Boosting and Support Vector Machines

Trevor Hastie
Statistics Department
Stanford University

Collaborators: Jerome Friedman, Saharon Rosset, Rob Tibshirani, Ji Zhu

Outline

- SVMs and regularized optimization
- Regularized logistic regression
- Idealized boosting and the lasso
- Connections

First part based on work by Vapnik (1996), Wahba (1990), Evgeniou, Pontil, and Poggio (1999); described in Hastie, Tibshirani and Friedman (2001) *Elements of Statistical Learning*, Springer, NY. Gunnar Rätsch and coworkers have also made connection between SVMs and Boosting.
Maximum Margin Classifier

Vapnik (1995)

\[ x_i \in \mathbb{R}^p, \ y_i \in \{-1, 1\} \]

\[ x^T \beta + \beta_0 = 0 \]

\[
\max_{\beta, \beta_0, \|\beta\|=1} C \\
\text{subject to} \quad y_i(x_i^T \beta + \beta_0) \geq C, \ i = 1, \ldots, N.
\]
- Modifications allow nonseparable classes
- Fit more general functions $f(x) = h(x)^T \beta + \beta_0$ through basis expansions and regularization.
- Exploits the “kernel trick” — a computational shortcut known in the RKHS literature; $h(x)$ is generated by kernel $K(x, x')$. 
With \( f(x) = h(x)^T \beta + \beta_0 \) and \( y_i \in \{-1, 1\} \), SVM solves penalized “hinge loss” problem

\[
\min_{\beta_0, \beta} \sum_{i=1}^{N} \left[ 1 - y_i f(x_i) \right]_+ + \lambda \|\beta\|^2
\]

\[
\min_{\beta_0, \beta} \sum_{i=1}^{N} L[y_i, f(x_i)] + \lambda \|\beta\|^2
\]

- Fig. shows \( L[y, f(x)] = \log (1 + e^{-yf(x)}) \), binomial log-likelihood.
- In separable case, both yield opt. separating hyperplane as \( \lambda \downarrow 0 \).
- Hinge loss estimates \( f(x) = \text{sign}[P(Y = 1|x) - \frac{1}{2}] \)
- Binomial log-likelihood estimates \( f(x) = \text{logit}[P(Y = 1|x)] \)
Final Classifier

\[ G(x) = \text{sign} \left[ \sum_{m=1}^{M} \alpha_m G_m(x) \right] \]

Classifiers are trained on weighted versions of the dataset, and then combined to produce a final prediction.
AdaBoost (Freund & Schapire, 1996)

- Start with weights $w_i = 1/N \forall i = 1, \ldots , N$. $y_i \in \{-1, 1\}$.
- Repeat for $m = 1, 2, \ldots , M$:
  - Estimate the weak learner $f_m(x) \in \{-1, 1\}$ from the training data with weights $w_i$.
  - Compute $e_m = E_w[1(y \neq f_m(x))]$, $c_m = \log((1 - e_m)/e_m)$.
  - Set $w_i \leftarrow w_i \exp[c_m \cdot 1(y_i \neq f_m(x_i))]$, $i = 1, 2, \ldots N$, and renormalize so that $\sum_i w_i = 1$.
- Output the majority weight classifier $C(x) = \text{sign}[\sum_{m=1}^{M} c_m f_m(x)]$. 
Boosting builds a sequence of models \( f_J(x) = \sum_{j=1}^{J} g_j(x) \), where each \( g_j(x) \) is a “weak” classifier fit to weighted training data.

Even though at stage \( J \), \( f_J(x) \) may have zero training errors, boosting increases the “margin”, \( yf(x) \).

Actually, boosting is fitting the model
\[
f(x) = \log \frac{\Pr(Y = 1 | x)}{\Pr(Y = -1 | x)}
\]
by stagewise optimization of the loss function
\[
L[Y, f(X)] = \exp[-Yf(X)]
\]
Boosting and $L_1$ Penalized Fitting

In a restricted setting where

- the base learners are chosen from a fixed set of basis functions;
- the increments at each boosting step are shrunk towards zero;
- + a few mild assumptions (yeah, right!),

the boosting sequence corresponds to a sequence (as $\lambda$ varies) of solutions to the $L_1$ penalized optimization problem

$$
\min_{\beta} \sum_{i=1}^{N} L[y_i, f(x_i)] + \lambda \|\beta\|_1
$$

where $L[Y, f(X)] = \exp[-Yf(X)]$.

- As $\lambda \downarrow 0$, $\hat{\beta} \to \beta^*$, the $L_1$ optimal margin separator.
Details: $\epsilon$ Boosting

$\epsilon$ Forward Stagewise

- Given a family of basis functions $h_1(x), \ldots, h_M(x)$, and loss function $L$.

- Model at $k$th step is $F_k(x) = \sum_m \beta^k_m h_m(x)$.

- At step $k+1$, identify coordinate $m$ with largest $|\partial L/\partial \beta_m|$, and update $\beta^{k+1}_m \leftarrow \beta^k_m - \epsilon \cdot \text{sign}(\partial L/\partial \beta_m)$.

- Approximately equivalent to the lasso: $\min L(\beta) + \lambda_k \| \beta \|_1$

- As $\lambda_k \downarrow 0$, $\beta^k \rightarrow \beta^*$, the $L_1$ optimal margin separator. This is true for the exponential loss or the binomial log-likelihood.

- Typically a useful solution is found for some $\lambda > 0$, or equivalently somewhere along the boosted path.
Example and Illustration

Lasso

Forward Stagewise

\[ t = \sum_k |\alpha_k| \]

Coefficients vs. Iteration
Summary

- SVM can be viewed as regularized fitting with a particular loss function: hinge loss.

- Regularized logistic regression gives very similar fit, with added benefits. Also approaches a separating hyperplane. Uses binomial deviance as loss.

- Boosting can be viewed as $L_1$ regularized fitting (exponential or binomial loss); has optimal margin limiting behavior.

- In none of the cases above is it clear that we should go all the way towards the maximal margin solution. The regularization parameter $\lambda$ (equivalently stopping time in boosting) should be determined by, e.g., cross-validation.