Small-variance Asymptotics for Dirichlet Process Mixtures of SVMs

Yining Wang Jun Zhu

Tsinghua University

July, 2014

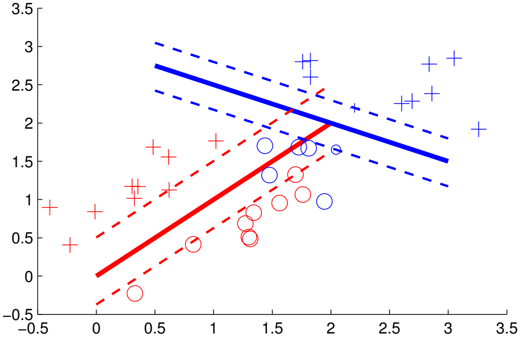
Outline

Outline

1. Infinite SVM (Zhu et al., 2011)
   - The Infinite SVM (iSVM) Model
   - Gibbs-iSVM
2. The Max-Margin DP-means (M^2DPM) Algorithm
   - Small-variance Asymptotic (SVA) Analysis
   - The M^2DPM Algorithm
3. Experiments
4. Summary

Infinite SVM (*Zhu et al., ICML 2011*)

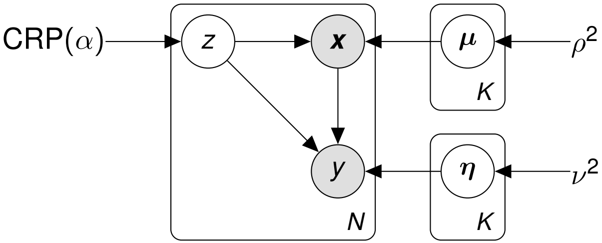
Solving clustering and classification simultaneously.



Motivation:

- Improve classification by finding underlying clusters.
- Improve clustering by incorporating supervised information
- Unknown # of clusters ⇒ a nonparametric treatment.

Infinite SVM



- Nonparametric prior:

$$p_0(z_i = k|\alpha, \boldsymbol{z}_{-i}) \propto \begin{cases} n_{-i,k}, & \text{if } n_{-i,k} > 0 \\ \alpha, & \text{otherwise} \end{cases}$$

- Max-margin classification model:

$$\phi(y_i|\boldsymbol{x}_i, \boldsymbol{\eta}, z_i = k) = \exp(-2c \cdot \max(0, 1 - y_i \boldsymbol{\eta}_k^\top \boldsymbol{x}_i))$$

Outline

1. Infinite SVM (Zhu et al., 2011)
   - The Infinite SVM (iSVM) Model
   - Gibbs-iSVM
2. The Max-Margin DP-means (M^2DPM) Algorithm
   - Small-variance Asymptotic (SVA) Analysis
   - The M^2DPM Algorithm
3. Experiments
4. Summary

Gibbs sampler for iSVM

Iteratively sample model parameters:

1. Mean parameter $\boldsymbol{\mu}_k$: Gaussian distribution.
2. Linear classifier $\boldsymbol{\eta}_k$: data augmentation.
3. Cluster assignments z_i: categorical distribution.

Problems: slow!

$$q(z_i = z_{new}) \propto \alpha \cdot \int \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}, \sigma^2 I) \mathrm{d}p_0(\boldsymbol{\mu}) \cdot \underbrace{\int \phi(y_i | \boldsymbol{x}_i, \boldsymbol{\eta}) \mathrm{d}p_0(\boldsymbol{\eta})}_{\text{difficult to evaluate}}$$

Outline

Small-variance Asymptotic (SVA) Analysis

1. **Assumptions:** model variance $\to 0$.
2. **Consequences:**
   - Connections between Bayesian posterior and deterministic loss.
   - Novel inference algorithms.
3. **Examples:**
   - Probabilistic PCA *vs.* PCA. (*Tipping et al., 1999*)
   - DP-means. (*Kulis et al., 2012*)
   - Efficient feature learning. (*Broderick et al., 2013*)
   - Asymp-iHMM. (*Roychowdhury et al., 2013*)

Small-variance Asymptotic (SVA) Analysis

1. Assumptions: model variance $\rightarrow$ 0.
2. Consequences:
   - Connections between Bayesian posterior and deterministic loss.
   - Novel inference algorithms.
3. Examples:
   - Probabilistic PCA *vs.* PCA. (*Tipping et al., 1999*)
   - DP-means. (*Kulis et al., 2012*)
   - Efficient feature learning. (*Broderick et al., 2013*)
   - Asymp-iHMM. (*Roychowdhury et al., 2013*)

Small-variance Asymptotic (SVA) Analysis

1. Assumptions: model variance $\rightarrow$ 0.
2. Consequences:
   - Connections between Bayesian posterior and deterministic loss.
   - Novel inference algorithms.
3. Examples:
   - Probabilistic PCA *vs.* PCA. (*Tipping et al., 1999*)
   - DP-means. (*Kulis et al., 2012*)
   - Efficient feature learning. (*Broderick et al., 2013*)
   - Asymp-iHMM. (*Roychowdhury et al., 2013*)

SVA on Gibbs-iSVM

SVA Assumptions: $\sigma^2, \nu^2 \to 0$, $c, \alpha \to \infty$.

- Existing clusters:

$$q(z_i = k) \propto n_{-i,k} \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^2}{\sigma^2} - 2c\max(0, 1 - y_i\boldsymbol{\eta}_k^\top \boldsymbol{x}_i)\right)$$

$$\Rightarrow Q_i(k) = \underbrace{s \cdot \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^2}_{\text{Clustering}} + \underbrace{2c(1 - y_i\boldsymbol{\eta}_k^\top \boldsymbol{x}_i)_+}_{\text{Classification}};$$

- Creating new clusters:

$$q(z_i = z_{new}) \propto \alpha \cdot \int \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}, \sigma^2 I)\mathrm{d}p_0(\boldsymbol{\mu}) \cdot \int \phi(y_i|\boldsymbol{x}_i, \boldsymbol{\eta})\mathrm{d}p_0(\boldsymbol{\eta})$$

$$\Rightarrow Q_i(new) = \underbrace{\lambda}_{\text{AIC}} + \underbrace{2c(1 - y_i\boldsymbol{\eta}^{*\top} \boldsymbol{x}_i)_+}_{\text{Classification}} + \underbrace{\|\boldsymbol{\eta}^*\|^2/\nu^2}_{\text{Regularization}}.$$

SVA on Gibbs-iSVM

SVA Assumptions: $\sigma^2, \nu^2 \to 0$, $c, \alpha \to \infty$.

- Existing clusters:

$$q(z_i = k) \propto n_{-i,k} \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^2}{\sigma^2} - 2c \max(0, 1 - y_i \boldsymbol{\eta}_k^\top \boldsymbol{x}_i)\right)$$

$$\Rightarrow Q_i(k) = \underbrace{s \cdot \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^2}_{\text{Clustering}} + \underbrace{2c(1 - y_i \boldsymbol{\eta}_k^\top \boldsymbol{x}_i)_+}_{\text{Classification}};$$

- Creating new clusters:

$$q(z_i = z_{new}) \propto \alpha \cdot \int \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}, \sigma^2 I) \mathrm{d}p_0(\boldsymbol{\mu}) \cdot \int \phi(y_i | \boldsymbol{x}_i, \boldsymbol{\eta}) \mathrm{d}p_0(\boldsymbol{\eta})$$

$$\Rightarrow Q_i(new) = \underbrace{\lambda}_{\text{AIC}} + \underbrace{2c(1 - y_i \boldsymbol{\eta}^{*\top} \boldsymbol{x}_i)_+}_{\text{Classification}} + \underbrace{\|\boldsymbol{\eta}^*\|^2/\nu^2}_{\text{Regularization}}.$$

Outline

The M^2DPM Algorithm

1. Repeat until convergence:
   - For each instance i: $z_i \leftarrow \text{argmin}_k Q_i(k)$.
   - For each cluster k: $\boldsymbol{\mu}_k \leftarrow \bar{\boldsymbol{x}}_k$.
   - Update $\{\boldsymbol{\eta}_k\}_{k=1}^K$ using data augmentation.
2. Deterministic loss function (via SVA on posterior):

$$\mathcal{L}(\boldsymbol{z}, \mu, \eta) = \underbrace{\sum_{k=1}^{K} \frac{\|\eta_k\|^2}{2\nu^2}}_{\text{Regularization}} + \underbrace{2c \sum_{i=1}^{n} (\zeta_i^{z_i})_+}_{\text{Classification}} + \underbrace{s \sum_{i=1}^{n} \|\boldsymbol{x}_i - \mu_{z_i}\|^2}_{\text{Clustering}} + \underbrace{\lambda \cdot K}_{\text{AIC}}.$$

3. Extension to exponential family distributions and multi-class classification.

The M²DPM Algorithm

1. Repeat until convergence:
   - For each instance i: $z_i \leftarrow \text{argmin}_k Q_i(k)$.
   - For each cluster k: $\boldsymbol{\mu}_k \leftarrow \bar{\boldsymbol{x}}_k$.
   - Update $\{\boldsymbol{\eta}_k\}_{k=1}^K$ using data augmentation.
2. Deterministic loss function (via SVA on posterior):

$$\mathcal{L}(\boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\eta}) = \underbrace{\sum_{k=1}^{K} \frac{\|\boldsymbol{\eta}_k\|^2}{2\nu^2}}_{\text{Regularization}} + \underbrace{2c \sum_{i=1}^{n} (\zeta_i^{z_i})_+}_{\text{Classification}} + \underbrace{s \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\mu}_{z_i}\|^2}_{\text{Clustering}} + \underbrace{\lambda \cdot K}_{\text{AIC}}.$$

3. Extension to exponential family distributions and multi-class classification.

The M^2DPM Algorithm

1. Repeat until convergence:
   - For each instance i: $z_i \leftarrow \text{argmin}_k Q_i(k)$.
   - For each cluster k: $\boldsymbol{\mu}_k \leftarrow \bar{\boldsymbol{x}}_k$.
   - Update $\{\boldsymbol{\eta}_k\}_{k=1}^K$ using data augmentation.
2. Deterministic loss function (via SVA on posterior):

$$\mathcal{L}(\boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\eta}) = \underbrace{\sum_{k=1}^{K} \frac{\|\boldsymbol{\eta}_k\|^2}{2\nu^2}}_{\text{Regularization}} + \underbrace{2c \sum_{i=1}^{n} (\zeta_i^{z_i})_+}_{\text{Classification}} + \underbrace{s \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\mu}_{z_i}\|^2}_{\text{Clustering}} + \underbrace{\lambda \cdot K}_{\text{AIC}}.$$

3. Extension to exponential family distributions and multi-class classification.

Outline

Datasets and Algorithms

1. Protein fold classification
   - 27 classes, 696 instances, 21 features
2. Pakinson's disease detection
   - 2 classes, 195 instances, 22 features
3. Algorithms
   - Classification models: MNL, Linear-SVM, RBF-SVM
   - Hybrid models: dpMNL (*Shahbaba et al., 2009*), DP+SVM, Gibbs-iSVM, M^2DPM.

Classification performance

	Protein		Parkinson	
Algo	F1 (%)	Times (s)	Acc. (%)	Times (s)
MNL	41.2	2.9	85.6	0.1
L-SVM	47.3	0.5	87.2	0.1
RBF-SVM	**49.5**	1.6	87.2	0.1
dpMNL	**49.5**	98.2	87.7	22.2
DP+SVM	47.9	0.2	86.2	0.1
Gibbs-iSVM	**50.1**	223.4	**88.9**	1.8
M^2DPM	**49.9**	8.1	**88.7**	0.1

Clustering performance

1. Nonparametric clustering on synthetic datasets: K_0 *vs.* K.

n_0	100	300	1000	3000	10000
K_0	8	9	11	13	14
K	8	8	11	12	14

2. Clustering on potential Parkinson's disease patients.

Group	I	II	III	IV	V
Avg. age	65.9	67.0	65.3	77.0	65.4
Avg. stage (0-4)	1.7	1.7	1.3	2.3	1.5

Clustering performance

1. Nonparametric clustering on synthetic datasets: K_0 *vs.* K.

n_0	100	300	1000	3000	10000
K_0	8	9	11	13	14
K	8	8	11	12	14

2. Clustering on potential Parkinson's disease patients.

Group	I	II	III	IV	V
Avg. age	65.9	67.0	65.3	77.0	65.4
Avg. stage (0-4)	1.7	1.7	1.3	2.3	1.5

Summary

- Infinite SVM: clustering + classification
- Small variance analysis
- M^2DPM: fast and accurate.

Summary

- Infinite SVM: clustering + classification
- Small variance analysis
- M^2DPM: fast and accurate.

Summary

- Infinite SVM: clustering + classification
- Small variance analysis
- M^2DPM: fast and accurate.

Questions

The data augmentation trick

- Data augmentation for SVM classifiers (*Polson and Scott, 2011*)

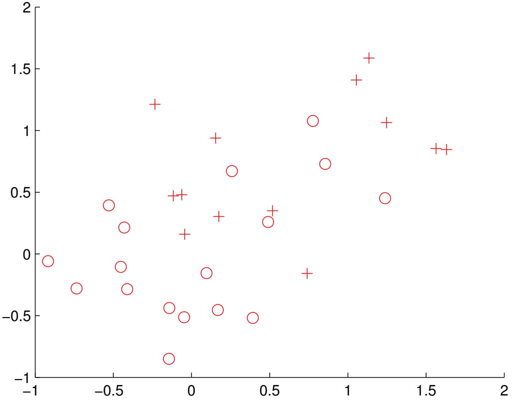
$$
\begin{aligned}
\phi(y_i|z_i = k, \boldsymbol{\eta}) &= \exp(-2c\max(0, 1 - y_i\boldsymbol{\eta}_k^\top \boldsymbol{x}_i)) \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi\omega_i}} \exp\left(-\frac{(\omega_i + c\zeta_i^k)^2}{2\omega_i}\right) \mathrm{d}\omega_i \\
&= \int_0^\infty \phi(y_i, \omega_i|z_i = k, \boldsymbol{\eta}) \mathrm{d}\omega_i
\end{aligned}
$$

Extension to multi-class classification

- Multi-class hinge loss: (*Crammer and Singer, 2001*)

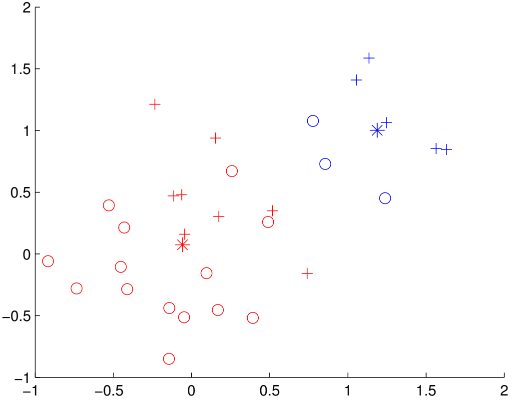
$$\phi^m(y_i | z_i = k, \boldsymbol{\eta}) = \exp(-2c \max_y(\delta_{y,y_i} + \boldsymbol{\eta}_{k,y}^\top \boldsymbol{x}_i - \boldsymbol{\eta}_{k,y_i}^\top \boldsymbol{x}_i))$$

1. Repeat until convergence:
   - For each instance i: $z_i \leftarrow \text{argmin}_k Q_i(k)$.
   - For each cluster k: $\mu_k \leftarrow \bar{\mathbf{x}}_k$.
   - Update $\{\eta_k\}_{k=1}^K$ using data augmentation.



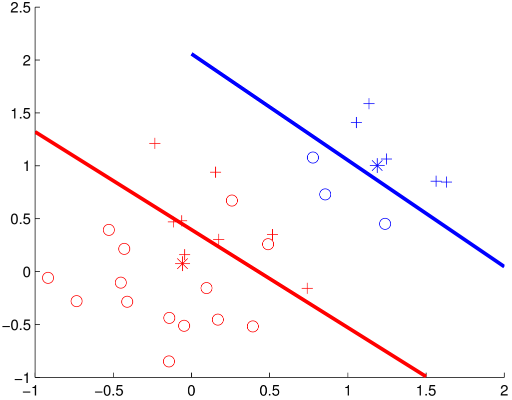
The M²DPM Algorithm

1. Repeat until convergence:
   - **For each instance** i**:** $z_i \leftarrow \textbf{argmin}_k Q_i(k)$**.**
   - **For each cluster** k**:** $\mu_k \leftarrow \bar{\mathbf{x}}_k$**.**
   - Update $\{\eta_k\}_{k=1}^K$ using data augmentation.



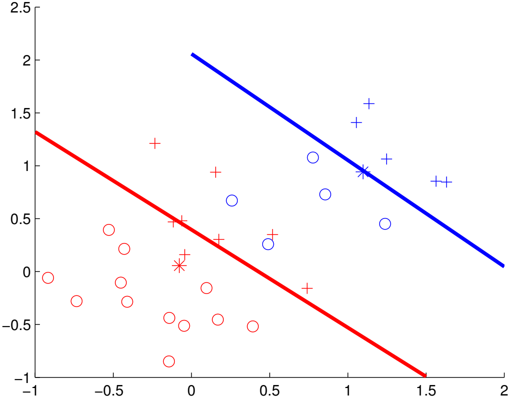
The M^{2}DPM Algorithm

1. Repeat until convergence:
   - For each instance i: $z_i \leftarrow \text{argmin}_k Q_i(k)$.
   - For each cluster k: $\mu_k \leftarrow \bar{\mathbf{x}}_k$.
   - **Update $\{\eta_k\}_{k=1}^K$ using data augmentation.**



The M²DPM Algorithm

1. Repeat until convergence:
   - **For each instance** i**:** $z_i \leftarrow \textbf{argmin}_k Q_i(k)$**.**
   - **For each cluster** k**:** $\mu_k \leftarrow \bar{\mathbf{x}}_k$**.**
   - Update $\{\eta_k\}_{k=1}^{K}$ using data augmentation.



The M^{2}DPM Algorithm

1. Repeat until convergence:
   - For each instance i: $z_i \leftarrow \text{argmin}_k Q_i(k)$.
   - For each cluster k: $\boldsymbol{\mu}_k \leftarrow \bar{\boldsymbol{x}}_k$.
   - Update $\{\boldsymbol{\eta}_k\}_{k=1}^K$ using data augmentation.