Chapter 10. Semi-Supervised Learning

Wei Pan

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

PubH 7475/8475

Outline

- Mixture model: L_1 penalization for variable selection
  Pan et al (2006, Bioinformatics)
  - Introduction: motivating example
  - Methods: standard and new ones
  - Simulation
  - Example
  - Discussion
- Transductive SVM (TSVM):
  Wang, Shen & Pan (2007, CM; 2009, JMLR)
- Constrained K-means: Wagstaff et al (2001)

Introduction

- Biology: Do human blood outgrowth endothelial cells (BOECs) belong to or are closer to large vessel endothelial cells (LVECs) or microvascular endothelial cells (MVECs)?
- Why important: BOECs are being explored for efficacy in endothelial-based gene therapy (Lin et al 2002), and as being useful for vascular diagnostic purposes (Hebbel et al 2005); in each case, it is important to know whether BOEC have characteristics of MVECs or of LVECs.
- Based on the expression of gene CD36, it seems reasonable to characterize BOECs as MVECs (Swerlick et al 1992).
- However, CD36 is expressed in endothelial cells, monocytes, some epidermal cells and a variety of cell lines; characterization of BOECs or any other cells using a single gene marker seems unreliable.

- Jiang (2005) conducted a genome-wide comparison: microarray gene expression profiles for BOEC, LVEC and MVEC samples were clustered; it was found that BOEC samples tended to cluster together with MVEC samples, suggesting that BOECs were closer to MVECs.
- Two potential shortcomings:
  1. Used hierarchical clustering; ignoring the known classes of LVEC and MVEC samples;
     Alternative? Semi-supervised learning: treating LVEC and MVEC as known while BOEC unknown (see McLachlan and Basford 1988; Zhu 2006 for reviews).
     Here it requires learning a novel class: BOEC may or may not belong to LVEC or MVEC.
  2. Used only 37 genes that best discriminate b/w LVEC and MVEC.
     Important: result may critically depend on the features or genes being used; the few genes might not reflect the whole picture.
     Alternative? Start with more genes; but ...
     A dilemma: too many genes might lead to covering true clustering structures; to be shown later.

- For high-dimensional data, necessary to have feature selection, preferably embedded within the learning framework – automatic/simultaneous feature selection.
- In contrast to sequential methods: first selecting features and then fitting/learning a model;
  Pre-selection may perform terribly;
  Why: selected features may not be relevant at all to uncovering interesting clustering structures, due to the separation between the two steps.
- We propose a penalized mixture model: semi-supervised learning; automatic variable selection simultaneously with model fitting.

- With more genes included in a starting model and with appropriate gene selection, BOEC samples are separate from LVEC and MVEC samples.
- Finite mixture models studied in the statistics and machine learning literature (McLachlan and Peel 2002; Nigam et al 2006), even applied to microarray data analysis (Alexandridis et al 2004), our proposal of using a penalized likelihood to realize automatic variable selection is novel; in fact, variable selection in this context is largely a neglected topic.
- This work extends the penalized unsupervised learning/clustering analysis method of Pan and Shen (2007) to semi-supervised learning.

Semi-Supervised Learning via Standard Mixture Model

- Data
  Given n K-dimensional obs's: $x_1, \ldots, x_n$; the first n_0 do not have class labels while the last n_1 have.
  There are $g = g_0 + g_1$ classes: the first g_0 unknown/novel classes to be discovered. while the last g_1 known.
  $z_{ij} = 1$ iff x_j is **known** to be in class i; $z_{ij} = 0$ o/w.
  Note: z_{ij}'s are missing for $1 \leq j \leq n_0$.
- A mixture model as a generative model:

$$f(x; \Theta) = \sum_{i=1}^{g} \pi_i f_i(x; \theta_i)$$

π_i: unknown prior prob's;
f_i: class-specific distribution with unknown parameters θ_i.

- For high-dim and low-sample-sized data, we propose

$$f_i(x_j;\theta_i) = \frac{1}{(2\pi)^{K/2}|V|^{1/2}} \exp\left(-\frac{1}{2}(x_j-\mu_i)'V^{-1}(x_j-\mu_i)\right),$$

  where $V = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_K^2)$, and $|V| = \prod_{k=1}^K \sigma_k^2$.
- Posterior prob of x_j's coming from class/component i:

$$\begin{aligned}\tau_{ij} &= \frac{\pi_i f_i(x_j;\theta_i)}{\sum_{l=1}^g \pi_l f_l(x_j;\theta_l)} \\ &= \frac{\pi_i \prod_{k=1}^K \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(x_{jk}-\mu_{ik})^2}{2\sigma_k^2}\right)}{\sum_{l=1}^g \pi_l \prod_{k=1}^K \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(x_{jk}-\mu_{lk})^2}{2\sigma_k^2}\right)},\end{aligned}$$

- Assign x_j to cluster $i_0 = \text{argmax}_i \tau_{ij}$.
- A key observation: if $\mu_{1k} = \mu_{2k} = ... = \mu_{gk}$ for some k, the terms involving x_{jk} will cancel out in τ_{ij}—-feature selection!

- Note: variable selection is possible under a common diagonal covariance matrix V across all clusters.
  E.g., if use V_i (or a non-diagonal V), even if $\mu_{1k} = \mu_{2k} = ... = \mu_{gk}$, x_{jk} is still informative; e.g., $N(0,1)$ vs $N(0,2)$.
- $\Theta = \{(\pi_i, \theta_i) : i = 1, ..., g\}$ need to be estimated; MLE
- The log-likelihood is

$$\log L(\Theta) = \sum_{j=1}^{n_0} \log[\sum_{i=1}^{g} \pi_i f_i(x_j; \theta_i)] + \sum_{j=n_0+1}^{n} \log[\sum_{i=1}^{g} z_{ij} f_i(x_j; \theta_i)].$$

- Common to use the EM (Dempster et al 1977) to get MLE; see below for details.

Penalized Mixture Model

- Penalized log-likelihood: use a weighted L_1 penalty;

$$\log L_P(\Theta) = \log L(\Theta) + \lambda \sum_i \sum_k w_{ik} |\mu_{ik}|,$$

  where w_{ik}'s are weights to be given later.
- Penalty: model regularization; Bayesian connection.
- Assume that the data have been standardized so that each feature has sample mean 0 and sample variance 1.
- Hence, for any k, if $\mu_{1k} = ... = \mu_{gk} = 0$, then feature k will not be used.
- L_1 penalty serves to obtain a sparse solution: μ_{ik}'s are automatically set to 0, realizing variable selection.

- EM algorithm: E-step and M-step for other parameters are the same as in the usual EM, except M-step for μ_{ik};

$$\hat{\pi}_i^{(m+1)} = \sum_{j=1}^{n} \tau_{ij}^{(m)} / n, \tag{1}$$

$$\hat{\sigma}_k^{2,(m+1)} = \sum_{i=1}^{g} \sum_{j=1}^{n} \tau_{ij}^{(m)} (x_{jk} - \hat{\mu}_{ik}^{(m)})^2 / n, \tag{2}$$

$$\hat{\mu}_i^{(m+1)} = \text{sign}(\tilde{\mu}_i^{(m+1)}) \left(|\tilde{\mu}_i^{(m+1)}| - \frac{\lambda}{\sum_j \tau_{ij}^{(m)}} V^{(m)} w_i \right)_+ \tag{3}$$

where

$$\tau_{ij}^{(m)} = \begin{cases} \frac{\pi_i^{(m)} f_i(x_j; \theta_i^{(m)})}{f(x_j; \Theta^{(m)})}, & \text{if } 1 \le j \le n_0 \\ z_{ij}, & \text{if } n_0 < j \le n \end{cases} \tag{4}$$

$$\tilde{\mu}_i^{(m+1)} = \sum_{j=1}^{n} \tau_{ij}^{(m)} x_j / \sum_{j=1}^{n} \tau_{ij}^{(m)} \tag{5}$$

- Soft-thresholding: If $\lambda w_{ik} > |\sum_{j=1}^{n} \tau_{ij}^{(m)} x_{jk}/\sigma_k^{2,(m)}|$, then $\hat{\mu}_{ik}^{(m+1)} = 0$; otherwise, $\hat{\mu}_{ik}^{(m+1)}$ is obtained by shrinking $\tilde{\mu}_{ik}^{(m+1)}$ by an amount $\lambda w_{ik}\sigma_k^{2,(m)}/\sum_{j=1}^{n} \tau_{ij}^{(m)}$.
- In the EM for the standard mixture model, use $\tilde{\mu}_i^{(m+1)}$; no shrinkage or thresholding.
- Zou (2005, 2006) proposed using the weighted L_1 penalty in the context of supervised learning; we extend the idea to the current context: using $w_{ij} = 1/|\tilde{\mu}_{ik}|^w$ with $w \geq 0$; the standard L_1 penalty corresponds to $w = 0$.
- The weighted penalty automatically realizes a data-adaptive penalization: it penalizes more on smaller μ_{ik} while penalizing less on, and thus reducing the bias for, larger μ_{ik}, leading to better feature selection and classification performance.
- As in Zou (2006), we tried $w \in \{0, 1, 2, 4\}$ and found only minor differences in results for $w > 0$; for simplicity we will present results only for $w = 0$ and $w = 1$.

Model Selection

- To determine g_0 (and λ), use BIC (Schwartz 1978)

$$BIC = -2 \log L(\hat{\Theta}) + \log(n) d,$$

where $d = g + K + gK - 1$ is the total number of unknown parameters in the model; the model with a minimum BIC is selected (Fraley and Raftery 1998).

- For the penalized mixture model, Pan and Shen (2007) proposed a modified BIC:

$$BIC = -2 \log L(\hat{\Theta}) + \log(n) d_e,$$

where $d_e = g + K + gK - 1 - q = d - q$ with $q = \#\{\hat{\mu}_{ik} : \hat{\mu}_{ik} = 0\}$, an estimate of the "effective" number of parameters.

- The idea was borrowed from Efron et al (2004) and Zou et al (2004) in penalized regression/LASSO.
- No proof yet...
- Data-based methods, such as cross-validation or data perturbation (Shen and Ye 2002; Efron 2004), can be also used; but computationally more demanding.
- Trials and errors to find a λ (and g_0).

Simulated Data

- Simulation set-ups:
  - Four non-null (i.e. $g_0 > 0$) cases;
  - 20 obs's in each of the $g_0 = 1$ unknown and $g_1 = 2$ known classes;
  - $K = 200$ independent attributes; only $2K_1$ were informative;
  - Each of the first K_1 informative attributes: indep $N(0, 1)$, $N(0, 1)$ and $N(1.5, 1)$ for 3 classes;
  - Each of the next K_1 informative ones: indep $N(1.5, 1)$, $N(0, 1)$ and $N(0, 1)$;
  - Each of the $K - 2K_1$ noise variables: $N(0, 1)$;
  - $K_1 = 10$, 15, 20 and 30.
  - Null case: $g_0 = 0$; only the first $K_1 = 30$ attributes were discriminatory as before, and others not.

- For each case, 100 independent datasets.
- Comparing standard method without variable selection (i.e. $\lambda = 0$) and penalized method with $w = 0$.
- For each dataset, the EM was run 10 times; its starting values were from the output of the K-means with random starts; final result was the one with the max (penalized) likelihood (for the given λ).
- $\lambda \in \Phi = \{0, 2, 4, 6, 8, 10, 12, 15, 20, 25\}$; for a given g_0, chose the one with min BIC.
- Comparison between the standard and penalized methods:

Set-up 1: $2K_1 = 20$, $g_0 = 1$							
	Standard		Penalized				
g_0	Freq	BIC	Freq	BIC	λ	#Zero1	#Zero0
0	100	12029	35	10793	10.3	19.8	180.0
		(4)		(3)	(.1)	(.2)	(.0)
1	0	12464	65	10779	9.4	0.0	169.4
		(5)		(6)	(.1)	(.0)	(.8)

Set-up 2: $2K_1 = 30$, $g_0 = 1$							
	Standard		Penalized				
g_0	Freq	BIC	Freq	BIC	λ	#Zero1	#Zero0
0	100	11876	13	10741	9.9	29.9	170.0
1	0	12225	87	10693	8.3	0.0	154.5

Set-up 3: $2K_1 = 40$, $g_0 = 1$							
	Standard		Penalized				
g_0	Freq	BIC	Freq	BIC	λ	#Zero1	#Zero0
0	100	11733	1	10688	9.1	40	160
1	0	11977	99	10590	8.0	0.0	142.9
Set-up 4: $2K_1 = 60$, $g_0 = 1$							
	Standard		Penalized				
g_0	Freq	BIC	Freq	BIC	λ	#Zero1	#Zero0
0	86	11433	0	10567	8.5	-	-
1	14	11483	100	10367	6.8	0.0	112.9

Set-up 5: $K_1 = 30$, $g_0 = 0$							
	Standard		Penalized				
g_0	Freq	BIC	Freq	BIC	λ	#Zero1	#Zero0
0	100	11583	100	10506	8.1	23.6	170
		(5)		(5)	(.1)	(.7)	(.0)
1	0	12196	0	10510	8.1	-	-
		(5)		(5)	(.1)		

- Comparison with pre-variable-selection:
  - Use F-statistics to rank the genes;
  - Treat unlabeled data as a separate class?
    F_2: ignore unlabeled data; use only labeled data.
    F_3: treat unlabeled data as a separate class.
  - How many top genes? i.e. K_0=?
  - Use BIC to select K_0?

Table: Frequencies of the selected numbers (g_0) of the cluster for unlabeled data in variable selection from 100 simulated datasets: top K_0 genes with the largest F-statistics based on labeled data (F_2), or both labeled and unlabeled data (F_3), were used in the standard mixture model; the last row was for the frequency of g_0 values selected when the best K_0 values were determined by BIC; true $g_0 = 1$.

	F_2		F_3	
K_0	$g_0 = 0$	$g_0 = 1$	$g_0 = 0$	$g_0 = 1$
5	83	1	1	15
15	36	0	0	64
20	20	0	0	80
30	1	0	0	99
40	0	0	0	100
50	0	0	0	100
60	0	0	0	100
$\hat{K}_0$	83	1	1	15

Summary

- No variable selection: tended to select $g_0 = 0$ because of the presence of many noise variables; correct in some sense!
- Pre-variable selection: tended to select $g_0 = 0$ because the selected model was indeed correct (based on a subset of non-informative variables) and most parsimonious, albeit of no interest!

Real Data

- 28 LVEC and 25 MVEC samples from Chi et al (2003); cDNA arrays.
- 27 BOEC samples; Affy arrays.
- Combined data: 9289 unique genes in both data.
- Need to minimize systematic bias due to different platforms.
- 6 human umbilical vein endothelial cell (HUVEC) samples from each of the two datasets.
- Jiang studied 64 possible combinations of a three-step normalization procedure and identified the one maximizing the extent of mixing of the 12 HUVEC samples.
- Normalized the data in the same way

- $g_0 = 0$ or 1; $g_1 = 2$.
- 6 models: 1) 3 methods: standard, penalized with $w = 0$, and penalized with $w = 1$; 2 values of g_0: 0 or 1.
- The EM randomly started 20 times with the starting values from the K-means output.
- At convergence, used the posterior probabilities to classify BOEC samples, as well as LVEC and MVEC samples.
- Used 3 sets of the genes in the starting model.
- Using 37 genes best discriminating LVEC and MVEC:

Table: Semi-supervised learning with 37 genes. The BIC values of the six models (from left to right and from top to bottom) were 2600, 2549, 2510, 2618, 2520 and 2467 respectively.

	$g_0 = 0,\ g_1 = 2$					
	$\lambda = 0$		$\lambda = 5,\ w = 0$		$\lambda = 2,\ w = 1$	
Sample	1	2	1	2	1	2
BOEC	1	26	6	21	0	27
LVEC	24	4	25	3	25	3
MVEC	2	23	3	22	2	23

	$g_0 = 1,\ g_1 = 2$								
	$\lambda = 0$			$\lambda = 6,\ w = 0$			$\lambda = 3,\ w = 1$		
Sample	1	2	3	1	2	3	1	2	3
BOEC	13	1	13	17	1	9	16	0	11
LVEC	1	24	3	2	24	2	1	25	2
MVEC	0	1	24	2	1	24	0	2	23

Table: Numbers of the 37 features with zero mean estimates.

	$g_0 = 0,\ g_1 = 2$					
	$\lambda = 5,\ w = 0$			$\lambda = 2,\ w = 1$		
Cluster	1	2	All	1	2	All
#Zeros	11	11	11	14	18	14
	$g_0 = 1,\ g_1 = 2$					
	$\lambda = 6,\ w = 0$				$\lambda = 3,\ w = 1$			
Cluster	1	2	3	All	1	2	3	All
#Zeros	21	10	11	5	24	18	20	12

- Using top 1000 genes discriminating LVEC and MVEC;
- Using top 1000 genes with largest sample variances;
- —-similar results!

Discussion

- As expected, results depend on which features are being used.
- For our motivating example, with various larger sets of genes, the BOEC samples seemed to be different from both LVEC and MVEC samples, and formed a new class.
- However, the result might owe to different microarray chips used.
- Our major contribution: use of penalized mixture model for semi-supervised learning.
- Lesson: As in clustering (Pan adn Shen 2007), variable selection in semi-supervised learning is both critical and challenging; either skipping variable selection or pre-selection may not work well, even though a *correct* model *of no interest* can be identified!

- Comparison to nearest shrunken centroids (NSC) (Tibshirani et al 2002; 2003)
  - Similar: 1. aim to handle high-dimensional (and low-sample-sized) data; 2. assume a Normal distribution for each cluster or class; 3. adopt a common diagonal covariance matrix for all the clusters/classes; for simplicity and for variable selection; 4. use soft-thresholding to realize variable selection.
  - Diff: 1. for supervised and semi-supervised respectively; 2. penalization: ad hoc in NSC; here in the general and unified framework of penalized likelihood.
- Here a single Normal distribution for each class; a mixture of Normals can be also used (Nigam et al 2006).
- Is model-based easier to incorporate the idea of "tight clustering" (Tseng and Wong 2005)?
- Other extensions in clustering: grouped VS (Xie, Pan & Shen 2008, Biometrics); cluster-specific diagonal cov matrices (Xie, Pan & Shen 2008, EJS); unconstrained covariance structures by glasso (Zhou, Pan & Shen 2009, EJS)...

TSVM

- Labeled data: (x_i, y_i), $i = 1, ..., n_l$;
  Unlabeled data: (x_i), $i = n_l + 1, ..., n$.
- SVM: consider linear kernel; i.e.

$$f(x) = \beta_0 + \beta' x.$$

- Estimation in SVM:

$$\min_{\beta_0, \beta} \sum_{i=1}^{n_l} L(y_i f(x_i)) + \lambda_1 ||\beta||^2$$

- TSVM: aim the same $f(x) = \beta_0 + \beta' x$.

- Estimation in TSVM:

$$\min_{\{y^*_{n_l+1},\ldots,y^*_n\},\beta_0,\beta} \sum_{i=1}^{n_l} L(y_i f(x_i)) + \lambda_1 ||\beta||^2 + \lambda_2 \sum_{i=n_l+1}^{n} L(y_i^* f(x_i))$$

- Equivalently (Wang, Shen & Pan 2007; 2009, JMLR),

$$\min_{\beta_0,\beta} \sum_{i=1}^{n_l} L(y_i f(x_i)) + \lambda_1 ||\beta||^2 + \lambda_2 \sum_{i=n_l+1}^{n} L(|f(x_i)|)$$

- Computational algorithms DO matter!
- Very active research going on...

Table: **Linear learning:** Averaged test errors as well as the estimated standard errors (in parenthesis) of SVM with labeled data alone, TSVM^{Light}, and TSVM^{DCA}, over 100 pairs of training and testing samples, in the simulated and benchmark examples.

Data	SVM	TSVM^{Light}	TSVM^{DCA}
Example 1	.345(.0081)	.230(.0081)	.220(.0103)
Example 2	.333(.0129)	.222(.0128)	.203(.0088)
WBC	.053(.0071)	.077(.0113)	.037(.0024)
Pima	.328(.0092)	.316(.0121)	.314(.0086)
Ionosphere	.257(.0097)	.295(.0085)	.197(.0071)
Mushroom	.232(.0135)	.204(.0113)	.206(.0113)
Email	.216(.0097)	.227(.0120)	.196(.0132)

Table: **Nonlinear learning with Gaussian kernel:** Averaged test errors as well as the estimated standard errors (in parenthesis) of SVM with labeled data alone, TSVM^{Light}, and TSVM^{DCA}, over 100 pairs of training and testing samples, in the simulated and benchmark examples.

Data	SVM	TSVM^{Light}	TSVM^{DCA}
Example 1	.385(.0099)	.267(.0132)	.232(.0122)
Example 2	.347(.0119)	.258(.0157)	.205(.0091)
WBC	.047(.0038)	.037(.0015)	.037(.0045)
Pima	.353(.0089)	.362(.0144)	.330(.0107)
Ionosphere	.232(.0088)	.214(.0097)	.183(.0103)
Mushroom	.217(.0135)	.217(.0117)	.185(.0080)
Email	.226(.0108)	.275(.0158)	.192(.0110)

Constrained K-means

- Ref: Wagstaff et al (2001); COP-k-means
- K-means with two types of constraints:
  1. Must-link: two obs's have to be in the same cluster;
  2. Cannot-link: two obs's cannot be in the same cluster.
- May not be feasible, or even reasonable.
  Many modifications.
- Constrained spectral clustering (Liu, Pan & Shen 2013, Front Genet).