A Bahadur Type Representation of the Linear Support Vector Machine and its Relative Efficiency

Yoonkyung Lee
Department of Statistics
The Ohio State University

June 8, 2009
Machine Learning Summer School/Workshop
University of Chicago
Outline

- Support Vector Machines
- Statistical Properties of SVM
- Main Results
  (asymptotic analysis of the linear SVM)
- An Illustrative Example
- Asymptotic Relative Efficiency
- Discussion
Applications

- Handwritten digit recognition
- Cancer diagnosis with microarray data
- Text categorization
Classification

- \( \mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d \)
- \( y \in \mathcal{Y} = \{1, \ldots, k\} \)
- Learn a rule \( \phi : \mathbb{R}^d \rightarrow \mathcal{Y} \) from the training data \( \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\} \), where \((\mathbf{x}_i, y_i)\) are i.i.d. with \( P(X, Y) \).
- The 0-1 loss function:

\[
\mathcal{L}(y, \phi(\mathbf{x})) = l(y \neq \phi(\mathbf{x}))
\]
Methods of Regularization (Penalization)

Find $f(x) \in F$ minimizing

$$\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i)) + \lambda J(f).$$

- Empirical risk + penalty
- $\mathcal{F}$: a class of candidate functions
- $J(f)$: complexity of the model $f$
- $\lambda > 0$: a regularization parameter
- Without the penalty $J(f)$, ill-posed problem
Maximum Margin Hyperplane

Margin = \frac{2}{||w||}
Support Vector Machines

Boser, Guyon, and Vapnik (1992)
Vapnik (1995), *The Nature of Statistical Learning Theory*
Schölkopf and Smola (2002), *Learning with Kernels*

- $y_i \in \{-1, 1\}$, class labels in the binary case
- Find $f \in \mathcal{F} = \{ f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b \mid \mathbf{w} \in \mathbb{R}^d \text{ and } b \in \mathbb{R} \}$ minimizing

$$\frac{1}{n} \sum_{i=1}^{n} (1 - y_i f(\mathbf{x}_i))_+ + \lambda \| \mathbf{w} \|^2,$$

where $\lambda$ is a regularization parameter.
- Classification rule : $\phi(\mathbf{x}) = \text{sign}(f(\mathbf{x}))$
(1 − yf(x))_+ is an upper bound of the misclassification loss function $l(y \neq \phi(x)) = [-yf(x)]_* \leq (1 − yf(x))_+$ where $[t]_* = l(t \geq 0)$ and $(t)_+ = \max\{t, 0\}$. 

Hinge Loss
SVM in General

Find \( f(x) = b + h(x) \) with \( h \in \mathcal{H}_K \) (RKHS) minimizing

\[
\frac{1}{n} \sum_{i=1}^{n} (1 - y_i f(x_i))_+ + \lambda \|h\|_{\mathcal{H}_K}^2.
\]

- **Linear SVM:**
  \[
  \mathcal{H}_K = \{ h(x) = w^\top x \mid w \in \mathbb{R}^d \} \text{ with }
  \begin{align*}
  &i) \ K(x, x') = x^\top x' \\
  &ii) \ \|h\|_{\mathcal{H}_K}^2 = \|w^\top x\|_{\mathcal{H}_K}^2 = \|w\|^2
  \end{align*}
  
- **Nonlinear SVM:** \( K(x, x') = (1 + x^\top x')^d, \)
  \[
  \exp(-\|x - x'\|^2/2\sigma^2), \ldots
  \]
Statistical Properties of SVM

- Fisher consistency (Lin, DM & KD 2002)
  \[ \arg_f \min E[(1 - Yf(X))^+ | X = x] = \text{sign}(p(x) - 1/2) \]
  where \( p(x) = P(Y = 1 | X = x) \)

- SVM approximates the Bayes decision rule
  \[ \phi_B(x) = \text{sign}(p(x) - 1/2). \]

  \[ R(\hat{f}_{SVM}) \to R(\phi_B) \text{ in prob.} \]
  under universal approximation condition on \( \mathcal{H}_K \)

- Rate of convergence (Steinwart et al., AOS 2007)
Main Questions

- Recursive Feature Elimination (Guyon et al., ML 2002): backward elimination of variables based on the fitted coefficients of the linear SVM

- What is the statistical behavior of the coefficients?
- What determines their variances?
- Study asymptotic properties of the coefficients of the linear SVM.
- What about its relative efficiency?
Something New, Old, and Borrowed

- The hinge loss:
  not everywhere differentiable,
  no closed form expression for the solution

- Useful link:
  $\text{sign}(p(x) - 1/2)$, the population minimizer w.r.t. the hinge loss is the median of $Y$ at $x$.

- Asymptotics for least absolute deviation (LAD) estimators (Pollard, ET 1991)

- Convexity of the loss
Preliminaries

- $(X, Y)$: a pair of random variables with $X \in \mathcal{X} \subset \mathbb{R}^d$ and $Y \in \{1, -1\}$
- $P(Y = 1) = \pi_+$ and $P(Y = -1) = \pi_-$
- Let $f$ and $g$ be the densities of $X$ given $Y = 1$ and $-1$.
- With $\tilde{x} = (1, x_1, \ldots, x_d)^\top$ and $\beta = (b, w^\top)^\top$, 
  \[ h(x; \beta) = w^\top x + b = \tilde{x}^\top \beta \]
- $\hat{\beta}_{\lambda, n} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (1 - y_i h(x_i; \beta))_+ + \lambda \|w\|^2$
Population Version

- \( L(\beta) = \mathbb{E}[1 - Yh(X; \beta)]_+ \)
- \( \beta^* = \arg \min_{\beta} L(\beta) \)
- The gradient of \( L(\beta) \):
  \[
  S(\beta) = -\mathbb{E} \left( \psi(1 - Yh(X; \beta)) Y \tilde{X} \right)
  \]
  where \( \psi(t) = I(t \geq 0) \)
- The Hessian matrix of \( L(\beta) \):
  \[
  H(\beta) = \mathbb{E} \left( \delta(1 - Yh(X; \beta)) \tilde{X} \tilde{X}^\top \right)
  \]
  where \( \delta \) is the Dirac delta function.
More on $H(\beta)$

- The Hessian matrix of $L(\beta)$:

$$H(\beta) = \mathbb{E}\left(\delta(1 - Yh(X; \beta))\tilde{X}\tilde{X}^\top\right)$$

$$H_{j,k}(\beta) = \mathbb{E}\left(\delta(1 - Yh(X; \beta))X_jX_k\right) \text{ for } 0 \leq j, k \leq d$$

$$= \pi_+ \int_X \delta(1 - b - w^\top x)x_jx_k f(x) dx$$

$$+ \pi_- \int_X \delta(1 + b + w^\top x)x_jx_k g(x) dx.$$  

- For a function $s$ on $\mathcal{X}$, define the Radon transform $\mathcal{R}s$ of $s$ for $p \in \mathbb{R}$ and $\xi \in \mathbb{R}^d$ as

$$(\mathcal{R}s)(p, \xi) = \int_\mathcal{X} \delta(p - \xi^\top x)s(x) dx.$$  

(the integral of $s$ over hyperplanes $\xi^\top x = p$)

- $H_{j,k}(\beta) = \pi_+(\mathcal{R}f_{j,k})(1 - b, w) + \pi_-(\mathcal{R}g_{j,k})(1 + b, -w)$, where $f_{j,k}(x) = x_jx_k f(x)$ and $g_{j,k}(x) = x_jx_k g(x)$.  

Regularity Conditions

(A1) The densities $f$ and $g$ are continuous and have finite second moments.

(A2) There exists $B(x_0, \delta_0)$ such that $f(x) > C_1$ and $g(x) > C_1$ for every $x \in B(x_0, \delta_0)$.

(A3) For some $1 \leq i^* \leq d$,

$$\int_{\mathcal{X}} \{ x_{i^*} \geq G_{i^*}^{-} \} x_{i^*} g(x) \, dx < \int_{\mathcal{X}} \{ x_{i^*} \leq F_{i^*}^{+} \} x_{i^*} f(x) \, dx$$

or

$$\int_{\mathcal{X}} \{ x_{i^*} \leq G_{i^*}^{+} \} x_{i^*} g(x) \, dx > \int_{\mathcal{X}} \{ x_{i^*} \geq F_{i^*}^{-} \} x_{i^*} f(x) \, dx.$$  

(when $\pi_+ = \pi_-$, it says that the means are different.)

(A4) Let $M^+ = \{ x \in \mathcal{X} \mid \tilde{x}^\top \beta^* = 1 \}$ and $M^- = \{ x \in \mathcal{X} \mid \tilde{x}^\top \beta^* = -1 \}$. There exist two subsets of $M^+$ and $M^-$ on which the class densities $f$ and $g$ are bounded away from zero.
Bahadur Representation

- Bahadur (1966), *A Note on Quantiles in Large Samples*
- A statistical estimator is approximated by a sum of independent variables with a higher-order remainder.
- Let $\xi = F^{-1}(p)$ be the $p$th quantile of distribution $F$. For $X_1, \ldots, X_n \sim iid F$, the sample $p$th quantile is

$$\xi + \left[ \sum_{i=1}^{n} I(X_i > \xi) - n(1 - p) \right] / nf(\xi) + R_n,$$

where $f(x) = F'(x)$. 
Bahadur-type Representation of the Linear SVM

**Theorem**
*Suppose that (A1)-(A4) are met. For \( \lambda = o(n^{-1/2}) \), we have*

\[
\sqrt{n} (\hat{\beta}_\lambda, n - \beta^*) = -\frac{1}{\sqrt{n}} H(\beta^*)^{-1} \sum_{i=1}^{n} \psi(1 - Y_i h(X_i; \beta^*)) Y_i \tilde{X}_i + o_P(1).
\]

**Recall** that \( H(\beta^*) = \mathbb{E}\left( \delta(1 - Y h(X; \beta^*)) \tilde{X} \tilde{X}^\top \right) \) and \( \psi(t) = I(t \geq 0) \).
Asymptotic Normality of $\hat{\beta}_{\lambda,n}$

**Theorem**

Suppose (A1)-(A4) are satisfied. For $\lambda = o(n^{-1/2})$,

$$
\sqrt{n} \left( \hat{\beta}_{\lambda,n} - \beta^* \right) \to N \left( 0, H(\beta^*)^{-1} G(\beta^*) H(\beta^*)^{-1} \right)
$$

in distribution, where

$$
G(\beta) = \mathbb{E} \left( \psi(1 - Yh(X; \beta)) \tilde{X} \tilde{X}^\top \right).
$$

**Corollary**

Under the same conditions as in Theorem,

$$
\sqrt{n} \left( h(x; \hat{\beta}_{\lambda,n}) - h(x; \beta^*) \right) \to N \left( 0, \tilde{x}^\top H(\beta^*)^{-1} G(\beta^*) H(\beta^*)^{-1} \tilde{x} \right)
$$

in distribution.
An Illustrative Example

- Two multivariate normal distributions in $\mathbb{R}^d$ with mean vectors $\mu_f$ and $\mu_g$ and a common covariance matrix $\Sigma$
- $\pi_+ = \pi_- = 1/2$.
- What is the relation between the Bayes decision boundary and the optimal hyperplane by the SVM, $h(x; \beta^*) = 0$?
Example

- The Bayes decision boundary (Fisher’s LDA):
  \[
  \left\{ \Sigma^{-1} (\mu_f - \mu_g) \right\}^\top \left\{ x - \frac{1}{2} (\mu_f + \mu_g) \right\} = 0.
  \]

- The hyperplane determined by the SVM:
  \[
  \tilde{x}^\top \beta^* = 0 \text{ with } S(\beta^*) = 0.
  \]

- \( \beta^* \) balances two classes within the margin
  \[
  \mathbb{E}\left( \psi(1 - Y h(X; \beta^*)) Y \tilde{X} \right) = 0
  \]

\[
\begin{align*}
P(h(X; \beta^*) \leq 1 | Y = 1) &= P(h(X; \beta^*) \geq -1 | Y = -1) \\
\mathbb{E}\left( I\{ h(X; \beta^*) \leq 1 \} X_j | Y = 1 \right) &= \mathbb{E}\left( I\{ h(X; \beta^*) \geq -1 \} X_j | Y = -1 \right)
\end{align*}
\]
Example

Direct calculation shows that
\[ \beta^* = C(d_\Sigma(\mu_f, \mu_g)) \left[ -\frac{1}{2}(\mu_f + \mu_g)^\top \right] \Sigma^{-1}(\mu_f - \mu_g), \]
where \( d_\Sigma(\mu_f, \mu_g) \) is the Mahalanobis distance between the two distributions.

The linear SVM is equivalent to Fisher’s LDA asymptotically in this case.

The assumptions (A1)-(A4) are satisfied. So, the main theorem applies.

Consider \( d = 1, \mu_f + \mu_g = 0, \sigma = 1, \) and \( d_\Sigma(\mu_f, \mu_g) = |\mu_f - \mu_g|. \)
Distance and Margin

\[ d=0.5 \]

\[ d=1 \]

\[ d=3 \]

\[ d=6 \]
Figure: The asymptotic variabilities of the intercept and the slope for the optimal hyperplane as a function of the Mahalanobis distance.
A Bivariate Normal Example

- \( \mu_f = (1, 1)^\top, \mu_g = (-1, -1)^\top \) and \( \Sigma = \text{diag}(1, 1) \)
- \( d_\Sigma(\mu_f, \mu_g) = 2\sqrt{2} \) and the Bayes error rate is 0.07865.
- Find \( (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (1 - y_i \tilde{x}_i^\top \beta)_+ \)

<table>
<thead>
<tr>
<th>Estimates</th>
<th>Sample size n</th>
<th>Optimal coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>( \hat{\beta}_0 )</td>
<td>0.0006</td>
<td>-0.0013</td>
</tr>
<tr>
<td>( \hat{\beta}_1 )</td>
<td>0.7709</td>
<td>0.7450</td>
</tr>
<tr>
<td>( \hat{\beta}_2 )</td>
<td>0.7749</td>
<td>0.7459</td>
</tr>
</tbody>
</table>

Table: Averages of estimated optimal coefficients over 1000 replicates.
Sampling Distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$

Figure: Estimated sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ ($n = 500$)
Figure: The median values of the type I error rates in variable selection when $\mu_f = (1_{d/2}, 0_{d/2})^\top$, $\mu_g = 0_d^\top$, and $\Sigma = I_d$
The canonical LDA setting:
\[ X \sim N((\Delta/2)e_1, I) \] for \( Y = 1 \) with probability \( \pi_+ \)
\[ X \sim N(-(\Delta/2)e_1, I) \] for \( Y = -1 \) with probability \( \pi_- \)

Fisher’s linear discriminant function is
\[ \ell(x) = \beta^\top x, \]
where \( \beta_{*0} = \log(\pi_+/\pi_-) \) and \( (\beta_{*1}, \ldots, \beta_{*d})^\top = \Delta e_1 \).

For a linear discriminant method \( \hat{\ell} \) with coefficient vector \( \hat{\beta} \), if \( \sqrt{n}(\hat{\beta} - \beta_*) \rightarrow N(0, \Sigma), E(R(\hat{\ell}) - R(\phi_B)) \) is given by
\[ \frac{\pi_+\phi(D_1)}{2\Delta n} \left[ \sigma_{00} - \frac{2\beta_{*0}}{\Delta} \sigma_{01} + \frac{\beta_{*0}^2}{\Delta^2} \sigma_{11} + \sigma_{22} + \cdots + \sigma_{dd} \right], \]
where \( D_1 = \Delta/2 + (1/\Delta) \log(\pi_+/\pi_-). \)
Relative Efficiency

- Efron (1975) studied the Asymptotic Relative Efficiency (ARE) of logistic regression (LR) to normal discrimination (LDA) defined as

\[
\lim_{n \to \infty} \frac{E(R(\hat{\ell}_{LDA}) - R(\phi_B))}{E(R(\hat{\ell}_{LR}) - R(\phi_B))}.
\]

- Logistic regression is shown to be between one half and two thirds as effective as normal discrimination typically.
Asymptotic Relative Efficiencies of SVM to LDA

Under the canonical LDA setting with $\pi_+ = \pi_- = 0.5$, the ARE of the linear SVM to LDA is given by

$$\text{Eff} = \frac{2}{\Delta} (1 + \frac{\Delta^2}{4}) \phi(a^*),$$

where $a^*$ is the constant satisfying $\phi(a^*)/\Phi(a^*) = \Delta/2$.

<table>
<thead>
<tr>
<th>$\Delta$</th>
<th>$R(\phi_B)$</th>
<th>$a^*$</th>
<th>SVM</th>
<th>LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>0.1587</td>
<td>-0.3026</td>
<td>0.7622</td>
<td>0.899</td>
</tr>
<tr>
<td>2.5</td>
<td>0.1056</td>
<td>-0.6466</td>
<td>0.6636</td>
<td>0.786</td>
</tr>
<tr>
<td>3.0</td>
<td>0.0668</td>
<td>-0.9685</td>
<td>0.5408</td>
<td>0.641</td>
</tr>
<tr>
<td>3.5</td>
<td>0.0401</td>
<td>-1.2756</td>
<td>0.4105</td>
<td>0.486</td>
</tr>
<tr>
<td>4.0</td>
<td>0.0228</td>
<td>-1.5718</td>
<td>0.2899</td>
<td>0.343</td>
</tr>
</tbody>
</table>
A Mixture of Two Gaussian Distributions

Figure: $\Delta_W$ and $\Delta_B$ indicate the mean difference between two Gaussian components within each class and the mean difference between two classes.
As $\Delta_W$ Varies

Figure: $\Delta_B = 2$, $d = 5$, $\pi_+ = \pi_-$, $\pi_1 = \pi_2$, and $n = 100$
As $\Delta B$ Varies

Error Rate

$\Delta W = 1, \ d = 5, \ \pi_+ = \pi_-, \ \pi_1 = \pi_2, \ \text{and} \ n = 100$
As Dimension $d$ Varies

$\Delta_W = 1$, $\Delta_B = 2$, $\pi_+ = \pi_-$, $\pi_1 = \pi_2$, and $n = 100$
Concluding Remarks

- Examine asymptotic properties of the coefficients of variables in the linear SVM and study its relative efficiency.
- Establish Bahadur type representation of the coefficients.
- How the margins of the optimal hyperplane and the underlying probability distribution characterize their statistical behavior.
- Variable selection for the SVM in the framework of hypothesis testing.
- For practical applications, need consistent estimators of $G(\beta^*)$ and $H(\beta^*)$.
- Explore a different scenario where $d$ also grows with $n$.
- Extension of the SVM asymptotics to the nonlinear case.
Reference


  Available at www.stat.osu.edu/~yklee or http://www.jmlr.org/.

- Relative efficiency analysis and related results are from joint work with Rui Wang (in progress).