
  \[ L_d = \frac{1}{2} \|\beta\|^2 + \sum_{i=1}^{n} (C \omega_i - \alpha_i - \gamma_i) \xi_i - \sum_{i=1}^{n} \alpha_i y_i \{b + x_i \beta\} + \sum_{i=1}^{n} \alpha_i \]
  \[ = -\frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j x_i^T x_j + \sum_{i=1}^{n} \alpha_i \quad \text{by (6)} \]

  under the constraint
  \[ \sum_{i=1}^{n} \alpha_i y_i = 0 \]
  \[ 0 \leq \alpha_i \leq C \omega_i \]

- The KKT complementary conditions are
  \[ \alpha_i \{y_i (b + x_i \beta) - 1 + \xi_i\} = 0 \]
  \[ (w_i C - \alpha_i) \xi_i = 0 \]
A decomposition algorithm chooses a subset of variables to optimize at each iteration.

- **Working set size and selection strategy**
  - SVMlight defaults to 10
    - according to steepest gradient, while satisfying all constraints (Joachims 1998)
    - once in a while, select a somewhat random working set to escape ‘dead zone’ (svmlight code)
  - libsvm only supports 2. First var is chosen based on gradient, and the second var is chosen based on second order information (Fan et al 2005)
  - Burges (1998) mentions conjugate gradient
  - In our experience, set size of 2 works better than set size of 1 (as illustrated on the next slide)
Decomposition Algorithm (cont’d)

- Other heuristics/tricks
  - Shrinking (Joachims, 1998). Only choose working set from a subset of total variables
  - Caching. This happens at several levels.
  - Burges (1998)

- To optimize a working set is to solve a constrained quadratic problem. Many optimizers can be used.
  - SVMlight uses Hideo’s optimizer
  - minQuad (explained in the next few slides)
SVM software

- svmlight
  - main advantage is the subproblem is not restricted to two variables
  - implements null bias/intercept
  - klaR R package

- libsvm
  - handles more types of svm models (but none of the extended model solves $\alpha^T Q \alpha + b^T \alpha$, where $b \neq 1$
  - does not implement null bias/intercept
  - e1071 R package

- svmw

- aucm
One-class SVM

Usual SVM

\[
\begin{align*}
\text{minimize} & & \| w \|^2 + C \sum_{i=1}^{n} \xi_i \\
\text{subject to} & & y_i (\langle \Phi (x_i), w \rangle + b) \geq 1 - \xi_i \\
& & \xi_i \geq 0
\end{align*}
\]

One-class SVM

minimizes \[
\| w \|^2 - \rho + C \sum_{i=1}^{n} \xi_i
\]
subject to \[
y_i (\langle \Phi (x_i), w \rangle + b) \geq \rho - \xi_i \\
\xi_i \geq 0
\]
Acknowledgement

- Shuxin Yin
- Krisztian Sebestyen


