Random Matrix Analysis to Balance between Supervised and Unsupervised Learning

Vasilii Feofanov*, Malik Tiomoko*, Aladin Virmaux*

Huawei Paris Noah's Ark Lab
firstname.lastname@huawei.com
*equal contribution

CAp RFIAP 2024, July 2
Published in ICML 2023

Introduction

In some applications data acquisition is cheaper than labeling,

Introduction

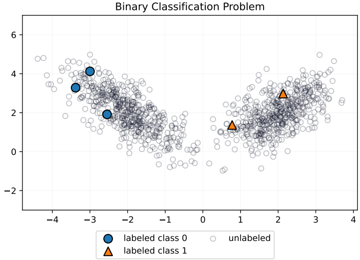
And supervised learning is inefficient.

Introduction

Semi-supervised learning: learn with both few labeled and many unlabeled training examples.

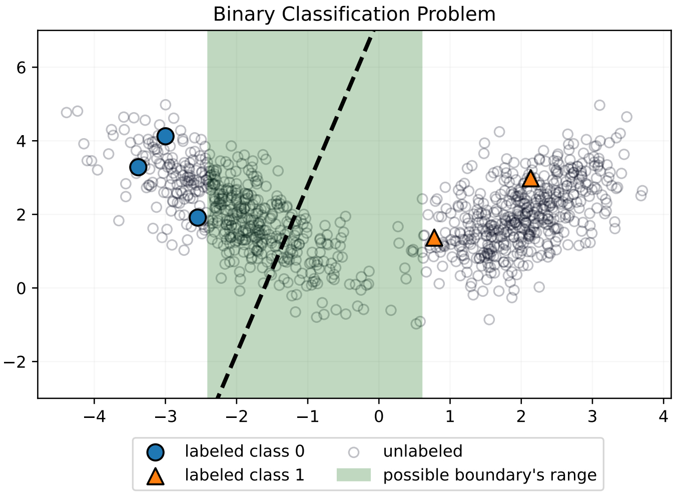
Family of SSL Methods:

- **Pseudo-labeling**,
- Graph-based algorithms,
- Cluster-then-label,
- Unsupervised feature learning.



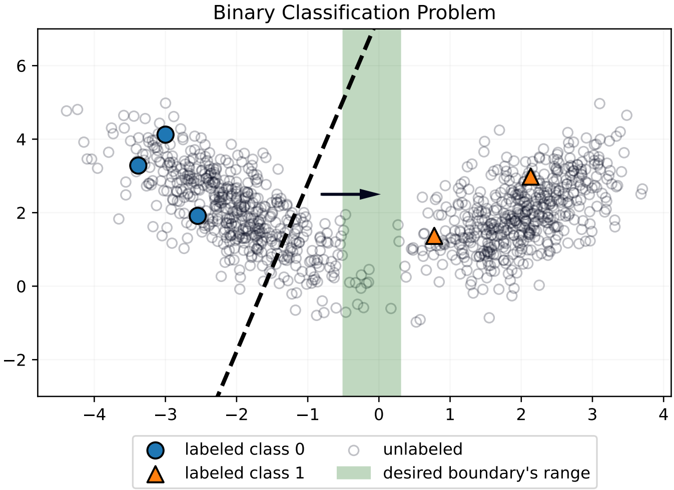
Low Density Separation

Range of possible supervised classifiers is vast: we need to make assumptions.



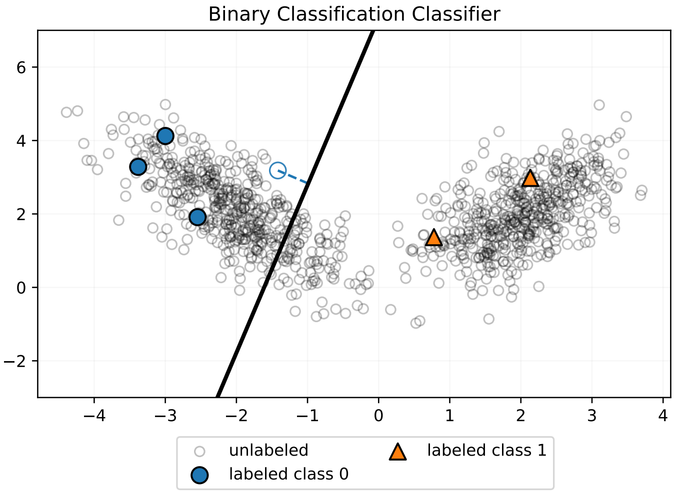
Low Density Separation

Low Density Separation (LDS) assumption: decision boundary is far away from dense regions of unlabeled data.



Pseudo-labeling

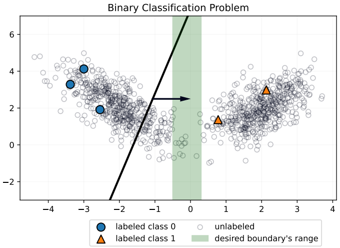
Implementation of LDS: push the boundary away from unlabeled data with high confident scores.



We introduce the QLDS classifier defined by solving:

$$\underset{\omega}{\operatorname{argmin}} \underbrace{\frac{\alpha_\ell}{2}\sum_{i=1}^{n_\ell}\left(y_i - \frac{\omega^\top \mathbf{x}_i}{\sqrt{n}}\right)^2}_{\text{labeled loss}} - \underbrace{\frac{\alpha_u}{2}\sum_{i=n_\ell+1}^{n_\ell+n_u}\left(\frac{\omega^\top \mathbf{x}_i}{\sqrt{n}}\right)^2}_{\text{unlabeled LDS loss}} + \underbrace{\frac{\lambda}{2}\|\omega\|^2}_{\text{regularization}}$$

- Linear classification:
  for $\mathbf{x}$, output $\operatorname{sign}(\omega^\top \mathbf{x})$;

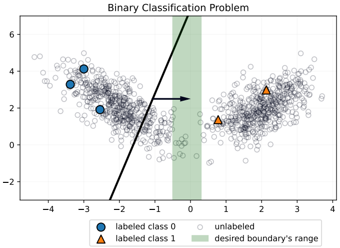


QLDS Classifier

We introduce the QLDS classifier defined by solving:

$$\operatorname*{argmin}_{\omega} \underbrace{\frac{\alpha_\ell}{2}\sum_{i=1}^{n_\ell}\left(y_i - \frac{\omega^\top \mathbf{x}_i}{\sqrt{n}}\right)^2}_{\text{labeled loss}} - \underbrace{\frac{\alpha_u}{2}\sum_{i=n_\ell+1}^{n_\ell+n_u}\left(\frac{\omega^\top \mathbf{x}_i}{\sqrt{n}}\right)^2}_{\text{unlabeled LDS loss}} + \underbrace{\frac{\lambda}{2}\|\omega\|^2}_{\text{regularization}}$$

- Linear classification:
  for $\mathbf{x}$, output $\text{sign}(\omega^\top \mathbf{x})$;
- Square margin maximization;

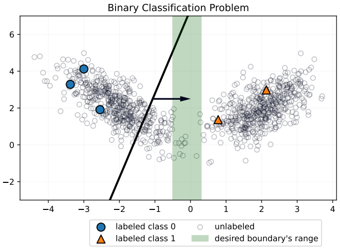


QLDS Classifier

We introduce the QLDS classifier defined by solving:

$$\underset{\omega}{\operatorname{argmin}} \underbrace{\frac{\alpha_\ell}{2}\sum_{i=1}^{n_\ell}\left(y_i - \frac{\omega^\top \mathbf{x}_i}{\sqrt{n}}\right)^2}_{\text{labeled loss}} - \underbrace{\frac{\alpha_u}{2}\sum_{i=n_\ell+1}^{n_\ell+n_u}\left(\frac{\omega^\top \mathbf{x}_i}{\sqrt{n}}\right)^2}_{\text{unlabeled LDS loss}} + \underbrace{\frac{\lambda}{2}\|\omega\|^2}_{\text{regularization}}$$

- Linear classification: for $\mathbf{x}$, output sign$(\omega^\top \mathbf{x})$;
- Square margin maximization;
- Hyperparameters α_ℓ, α_u, λ to balance the components.



QLDS vs TSVM

$$\underset{\omega}{\operatorname{argmin}} \underbrace{\frac{\alpha_\ell}{2}\sum_{i=1}^{n_\ell}\left(y_i - \frac{\omega^\top \mathbf{x}_i}{\sqrt{n}}\right)^2}_{\text{labeled loss}} - \underbrace{\frac{\alpha_u}{2}\sum_{i=n_\ell+1}^{n_\ell+n_u}\left(\frac{\omega^\top \mathbf{x}_i}{\sqrt{n}}\right)^2}_{\text{unlabeled LDS loss}} + \underbrace{\frac{\lambda}{2}\|\omega\|^2}_{\text{regularization}}$$

Difference between QLDS and Transductive SVM:

- Quadratic loss instead of hinge loss;

QLDS vs TSVM

$$\underset{\omega}{\operatorname{argmin}} \underbrace{\frac{\alpha_\ell}{2}\sum_{i=1}^{n_\ell}\left(y_i - \frac{\omega^\top \mathbf{x}_i}{\sqrt{n}}\right)^2}_{\text{labeled loss}} - \underbrace{\frac{\alpha_u}{2}\sum_{i=n_\ell+1}^{n_\ell+n_u}\left(\frac{\omega^\top \mathbf{x}_i}{\sqrt{n}}\right)^2}_{\text{unlabeled LDS loss}} + \underbrace{\frac{\lambda}{2}\|\omega\|^2}_{\text{regularization}}$$

Difference between QLDS and Transductive SVM:

- Quadratic loss instead of hinge loss;
- Margin squared $(\omega^\top \mathbf{x})^2$ instead of $|\omega^\top \mathbf{x}|$, regularization instead of pseudo-labeling loss;

QLDS vs TSVM

$$\underset{\omega}{\operatorname{argmin}} \underbrace{\frac{\alpha_\ell}{2}\sum_{i=1}^{n_\ell}\left(y_i - \frac{\omega^\top \mathbf{x}_i}{\sqrt{n}}\right)^2}_{\text{labeled loss}} - \underbrace{\frac{\alpha_u}{2}\sum_{i=n_\ell+1}^{n_\ell+n_u}\left(\frac{\omega^\top \mathbf{x}_i}{\sqrt{n}}\right)^2}_{\text{unlabeled LDS loss}} + \underbrace{\frac{\lambda}{2}\|\omega\|^2}_{\text{regularization}}$$

Difference between QLDS and Transductive SVM:

- Quadratic loss instead of hinge loss;
- Margin squared $(\omega^\top \mathbf{x})^2$ instead of $|\omega^\top \mathbf{x}|$, regularization instead of pseudo-labeling loss;
- TSVM loss is non-convex and difficult to optimize, QLDS loss is convex and has a closed form solution;

Particular Cases of QLDS

$$\underset{\omega}{\operatorname{argmin}} \underbrace{\frac{\alpha_\ell}{2}\sum_{i=1}^{n_\ell}\left(y_i - \frac{\omega^\top \mathbf{x}_i}{\sqrt{n}}\right)^2}_{\text{labeled loss}} - \underbrace{\frac{\alpha_u}{2}\sum_{i=n_\ell+1}^{n_\ell+n_u}\left(\frac{\omega^\top \mathbf{x}_i}{\sqrt{n}}\right)^2}_{\text{unlabeled LDS loss}} + \underbrace{\frac{\lambda}{2}\|\omega\|^2}_{\text{regularization}}$$

Particular cases of QLDS are:

- $\alpha_u = 0 \Rightarrow$ Least-Square SVM (supervised regime);

Particular Cases of QLDS

$$\underset{\omega}{\operatorname{argmin}} \underbrace{\frac{\alpha_\ell}{2} \sum_{i=1}^{n_\ell} \left(y_i - \frac{\omega^\top \mathbf{x}_i}{\sqrt{n}} \right)^2}_{\text{labeled loss}} - \underbrace{\frac{\alpha_u}{2} \sum_{i=n_\ell+1}^{n_\ell+n_u} \left(\frac{\omega^\top \mathbf{x}_i}{\sqrt{n}} \right)^2}_{\text{unlabeled LDS loss}} + \underbrace{\frac{\lambda}{2} \|\omega\|^2}_{\text{regularization}}$$

Particular cases of QLDS are:

- $\alpha_u = 0 \Rightarrow$ Least-Square SVM (supervised regime);
- $\alpha_l \to 0 \Rightarrow$ Graph-based SSL (Mai, X. and Couillet, R., 2018);

Particular Cases of QLDS

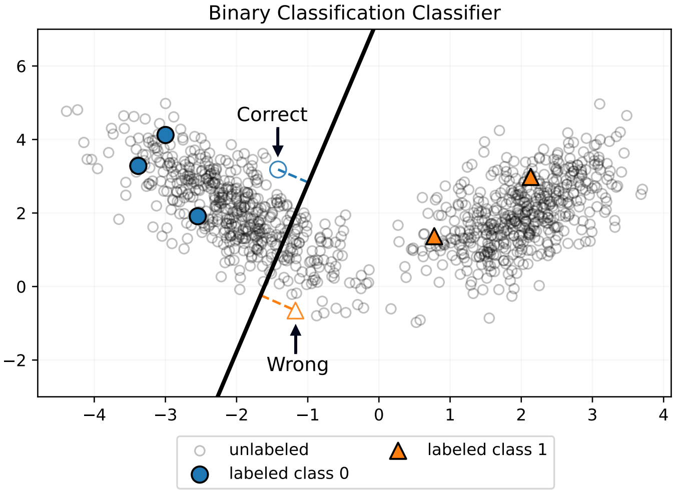
$$\operatorname*{argmin}_{\omega} \underbrace{\frac{\alpha_\ell}{2}\sum_{i=1}^{n_\ell}\left(y_i - \frac{\omega^\top \mathbf{x}_i}{\sqrt{n}}\right)^2}_{\text{labeled loss}} - \underbrace{\frac{\alpha_u}{2}\sum_{i=n_\ell+1}^{n_\ell+n_u}\left(\frac{\omega^\top \mathbf{x}_i}{\sqrt{n}}\right)^2}_{\text{unlabeled LDS loss}} + \underbrace{\frac{\lambda}{2}\|\omega\|^2}_{\text{regularization}}$$

Particular cases of QLDS are:

- $\alpha_u = 0 \Rightarrow$ Least-Square SVM (supervised regime);
- $\alpha_l \to 0 \Rightarrow$ Graph-based SSL (Mai, X. and Couillet, R., 2018);
- $\alpha_l \to 0$ and $\lambda \to \lambda_{\max}(\mathbf{X}_u) \Rightarrow$ Linear spectral clustering.

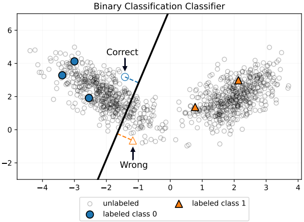
Research Questions

Problem: It is not safe since prediction can be wrong.



Problem: It is not safe since prediction can be wrong.

$$\frac{\alpha_\ell}{2}\sum_{i=1}^{n_\ell}\left(y_i-\frac{\boldsymbol{\omega}^\top \mathbf{x}_i}{\sqrt{n}}\right)^2-\frac{\alpha_u}{2}\sum_{i=n_\ell+1}^{n_\ell+n_u}\left(\frac{\boldsymbol{\omega}^\top \mathbf{x}_i}{\sqrt{n}}\right)^2+\frac{\lambda}{2}\|\boldsymbol{\omega}\|$$



Questions:

1. What are generalization guarantees of the classifier?
2. How to properly choose the hyperparameters?

Random Matrix Analysis

Three types of asymptotic analysis:

- Traditional: d is fixed, n is large ($n \gg d, n \to \infty$);

Three types of asymptotic analysis:

- Traditional: d is fixed, n is large ($n \gg d, n \to \infty$);
- Small-data: n is small, d is large, ($n \ll d$, $d \to \infty$);

Random Matrix Analysis

Three types of asymptotic analysis:

- Traditional: d is fixed, n is large ($n \gg d, n \to \infty$);
- Small-data: n is small, d is large, ($n \ll d$, $d \to \infty$);
- Large-dimensional: n, d are both large ($d = \mathcal{O}(n)$, $(n, d) \to \infty$).

Random Matrix Analysis

Three types of asymptotic analysis:

- Traditional: d is fixed, n is large ($n \gg d, n \to \infty$);
- Small-data: n is small, d is large, ($n \ll d$, $d \to \infty$);
- Large-dimensional: n, d are both large ($d = \mathcal{O}(n)$, $(n, d) \to \infty$).

Assumptions on data distribution:

- Gaussian Mixture Model (GMM);

Random Matrix Analysis

Three types of asymptotic analysis:

- Traditional: d is fixed, n is large ($n \gg d, n \to \infty$);
- Small-data: n is small, d is large, ($n \ll d$, $d \to \infty$);
- Large-dimensional: n, d are both large ($d = \mathcal{O}(n)$, $(n, d) \to \infty$).

Assumptions on data distribution:

- Gaussian Mixture Model (GMM);
- Concentrated Data (Louart, C. and Couillet, R., 2018): variance of $\boldsymbol{\omega}^\top \mathbf{x}$ does not grow with dimension d,

Random Matrix Analysis

Three types of asymptotic analysis:

- Traditional: d is fixed, n is large ($n \gg d, n \to \infty$);
- Small-data: n is small, d is large, ($n \ll d$, $d \to \infty$);
- Large-dimensional: n, d are both large ($d = \mathcal{O}(n)$, $(n, d) \to \infty$).

Assumptions on data distribution:

- Gaussian Mixture Model (GMM);
- Concentrated Data (Louart, C. and Couillet, R., 2018): variance of $\boldsymbol{\omega}^\top \mathbf{x}$ does not grow with dimension d,
  Particular cases:
  - Standard Gaussian distribution,
  - Lipschitz transformation of Gaussian (e.g., GAN images),
  - Open question: learned features by DNN?

Random Matrix Theorem

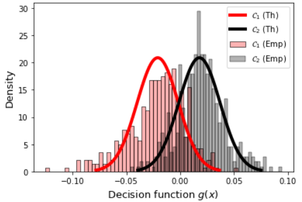
Assumptions:

1. Large-dimensional regime $d = \mathcal{O}(n)$;
2. Concentrated data distribution: variance of $\boldsymbol{\omega}^\top \mathbf{x}$ does not grow with dimension d.

Theorem

Under the assumptions, we have:

- $\omega^\top \mathbf{x} | y = -1$ *and* $\omega^\top \mathbf{x} | y = +1$ *are asymptotically normally distributed with known parameters;*
- *Classification error is explicitly evaluated;*



- *The classification problem concentrates into two-dimensional sufficient statistics.*

Comments on Theorem

- Sketch of Proof.
  - CLT: if $\mathbf{x}$ is a concentrated random vector, then $\boldsymbol{\omega}^\top \mathbf{x}$ is asymptotically Gaussian,
  - Following Marchenko, V. A. and Pastur, L. A. (1967), for each class $\mathcal{C}_j$, we compute:

$$\mathbb{E}_{\mathbf{X}_\ell, \mathbf{X}_u} \left[(\boldsymbol{\omega}^*(\mathbf{X}_\ell, \mathbf{X}_u))^\top \mathbf{x} \middle| \mathbf{x} \in \mathcal{C}_j \right],$$
$$\mathrm{Var}_{\mathbf{X}_\ell, \mathbf{X}_u} \left[(\boldsymbol{\omega}^*(\mathbf{X}_\ell, \mathbf{X}_u))^\top \mathbf{x} \middle| \mathbf{x} \in \mathcal{C}_j \right].$$

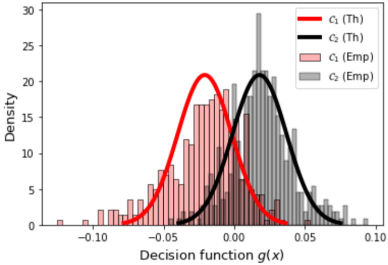
Comments on Theorem

- Sketch of Proof.
  - CLT: if $\mathbf{x}$ is a concentrated random vector, then $\boldsymbol{\omega}^\top \mathbf{x}$ is asymptotically Gaussian,
  - Following Marchenko, V. A. and Pastur, L. A. (1967), for each class $\mathcal{C}_j$, we compute:

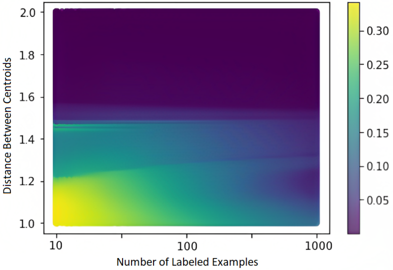
$$\mathbb{E}_{\mathbf{X}_\ell, \mathbf{X}_u}\left[(\boldsymbol{\omega}^*(\mathbf{X}_\ell, \mathbf{X}_u))^\top \mathbf{x} \middle| \mathbf{x} \in \mathcal{C}_j \right],$$
$$\mathrm{Var}_{\mathbf{X}_\ell, \mathbf{X}_u}\left[(\boldsymbol{\omega}^*(\mathbf{X}_\ell, \mathbf{X}_u))^\top \mathbf{x} \middle| \mathbf{x} \in \mathcal{C}_j \right].$$

- In the final expression, some quantities depend on class mean and covariance and need to be estimated from data.
  - For example, $[\mu_{-1}, \mu_{+1}]^\top [\mu_{-1}, \mu_{+1}]$ is better to estimate directly rather than estimating μ_{-1}, μ_{+1} separately.

Illustration of Theorem



(a) Theoretical vs Empirical Distribution.



(b) Utility of Unlabeled Data.

- Theory can fit the empirical distribution of $\boldsymbol{\omega}^\top \mathbf{x}$.
- The theoretical expression can be viewed as a function of different variables: $n_\ell, \alpha_l, \alpha_u$, etc.

Application to Model Selection

- Find α_ℓ and α_u automatically based on asymptotic error given by Theorem.

Application to Model Selection

- Find α_ℓ and α_u automatically based on asymptotic error given by Theorem.
- Experimental results showed that

Data set	Baselines		Model Selection	
	`QLDS(1,0)` `(LS-SVM)`	`QLDS(0,1)` `(Graph SSL)`	`QLDS(cv)`	`QLDS(th)`
`books`	$37.47^{\downarrow} \pm 2.25$	26.47 ± 0.72	27.91 ± 3.32	$\mathbf{26.03} \pm 0.79$
`dvd`	$38.33^{\downarrow} \pm 1.72$	29.12 ± 1.35	29.53 ± 3.48	$\mathbf{28.53} \pm 1.33$
`electronics`	$34.15^{\downarrow} \pm 3.25$	19.4 ± 0.29	$20.1^{\downarrow} \pm 1.03$	$\mathbf{19.41} \pm 0.46$
`kitchen`	$32.39^{\downarrow} \pm 3.02$	19.31 ± 0.16	$19.98^{\downarrow} \pm 2.28$	$\mathbf{19.11} \pm 0.32$
`splice`	$39.81^{\downarrow} \pm 2.93$	35.48 ± 0.86	37.02 ± 3.04	$\mathbf{35.35} \pm 1.26$
`adult`	33.35 ± 0.68	$36.28^{\downarrow} \pm 0.06$	$\mathbf{32.25} \pm 1.92$	32.88 ± 2.46
`mushrooms`	$6.55^{\downarrow} \pm 2.07$	$11.33^{\downarrow} \pm 0.04$	$\mathbf{2.57} \pm 1.86$	$8.49^{\downarrow} \pm 3.63$

- Find α_ℓ and α_u automatically based on asymptotic error given by Theorem.
- Experimental results showed that

Data set	Baselines		Model Selection	
	`QLDS(1,0)` `(LS-SVM)`	`QLDS(0,1)` `(Graph SSL)`	`QLDS(cv)`	`QLDS(th)`
`books`	$37.47^{\downarrow} \pm 2.25$	26.47 ± 0.72	27.91 ± 3.32	$\mathbf{26.03} \pm 0.79$
`dvd`	$38.33^{\downarrow} \pm 1.72$	29.12 ± 1.35	29.53 ± 3.48	$\mathbf{28.53} \pm 1.33$
`electronics`	$34.15^{\downarrow} \pm 3.25$	19.4 ± 0.29	$20.1^{\downarrow} \pm 1.03$	$\mathbf{19.41} \pm 0.46$
`kitchen`	$32.39^{\downarrow} \pm 3.02$	19.31 ± 0.16	$19.98^{\downarrow} \pm 2.28$	$\mathbf{19.11} \pm 0.32$
`splice`	$39.81^{\downarrow} \pm 2.93$	35.48 ± 0.86	37.02 ± 3.04	$\mathbf{35.35} \pm 1.26$
`adult`	33.35 ± 0.68	$36.28^{\downarrow} \pm 0.06$	$\mathbf{32.25} \pm 1.92$	32.88 ± 2.46
`mushrooms`	$6.55^{\downarrow} \pm 2.07$	$11.33^{\downarrow} \pm 0.04$	$\mathbf{2.57} \pm 1.86$	$8.49^{\downarrow} \pm 3.63$

  - Model selection outperforms both LS-SVM and Graph-based SSL;

Application to Model Selection

- Find α_ℓ and α_u automatically based on asymptotic error given by Theorem.
- Experimental results showed that

Data set	Baselines		Model Selection	
	QLDS (1,0) (LS-SVM)	QLDS (0,1) (Graph SSL)	QLDS (cv)	QLDS (th)
books	$37.47^{\downarrow} \pm 2.25$	26.47 ± 0.72	27.91 ± 3.32	$\mathbf{26.03} \pm 0.79$
dvd	$38.33^{\downarrow} \pm 1.72$	29.12 ± 1.35	29.53 ± 3.48	$\mathbf{28.53} \pm 1.33$
electronics	$34.15^{\downarrow} \pm 3.25$	19.4 ± 0.29	$20.1^{\downarrow} \pm 1.03$	$\mathbf{19.41} \pm 0.46$
kitchen	$32.39^{\downarrow} \pm 3.02$	19.31 ± 0.16	$19.98^{\downarrow} \pm 2.28$	$\mathbf{19.11} \pm 0.32$
splice	$39.81^{\downarrow} \pm 2.93$	35.48 ± 0.86	37.02 ± 3.04	$\mathbf{35.35} \pm 1.26$
adult	33.35 ± 0.68	$36.28^{\downarrow} \pm 0.06$	$\mathbf{32.25} \pm 1.92$	32.88 ± 2.46
mushrooms	$6.55^{\downarrow} \pm 2.07$	$11.33^{\downarrow} \pm 0.04$	$\mathbf{2.57} \pm 1.86$	$8.49^{\downarrow} \pm 3.63$

  - Model selection outperforms both LS-SVM and Graph-based SSL;
  - Selecting α_ℓ and α_u by cross-validation is more costly and can lead to over-confidence towards labeled data.

Thanks for your attention !