Chapter 8. Support Vector Machines

Wei Pan

Division of Biostatistics, School of Public Health, University of Minnesota,
Minneapolis, MN 55455
Email: weip@biostat.umn.edu

PubH 7475/8475

Introduction

- SVM: §4.5.2, 12.1-12.3; by Vapnik (1996).
- Training data: (Y_i, X_i), $Y_i = \pm 1$, $i = 1, ..., n$.
- Fig 4.14: with two separable classes, many possible separating hyperplanes, e.g. ,
  LSE (or LDA): 1 error;
  Perceptron: diff starting values;
  SVC: max the "separation" b/w two classes; Fig 4.16.

Elements of Statistical Learning (2nd Ed.) ©Hastie, Tibshirani & Friedman 2009 Chap 4

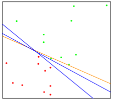


FIGURE 4.14. *A toy example with two classes separable by a hyperplane. The orange line is the least squares solution, which misclassifies one of the training points. Also shown are two blue separating hyperplanes found by the* perceptron learning algorithm *with different random starts.*

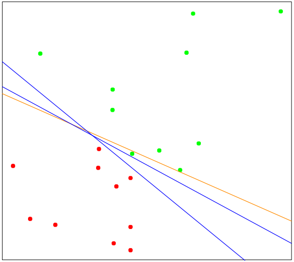


FIGURE 4.14. *A toy example with two classes separable by a hyperplane. The orange line is the least squares solution, which misclassifies one of the training points. Also shown are two blue separating hyperplanes found by the* perceptron learning algorithm *with different random starts.*

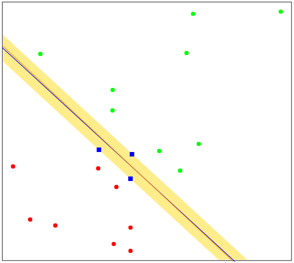
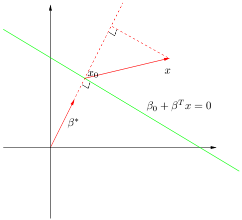


FIGURE 4.16. *The same data as in Figure 4.14. The shaded region delineates the maximum margin separating the two classes. There are three support points indicated, which lie on the boundary of the margin, and the optimal separating hyperplane (blue line) bisects the slab. Included in the figure is the boundary found using logistic regression (red line), which is very close to the optimal separating hyperplane (see Section 12.3.3).*

Review

- Hyperplane L: $f(x) = \beta_0 + \beta' x = 0$.
- 1) Any x_1, $x_2 \in L \Longrightarrow \beta'(x_1 - x_2) = 0$.
  $\beta \perp L$;
  $\beta^* = \beta / ||\beta||$, vector normal to L.
- 2) $x_0 \in L \Longrightarrow \beta_0 + \beta' x_0 = 0$.
- 3) The signed distance of any x to L is:
  $\beta^{*\prime}(x - x_0) = (\beta' x - \beta' x_0)/||\beta|| = (\beta' x + \beta_0)/||\beta||$.
  $\Longrightarrow f(x)/||\beta|| =$ signed dist of x to L
  and $f(x) \propto$ signed dist of x to L.

Elements of Statistical Learning (2nd Ed.) ©Hastie, Tibshirani & Friedman 2009 Chap 4



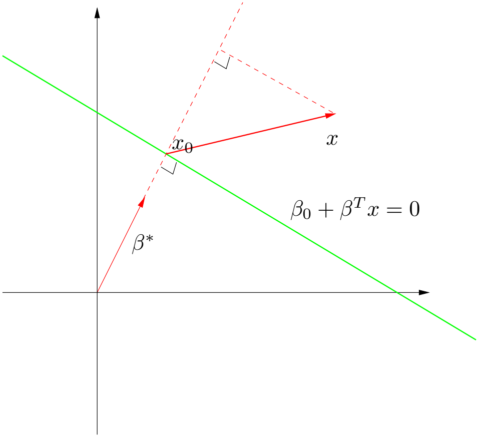


FIGURE 4.15. *The linear algebra of a hyperplane (affine set).*

Case I: two classes are separable

- WLOG, assume $||\beta|| = 1$ in $f(x) = \beta_0 + \beta' x$.
  Classifier: $G(x) = \text{sign}(f(x))$.
- Since the two claases are separable,
  1) Exists a $f(x) = \beta_0 + \beta' x = 0$ s.t. $Y_i f(X_i) > 0$ for all i;
  2) Exists a $f(x) = \beta_0 + \beta' x = 0$ s.t. the margin is maximized; Fig 12.1.
- Optimization problem
  $\max_{\beta_0, \beta, ||\beta||=1} M$
  s.t. $Y_i(\beta_0 + \beta' X_i) \geq M$ for $i = 1, ..., n$.
  Q: what is $\beta_0 + \beta' X_i$?
- Or, $\max_{\beta_0, \beta} M$
  s.t. $Y_i(\beta_0 + \beta' X_i)/||\beta|| \geq M$ for $i = 1, ..., n$.
- Set $||\beta|| = 1/M$, then
  $\min_{\beta_0, \beta} ||\beta||$ or $\min_{\beta_0, \beta} \frac{1}{2}||\beta||^2$
  s.t. $Y_i(\beta_0 + \beta' X_i) \geq 1$ for $i = 1, ..., n$.

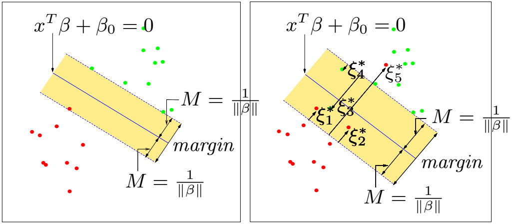


FIGURE 12.1. *Support vector classifiers. The left panel shows the separable case. The decision boundary is the solid line, while broken lines bound the shaded maximal margin of width* $2M = 2/\|\beta\|$. *The right panel shows the nonseparable (overlap) case. The points labeled* ξ_j^* *are on the wrong side of their margin by an amount* $\xi_j^* = M\xi_j$; *points on the correct side have* $\xi_j^* = 0$. *The margin is maximized subject to a total budget* $\sum \xi_i \leq$ *constant. Hence* $\sum \xi_j^*$ *is the total distance of points on the wrong side of their margin.*

- Convex programming: a quadratic obj with linear inequality constraints.
- Rewritten as a Lagrange function, ...
  β is defined by some *support points/vectors* X_i's.
  Fig 4.16: 3 SVs
- Remarks: 1) SVC: a large margin leads to better separation/prediction on test data!?
- 2) Robustness: β_0 and β determined only by SVs, but ...

Case II: non-separable

- Introduce some new variable ξ's:
  $\max_{\beta_0,\beta,||\beta||=1} M$
  s.t. $Y_i(\beta_0 + \beta' X_i) \geq M(1 - \xi_i)$
  and $\xi_i \geq 0$ and $\sum_{i=1}^{n} \xi_i \leq B$
  for $i = 1, ..., n$,
- Rewrite
  $\min_{\beta_0,\beta} \frac{1}{2}||\beta||^2 + C \sum_{i=1}^{n} \xi_i$
  s.t. $\xi_i \geq 0$ and $Y_i(\beta_0 + \beta' X_i) \geq 1 - \xi_i \ \forall i$,
  where C is the "cost", a tuning parameter.
  Fig 12.2
- ....similar results as before (e.g. convex programming, SVs)
  (8000): Computing §12.2.1.

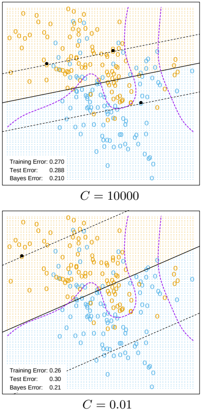
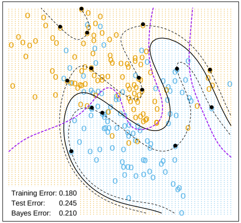


FIGURE 12.2. *The linear support vector boundary for the mixture data example with two overlapping classes, for two different values of C. The broken lines*

SVM

- SVC: linear in the input space.
- Basis expansion: use $h(X_i) = (h_1(X_i), ..., h_M(X_i))'$;
  $f(x) = \beta_0 + \beta' h(x)$.
- "Kernel trick": it turns out no need to "first transform then fit"; just use some kernel function $K(.,.)$:
  $K(x,z) =< h(x), h(z) >$.
- Three popular choices:
  0) Linear: $K(x,z) =< x, z >= x'z$.
  1) dth degree polynomial: $K(x,z) = (1 + \gamma < x, z >)^d$.
  2) radial basis: $K(x,z) = \exp(-\gamma ||x - z||^2)$.
  3) neural network/sigmoid: $K(x,z) = tanh(\gamma < x, z > +c)$.
  Logistic: $l(x) = \frac{1}{1+e^{-x}}$;
  Hyperbolic tangent: $tanh(x) = \frac{1-e^{-x}}{1+e^{-x}} = 2l(x) - 1$.
- Kernel: influences the performance; Fig 12.3.
- How to choose a kernel (and its parameters)? 1) prior; 2) tuning.

SVM - Degree-4 Polynomial in Feature Space



SVM - Radial Kernel in Feature Space

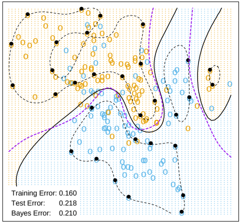


FIGURE 12.3. *Two nonlinear SVMs for the mixture data. The upper plot uses a 4th degree polynomial*

SVM as a penalization method

- SVM $f(x) = \beta_0 + \beta' h(x)$ is obtained from $\min_{\beta_0,\beta} \sum_{i=1}^{n} (1 - Y_i f(X_i))_+ + \lambda ||\beta||^2$.
  why?
- Marvelous!
  1) Explain why SVM is robust to high-dim data!
  2) Explain what SVM is doing:
- Hinge loss $L(Y, f(X)) = (1 - Y_i f(X_i))_+$
  $f^*(X) = \arg\min_f EL(Y, f(X)) = 1$ if $Pr(Y = 1|X) \geq 1/2$; $= -1$ o/w.
  SVM estimates the decision boundary directly!
  Need some effort to estimate the probabilities (Wang et al 2008, B'ka).
- AND 3) generalizations: ...

- Generalizations: use other loss or penalty functions, e.g. for $K > 2$ classes and for regression (SVR).
  - Fig 12.5: use binomial deviance; can directly estimate $P(Y = 1|X)$.
  - Extending to $K > 2$ classes (Wang et al 2007, JASA); rather than using "all pari-wise comparisons" with a binary SVM.
  - SVR: $V_\epsilon(r) = \max(|r| - \epsilon, 0) = (|r| - \epsilon)_+$; or Huber's; Fig 12.8.
  - Use the Lasso and others for VS (Wang et al 2007, JASA).
  - May even use even a better loss function (Shen et al 2003, JASA).
- Example code: ex8.1.R

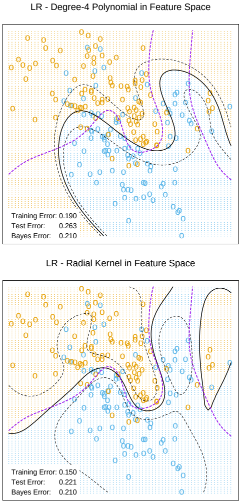


FIGURE 12.5. *The logistic regression versions of the SVM models in Figure 12.3, using the identical kernels*

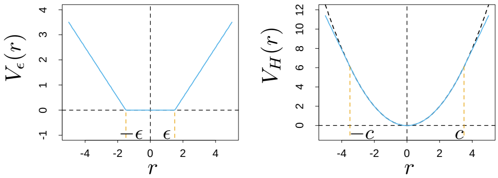


FIGURE 12.8. *The left panel shows the ϵ-insensitive error function used by the support vector regression machine. The right panel shows the error function used in Huber's robust regression (blue curve). Beyond $|c|$, the function changes from quadratic to linear.*

(8000): Computing §12.2.1, 12.3.1

- Recall that for SVC $f(x) = \beta_0 + \beta' x$,
  $\min_{\beta_0,\beta,\xi} \frac{1}{2}||\beta||^2 + C\sum_{i=1}^n \xi_i$
  s.t. $\xi_i \geq 0$ and $Y_i(\beta_0 + \beta' X_i) \geq 1 - \xi_i\ \forall i$.
- The Lagrangian (primal):
  $L_P = \frac{1}{2}||\beta||^2 + C\sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i[Y_i(\beta_0 + \beta' X_i) - (1 - \xi_i)] - \sum_{i=1}^n \mu_i \xi_i$, where C, α_i's, μ_i's and ξ_i's are all ≥ 0.
- Set the derivatives wrt β, β_0 and ξ_i to be 0, we have the Lagrangian (dual):
  $L_D = \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i=1}^n \sum_{i'=1}^n \alpha_i \alpha_{i'} Y_i Y_{i'} < X_i, X_{i'} >$.
- The solution satisfies: $\hat{\beta} = \sum_{i=1}^n \hat{\alpha}_i Y_i X_i$.
- Now, with a new SVM $f(x) = \beta_0 + \beta' h(x)$,
  $L_D = \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i=1}^n \sum_{i'=1}^n \alpha_i \alpha_{i'} Y_i Y_{i'} < h(x_i), h(x_{i'}) >$.
- The solution:
  $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}' h(x) = \hat{\beta}_0 + h(x)' \sum_{i=1}^n \hat{\alpha}_i Y_i h(X_i) = \hat{\beta}_0 + \sum_{i=1}^n \hat{\alpha}_i Y_i < h(x), h(X_i) >$.