

Feature Selection and Classification via a hybrid SVM

Hui Zou

School of Statistics

University of Minnesota

Agenda

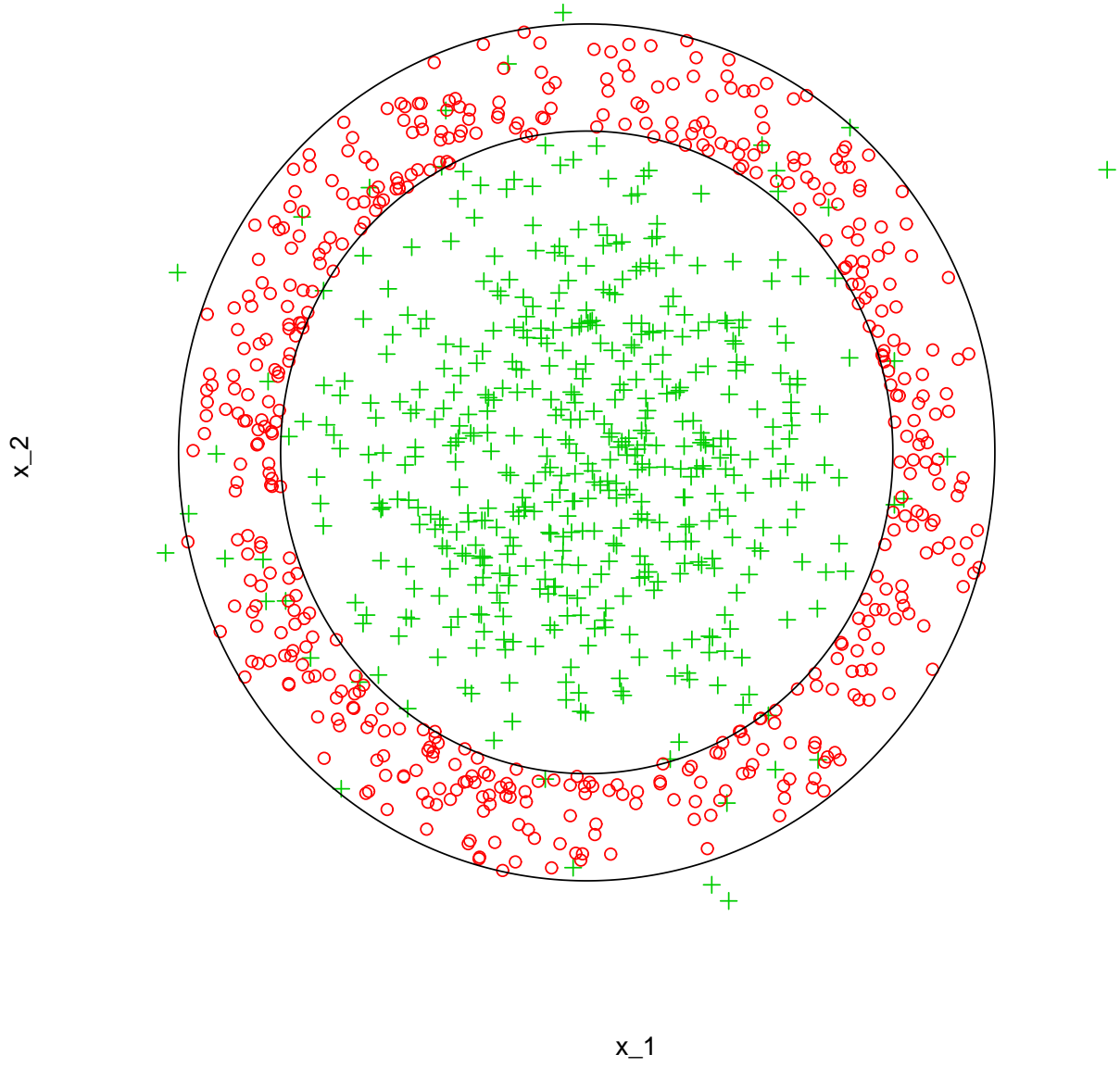
- Why sparse learning is important?
- Lasso (the ℓ_1 regularization)
- ℓ_1 -norm SVMs
- Adaptive Lasso
- The hybrid SVM

Sparse Learning

- learn the model structure
- improve the learning accuracy

Example: Orange Data Model

- We generated 50 observations in each of two classes.
- The first class ("+") has two independent standard normal inputs x_1, x_2 . The second class ("-") also has two standard normal independence inputs, but conditioned on $4.5 \leq x_1^2 + x_2^2 \leq 8$.
- Included q independent standard normal inputs in the model. We let q be 2, 4, 6, 8, 12, 16.
- We used the enlarged dictionary $D = \{\sqrt{2}x_j, \sqrt{2}x_jx_k, x_j^2, j, k = 1, 2, \dots, 2 + q\}$ to build the SVM classifiers.



Orange Data Model: the 2-norm SVM

	p	Misclassification Error
$q = 2$	14	9.97 (0.09) %
$q = 4$	27	12.87 (0.11) %
$q = 6$	44	16.17 (0.14) %
$q = 8$	75	19.21 (0.15) %
$q = 12$	90	24.30 (0.14) %
$q = 16$	152	27.81 (0.14) %

Some Observations

- the classification error of the SVM increases quickly with the dimension (p).
- noise features can greatly degrade the performance of the SVM.
- in order to improve its accuracy it is necessary to remove the noise features.

How?

Best subset selection plus model selection criteria (AIC, BIC etc.)

- Equivalent to using an ℓ_0 penalty
- A NP hard problem (Huo, 2005)
- Instability (Breiman, 1996)
- A better strategy: the ℓ_1 approach

Lasso

Data (\mathbf{X}, \mathbf{y}) . \mathbf{X} is the $n \times p$ predictor matrix of standardized variables; and \mathbf{y} is the response vector.

$$\arg \min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \sum_{j=1}^p |\beta_j|$$

$$\arg \min \sum_{i=1}^n -\log(L(y_i, \beta_0 + x_i^T \beta)) + \lambda \sum_{j=1}^p |\beta_j|.$$

Good Properties of Lasso

- Bias-variance tradeoff by a continuous shrinkage.
- Variable selection by the ℓ_1 penalization.
- A combination of ℓ_0 selection and ℓ_2 shrinkage.
- Flexible: Lasso can be used with general loss functions, even in unsupervised learning problems. Sparse PCA (Zou et al. 2004), Variable selection in Clustering (Pan and Shen, 2005).

A short literature review

- Wavelet Shrinkage; Donoho and Johnstone (1994, 1995)
- Basis pursuit; Chen, Donoho and Saunders (1995)
- Lasso; Tibshirani (1996, 1997)
- Least Angle Regression; Efron, Hastie, Johnstone and Tibshirani (2002)–**Solution paths of Lasso**
- The 1-norm support vector machine; Solution paths by Zhu, Rosset, Hastie and Tibshirani (2003)
- Many other extensions and applications

Lasso vs Ridge

$$\text{Ridge} \quad \arg \min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \sum_{j=1}^p |\beta_j|^2$$

- Ridge (ℓ_2) can produce very accurate predictive models, but it cannot deliver a sparse model.
- When there are a few noise variables, the performance of ridge can be severely damaged.
- Lasso (ℓ_1) is much more robust to noise variables.
- Ridge is better when all predictors contribute to the response, while lasso is preferred if there are noise variables.
- [the best-on-sparsity principle](#) (Friedman et al 2003).

SVMs and Ridge

- The standard SVM finds a hyperplane $(x^T \beta + \beta_0)$ that creates the biggest margin between the training points for class 1 and class -1 :

$$\begin{aligned} & \max_{\beta, \beta_0} \frac{1}{\|\beta\|_2} \\ \text{subject to} & \quad y_i(\beta_0 + x_i^T \beta) \geq 1 - \xi_i, \quad \forall i = 1, 2, \dots, n \\ & \quad \xi_i \geq 0, \quad \sum \xi_i \leq B. \end{aligned}$$

- An equivalent **loss + penalty** formulation

$$(\hat{\beta}, \hat{\beta}_0) = \arg \min_{\beta, \beta_0} \sum_{i=1}^n [1 - y_i(x_i^T \beta + \beta_0)]_+ + \lambda \|\beta\|_2^2.$$

- Like Ridge regression, the SVM uses the ℓ_2 penalty to control its estimation variance. So it is called the 2-norm SVM.

SVMs using the ℓ_1 penalty

- Let's replace the ℓ_2 penalty with the ℓ_1 penalty

$$\arg \min_{\beta, \beta_0} \sum_{i=1}^n [1 - y_i(x_i^T \beta + \beta_0)]_+ + \lambda \|\beta\|_1.$$

- Why would one do that?
 - The 2-norm SVM will use all the features in classification, so it suffers severe damage caused by noise features.
 - The 1-norm SVM is able to delete many noise features by estimating their coefficients by zero. Thus the 1-norm SVM is a better choice than the 2-norm SVM if the underlying model has a sparse presentation.

	p	2-norm	1-norm	
$q = 2$	14	9.97 (0.09) %	8.00 (0.04) %	5
$q = 4$	27	12.87 (0.11) %	8.21 (0.04) %	5
$q = 6$	44	16.17 (0.14) %	8.42 (0.01) %	6
$q = 8$	75	19.21 (0.15) %	8.52 (0.06) %	7
$q = 12$	90	24.30 (0.14) %	8.65 (0.05) %	7
$q = 16$	152	27.81 (0.14) %	8.61 (0.05) %	7

Orange data example: compare the 2-norm and 1-norm SVMs.

- The performance of 1-norm SVM is robust against the number of noise features, because it deletes most of them.
- The bet-on-sparsity principle works well in this example.
- There are still several noise features selected by the 1-norm SVM.
- Could we further improve its accuracy?

A Hybrid SVM

1. Solve the 2-norm SVM:

$$\hat{\beta}(\ell_2) = \arg \min_{\beta, \beta_0} \sum_{i=1}^n [1 - y_i(x_i^T \beta + \beta_0)]_+ + \lambda_2 \|\beta\|_2^2.$$

2. Construct a weighted ℓ_1 penalty:

$$w_j = |\hat{\beta}(\ell_2)_j|^{-\gamma} \quad j = 1, \dots, p, \quad \gamma > 0$$

$$\|\beta\|_{W1} = \sum_{j=1}^p w_j |\beta_j|.$$

3. Solve a new penalized hinge loss problem:

$$(\hat{\beta}, \hat{\beta}_0) = \arg \min_{\beta, \beta_0} \sum_{i=1}^n [1 - y_i(x_i^T \beta + \beta_0)]_+ + \lambda \|\beta\|_{W1}.$$

The fitted classifier is $\text{Sgn}(\hat{\beta}_0 + x^T \hat{\beta})$.

Several Remarks

- The weighted 1-norm penalty is a data-driven quantity
- Adaptively penalizes each component (adaptive lasso).
- In penalized likelihood setting, Zou (2005) showed adaptive lasso enjoys
 - Model selection and estimation: Oracle properties, not shared by Lasso.
 - Risk bound: Oracle inequality
- The hybrid SVM is expected to outperform the 1-norm SVM in the sparse setting.

Oracle Properties

$$\arg \min \sum_{i=1}^n -\log(L(y_i, \beta_0 + x_i^T \beta)) + \lambda \sum_{j=1}^p \frac{|\beta_j|}{|\hat{\beta}_j|^\gamma}$$

Theorem 1 *Suppose that $\lambda/\sqrt{n} \rightarrow 0$ and $\lambda n^{\frac{\gamma-1}{2}} \rightarrow \infty$, then the adaptive lasso estimates must satisfy:*

1. *Consistent selection:* $\lim_n P(\mathcal{A}_n^* = \mathcal{A}) = 1$

2. *Asymptotic Efficiency:* $\sqrt{n}(\hat{\beta}_{\mathcal{A}}^{*(n)} - \beta_{\mathcal{A}}^*) \rightarrow_d N(0, \sigma^2 \mathbf{I}_{\mathcal{A}}^{-1})$

Not hold for Lasso

- optimal estimation leads to inconsistent selection
- sometimes lasso can never be consistent in selection.

Adaptive Lasso Shrinkage

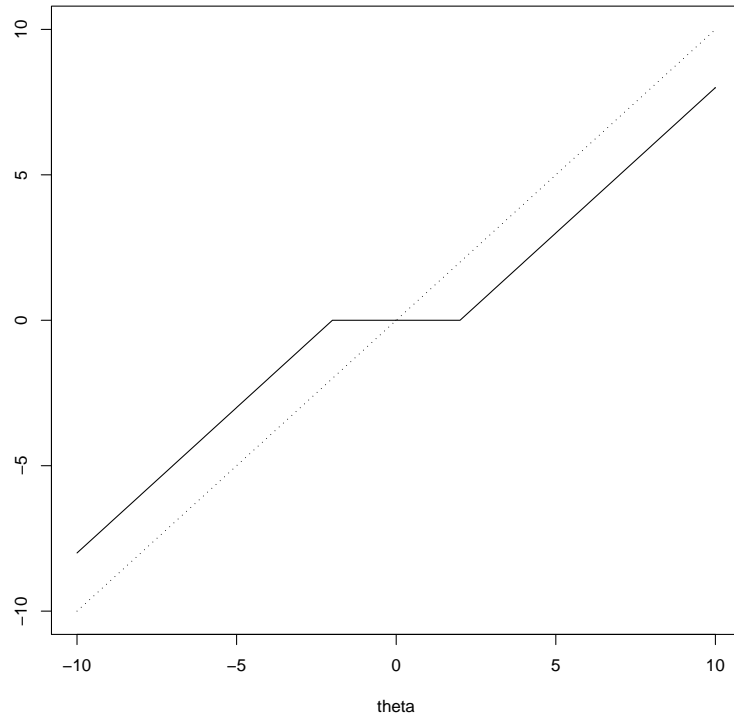
- Given n independent observations $\{y_i\}$ generated from

$$y_i = \mu_i + z_i \quad i = 1, 2, \dots, n$$

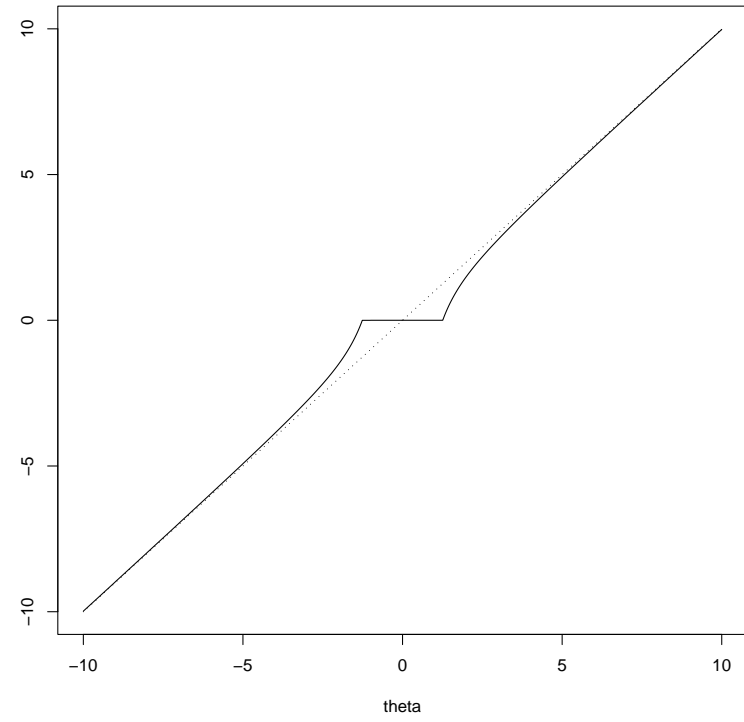
where z_i are i.i.d. $N(0, 1)$.

- Objective: estimate the mean vector μ .
- Estimation risk: $R(\hat{\mu}) = E[\sum_i^n (\hat{\mu}_i - \mu_i)^2]$.
- Stein says $\hat{\mu} = y$ is inadmissible for $m \geq 3$.

Lasso



Adaptive Lasso



- The **ideal risk** (by an Oracle) is $R(\text{ideal}) = \sum_i^n \min(\mu_i^2, 1)$ (Donoho and Johnstone 1994).
- Donoho and Johnstone (1994) showed

$$R(\text{Lasso}(\sqrt{2 \log n})) \leq (2 \log n + 1) (R(\text{ideal}) + 1).$$

$$\inf_{\hat{\mu}} \sup_{\mu} \frac{R(\hat{\mu})}{1 + R(\text{ideal})} \sim 2 \log n.$$

- **Theorem 2** Let $\lambda = (2 \log n)^{\frac{1+\gamma}{2}}$, then

$$R(\text{Adaptive Lasso}) \leq \left(2 \log n + 5 + \frac{4}{\gamma}\right) \left(R(\text{ideal}) + \frac{1}{2\sqrt{\pi}} (\log n)^{-1/2}\right).$$

Computational Considerations

- Computationally, the hybrid SVM is equivalent to the 1-norm SVM. Consider the working data (y, X_w) ($x_{wj} = x_j/w_j$).

$$(\hat{\beta}, \hat{\beta}_0) = \arg \min_{\beta, \beta_0} \sum_{i=1}^n [1 - y_i(x_{iw}^T \beta + \beta_0)]_+ + \lambda \|\beta\|_1.$$

- Tuning parameter selection
 - first find the best 2-norm SVM classifier for the data, then compute the weights using the 2-norm SVM classifier.
 - choose the optimal pair of (λ, γ) by cross-validation. In our experiments we choose γ from the set $\{1, 2, 4\}$.

	p	1-norm			hybrid		
		C	IC	PPS	C	IC	PPS
$q = 2$	14	2	3	0.206	2	0	0.724
$q = 4$	27	2	3	0.19	2	0	0.72
$q = 6$	44	2	4	0.11	2	0	0.69
$q = 8$	75	2	5	0.118	2	0	0.71
$q = 12$	90	2	5	0.064	2	0	0.668
$q = 16$	152	2	4	0.036	2	0	0.652

Orange data example: feature selection.

	p	1-norm	hybrid
$q = 2$	14	8.00 (0.04) %	7.27 (0.04) %
$q = 4$	27	8.21 (0.04) %	7.45 (0.04) %
$q = 6$	44	8.42 (0.01) %	7.63 (0.05) %
$q = 8$	75	8.52 (0.06) %	7.65 (0.05) %
$q = 12$	90	8.65 (0.05) %	7.66 (0.05) %
$q = 16$	152	8.61 (0.05) %	7.74 (0.06) %

Orange data example: classification error.

Simulation Model 2

- Simulated a training data set consisting of 100 observations from the model

$$y \sim \text{Bernoulli}\{p(x^T \beta + \beta_0)\},$$

where $p(u) = \exp(u)/(1 + \exp(u))$.

$\beta = (3, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 3)$ and $\beta_0 = 0$.

- x are standard normal, where the correlation between x_i and x_j is ρ . We considered both $\rho = 0.5$ and $\rho = 0$.
- The Bayes rule is to assign datum (x_1, \dots, x_{12}) to class $2I(x_1 + x_6 + x_{12}) - 1$.

Simulation model 2

	$\rho = 0.5$				$\rho = 0$			
	Error %	C	IC	PPS	Error %	C	IC	PPS
2-norm	10.41 (0.06)				13.61 (0.06)			
1-norm	9.56 (0.07)	3	2	0.126	12.20 (0.06)	3	2	0.158
hybrid	8.80 (0.06)	3	0	0.51	11.36 (0.05)	3	0	0.634
Bayes	7.38				10.19			

Simulation Model 3

- Simulated a training data set consisting of 100 observations from the model $y \sim \text{Bernoulli}\{p(x^T \beta + \beta_0)\}$, where $\beta = (3, 2, 0, 0, 0, 0, 0, 0, 0)$ and $\beta_0 = 1$.
- x are standard normal, where the correlation between x_i and x_j is $\rho^{|i-j|}$. We considered both $\rho = 0.5$ and $\rho = 0$.
- The Bayes rule is to assign datum (x_1, \dots, x_9) to class $2I(3x_1 + 2x_2 + 1) - 1$.

Simulation model 3

	$\rho = 0.5$				$\rho = 0$			
	Error %	C	IC	PPS	Error %	C	IC	PPS
2-norm	13.77 (0.05)				14.03 (0.05)			
1-norm	12.75 (0.04)	2	2	0.164	13.85 (0.08)	2	3	0.068
hybrid	12.63 (0.05)	2	0	0.616	13.09 (0.06)	2	0	0.568
Bayes	11.74				11.67			

A Dense Model

Simulated a training data set consisting of 100 observations from the model $y \sim \text{Bernoulli}\{p(x^T \beta + \beta_0)\}$, where $\beta = (3, 3, 3, 3, 3, 3, 3, 3)$ and $\beta_0 = 0$.

Bayes	2-norm	1-norm	hybrid
6.36 %	8.75 (0.05) %	9.44 (0.07) %	9.42 (0.08) %

From the simulation, we have observed

- The hybrid SVM
 - significantly dominates the 1-norm SVM in terms of classification accuracy.
 - has a greater tendency to delete all the noise features.
 - has a much higher probability of perfect selection.
- When all the features are significant, the 2-norm SVM can be more accurate than the sparse SVMs.
- In sparse settings the sparse SVMs have big advantages over the 2-norm SVM. So the bet-on-sparsity principle is a good rule to follow in classification.

Benchmark Datasets

	# of observations	# of features
spam	4601	57
ionosphere	351	34
WDBC	569	30

From UCI Machine Learning Repository.

- Spam has a test set of 1536 observations.
- For `ionosphere` and `WDBC` data, we randomly selected 2/3 data for training and the other 1/3 data as the test set. We repeated the randomization ten times.

	2-norm	1-norm		hybrid	
	Error %	Error %	NSV	Error %	NSV
spam	7.31 (0.66)	7.37 (0.66)	55	7.08 (0.65)	41
iono	13.81 (0.83)	13.47 (0.80)	11	12.16 (0.72)	8
wdbc	3.26 (0.31)	3.37 (0.25)	11	3.29 (0.22)	8

Conclusion

- The weighted ℓ_1 penalty improves the performance of the 1-norm SVM in variable selection and classification.
- When the underlying model is truly sparse, both the 1-norm SVM and hybrid SVM outperform the 2-norm SVM.
- The hybrid SVM seems to be a better choice than the 1-norm SVM in applying the bet-on-sparsity principle.