Semi-supervised Learning via Transductive Inference

HML Reading Group: Session 1

May 9, 2019

Vasilii Feofanov

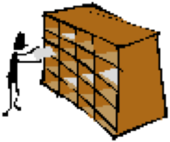
Université Grenoble Alpes

vasilii.feofanov@univ-grenoble-alpes.fr

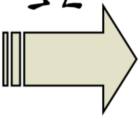
Outline

In many applications, labeling examples is prohibitive while huge number of unlabeled data are available.

$$\mathrm{Z}_{\mathcal{L}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$$



$$\mathrm{X}_{\mathcal{U}} = \{\mathbf{x}_i\}_{i=l+1}^{l+u}$$



Classifier:
$\mathcal{X} \to \mathcal{Y}$

Goal:
Small classification error

Introduction

- **Supervised Learning:**
  Labeled data $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$.

- **Supervised Learning:**
  Labeled data $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$.

- **Unsupervised Learning:**
  Unlabeled data $\{\mathbf{x}_i\}_{i=1}^{u}$.

- **Supervised Learning:**
  Labeled data $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$.

- **Semi-supervised Learning:**
  Both labeled $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ and unlabeled data $\{\mathbf{x}'_i\}_{i=l+1}^{l+u}$

- **Unsupervised Learning:**
  Unlabeled data $\{\mathbf{x}_i\}_{i=1}^{u}$.

- **Supervised Learning:**
  Labeled data $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$.

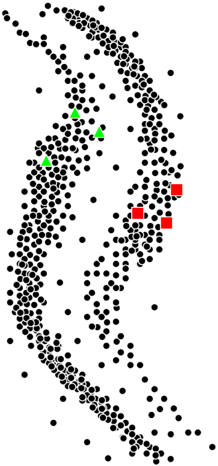
$\Downarrow$

- **Semi-supervised Learning:**
  Both labeled $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ and unlabeled data $\{\mathbf{x}'_i\}_{i=l+1}^{l+u}$

$\Uparrow$

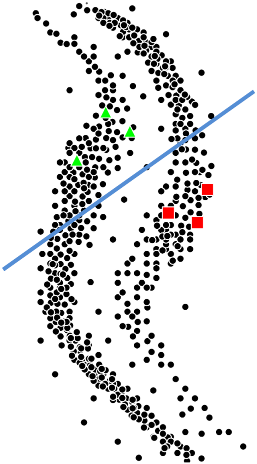
- **Unsupervised Learning:**
  Unlabeled data $\{\mathbf{x}_i\}_{i=1}^{u}$.

Supervised vs Semi-supervised Learning



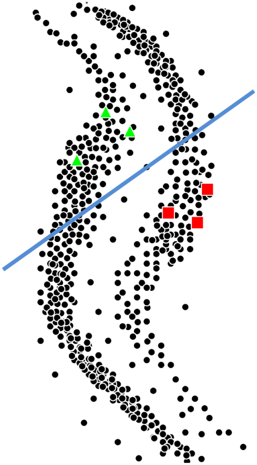
Example of partially labeled data

Supervised vs Semi-supervised Learning

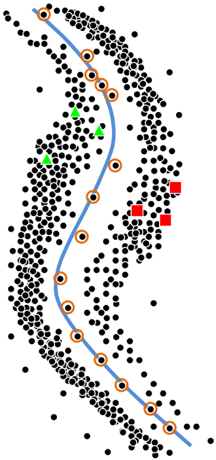


(a) Supervised classifier

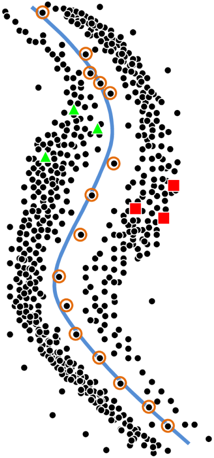
Supervised vs Semi-supervised Learning



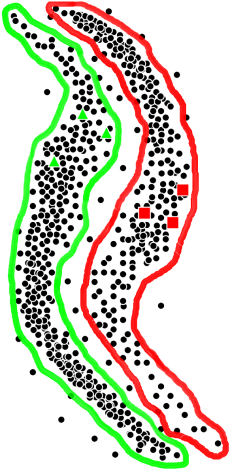
(a) Supervised classifier



(b) Semi-supervised classifier



(a) Low density separation



(b) Cluster assumption

Outline

SSL Classifiers

- Generative models:
  - Semi-supervised CEM [McLachlan, 1992]
  - Semi-supervised logistic regression [Amini and Gallinari, 2002]
  - Deep generative models [Kingma et al., 2014]

SSL Classifiers

- Generative models:
  - Semi-supervised CEM [McLachlan, 1992]
  - Semi-supervised logistic regression [Amini and Gallinari, 2002]
  - Deep generative models [Kingma et al., 2014]
- Graph-based algorithms:
  - Label propagation [Zhu and Ghahramani, 2002]
  - Label spreading [Zhou et al., 2004]

SSL Classifiers

- Generative models:
  - Semi-supervised CEM [McLachlan, 1992]
  - Semi-supervised logistic regression [Amini and Gallinari, 2002]
  - Deep generative models [Kingma et al., 2014]
- Graph-based algorithms:
  - Label propagation [Zhu and Ghahramani, 2002]
  - Label spreading [Zhou et al., 2004]
- Transductive Learning:
  - Transductive support vector machine [Joachims, 1999]
  - **Self-learning algorithm**
    [Tür et al., 2005, Amini et al., 2008, Feofanov et al., 2019]

Transductive Inference

"When solving a problem of interest, do not solve a more general problem as an intermediate step."

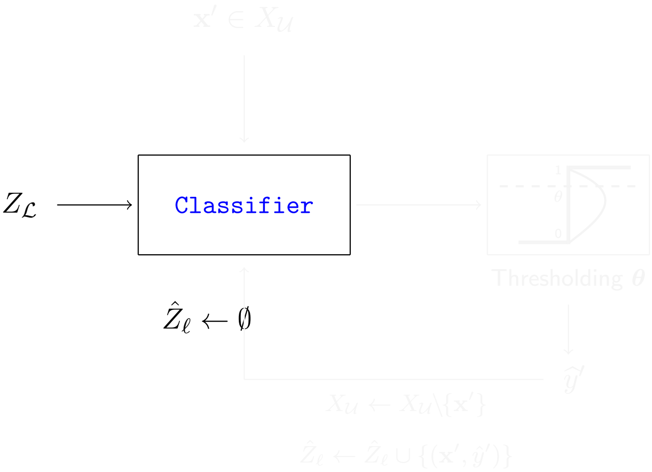
— Vladimir Vapnik

Transductive Inference

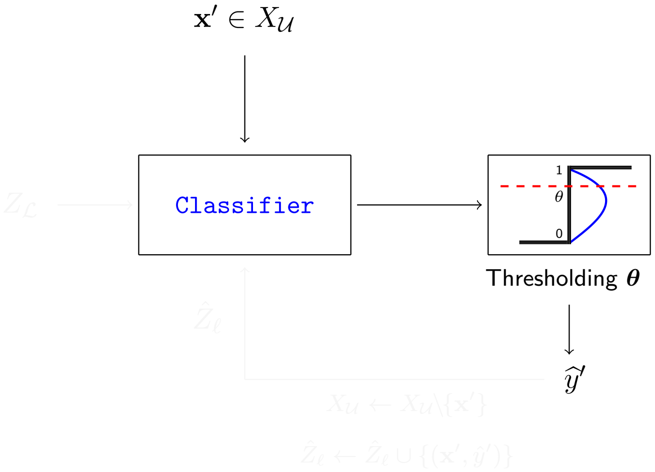
> *"When solving a problem of interest, do not solve a more general problem as an intermediate step."*
>
> — Vladimir Vapnik

- In transductive learning we concentrate on the object of interest $\Rightarrow$ unlabeled examples.
- Thus, the objective is not the generalization error, but the error computed on the unlabeled examples.

Self-learning Algorithm


$\mathbf{x}' \in X_{\mathcal{U}}$
$Z_{\mathcal{L}}$
Classifier
1
θ
0
Thresholding θ
$\hat{Z}_\ell \leftarrow \emptyset$
$\hat{y}'$
$X_{\mathcal{U}} \leftarrow X_{\mathcal{U}} \backslash \{\mathbf{x}'\}$
$\hat{Z}_\ell \leftarrow \hat{Z}_\ell \cup \{(\mathbf{x}', \hat{y}')\}$

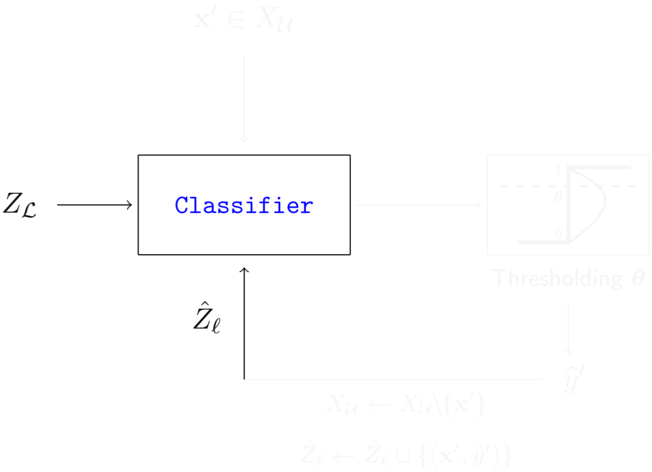
Self-learning Algorithm



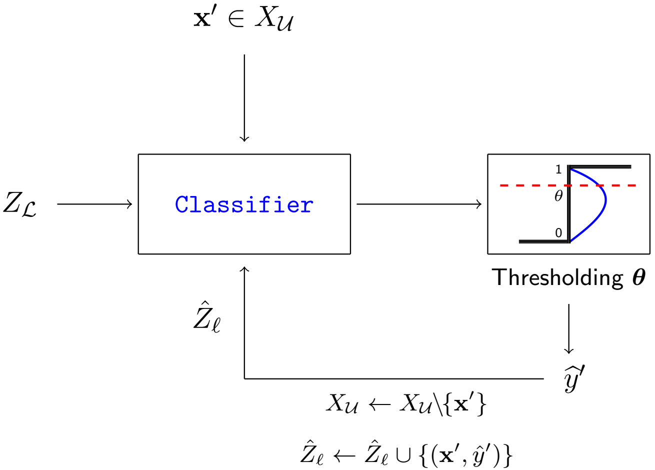
$$X_{\mathcal{U}} \leftarrow X_{\mathcal{U}} \backslash \{\mathbf{x}'\}$$

$$\hat{Z}_{\ell} \leftarrow \hat{Z}_{\ell} \cup \{(\mathbf{x}', \hat{y}')\}$$

Self-learning Algorithm



Self-learning Algorithm



Practical Issues in SSL

In practice, how do we

- Tune hyperparameters?
- Estimate our performance?

Practical Issues in SSL

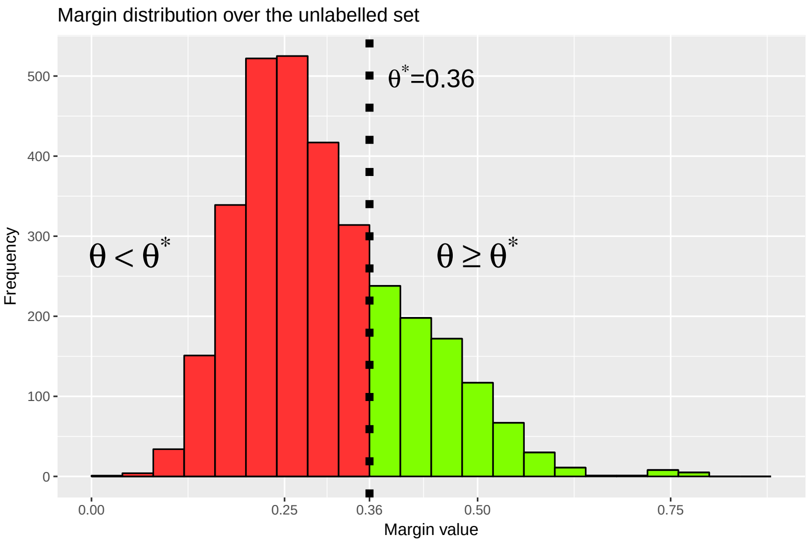
In practice, how do we

- Tune hyperparameters?
- Estimate our performance?

Possible solutions:

- Exhaustive analysis of your problem?
- Look at behavior of the algorithm on different data sets?
- Theoretical study of the algorithm?

Automatic Threshold Finding



Automatic Threshold Finding

We look for θ that minimizes:

$$\mathrm{E}_{\mathcal{U}|\theta}(h) := \frac{\mathrm{E}_{\mathcal{U}\wedge\boldsymbol{\theta}}(h)}{\pi(m(\mathbf{x}') \geq \theta)}.$$

A **trade-off** between:

- Transductive error (bound) on pseudo-labeled examples,
- Proportion of examples in the unlabeled set that will be pseudo-labeled.

We look for θ that minimizes:

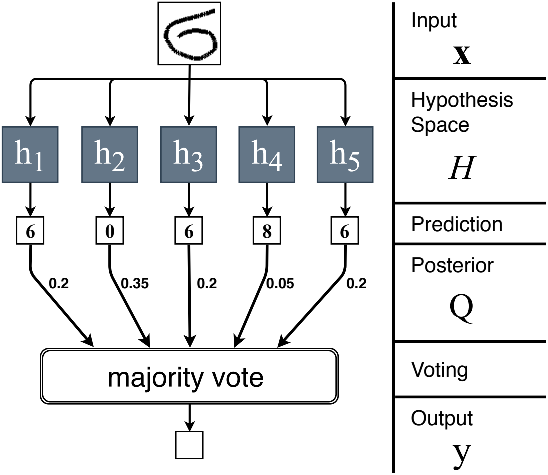
$$\mathrm{E}_{\mathcal{U}|\theta}(h) := \frac{\mathrm{E}_{\mathcal{U}\wedge\boldsymbol{\theta}}(h)}{\pi(m(\mathbf{x}') \geq \theta)}.$$

A **trade-off** between:

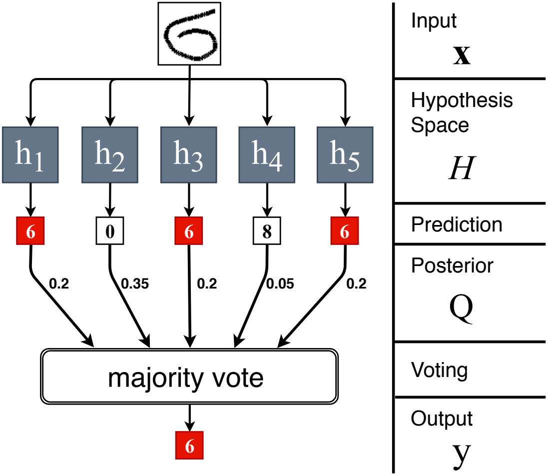
- Transductive bound for the binary majority vote classifier [Amini et al., 2008].
- Extension to the multi-class classification was proposed in [Feofanov et al., 2019].

Outline

$$B_Q(\mathbf{x}) := \operatorname{argmax}_{c \in \mathcal{Y}} \left[\mathbb{E}_{h \sim Q} \mathbb{I}(h(\mathbf{x}) = c) \right]$$

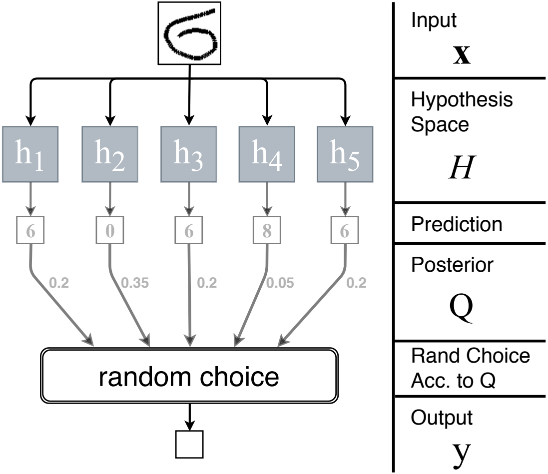


$$B_Q(\mathbf{x}) := \operatorname{argmax}_{c\in\mathcal{Y}} \left[\mathbb{E}_{h\sim Q}\mathbb{I}(h(\mathbf{x}) = c)\right]$$



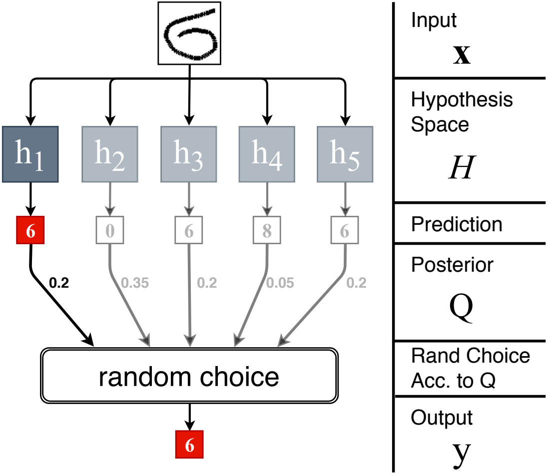
Gibbs Classifier

$G_Q(\mathbf{x}) := \text{rand}_{h \sim Q} h(\mathbf{x})$



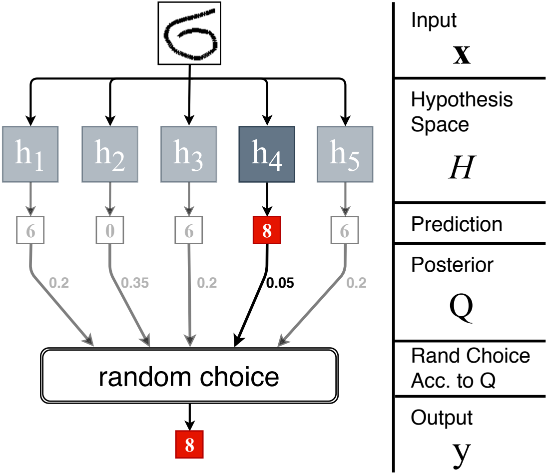
Gibbs Classifier

$G_Q(\mathbf{x}) := \text{rand}_{h\sim Q} h(\mathbf{x})$



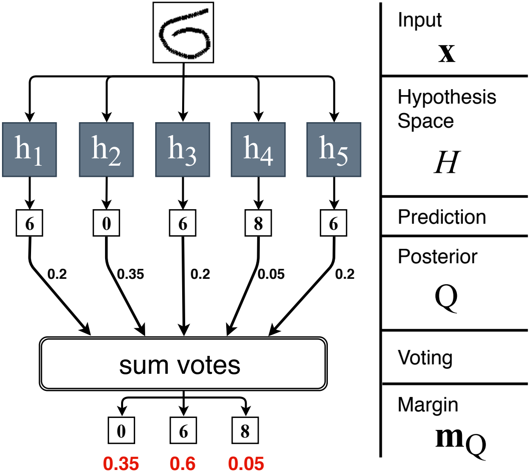
Gibbs Classifier

$G_Q(\mathbf{x}) := \text{rand}_{h \sim Q} h(\mathbf{x})$



Margin: Indicator of Confidence

$$m_Q(\mathbf{x}, c) = \mathbb{E}_{h\sim Q}\mathbb{I}(h(\mathbf{x}) = c)$$



Error Measures

Error rate:

- $\mathtt{E}_{\mathcal{U}}(h) := \frac{1}{u} \sum_{\mathbf{x}' \in \mathrm{X}_{\mathcal{U}}} \mathbb{I}(h(\mathbf{x}') \neq y'),$

Error rate:

- $\mathtt{E}_{\mathcal{U}}(h) := \frac{1}{u} \sum_{\mathbf{x}' \in \mathrm{X}_{\mathcal{U}}} \mathbb{I}(h(\mathbf{x}') \neq y'),$

Conditional risk:

- $R_{\mathcal{U}}(B_Q, i, j) := \frac{1}{u_i} \sum_{\mathbf{x}' \in \mathrm{X}_{\mathcal{U}}} \mathbb{I}(B_Q(\mathbf{x}') = j) \mathbb{I}(y' = i),$
- $R_{\mathcal{U}}(G_Q, i, j) := \frac{1}{u_i} \sum_{\mathbf{x}' \in \mathrm{X}_{\mathcal{U}}} \mathbb{E}_{h \sim Q} \mathbb{I}(h(\mathbf{x}') = j) \mathbb{I}(y' = i),$

The error to predict j given class i.

Error Measures

Error rate:

- $\mathtt{E}_{\mathcal{U}}(h) := \frac{1}{u} \sum_{\mathbf{x}' \in \mathrm{X}_{\mathcal{U}}} \mathbb{I}(h(\mathbf{x}') \neq y')$,

Conditional risk:

- $R_{\mathcal{U}}(B_Q, i, j) := \frac{1}{u_i} \sum_{\mathbf{x}' \in \mathrm{X}_{\mathcal{U}}} \mathbb{I}(B_Q(\mathbf{x}') = j)\mathbb{I}(y' = i)$,
- $R_{\mathcal{U}}(G_Q, i, j) := \frac{1}{u_i} \sum_{\mathbf{x}' \in \mathrm{X}_{\mathcal{U}}} \mathbb{E}_{h \sim Q}\mathbb{I}(h(\mathbf{x}') = j)\mathbb{I}(y' = i)$,

Confusion matrix:

- $\mathbf{C}_h^{\mathcal{U}} := (c_{i,j})_{i,j=\{1,\dots,K\}^2}, \quad c_{i,j} = \begin{cases} 0 & i = j \\ R_{\mathcal{U}}(h, i, j) & i \neq j \end{cases}.$
  – [Morvant et al., 2012]

Error Measures

Error rate:

- $\mathrm{E}_{\mathcal{U}}(h) := \frac{1}{u} \sum_{\mathbf{x}' \in \mathrm{X}_{\mathcal{U}}} \mathbb{I}(h(\mathbf{x}') \neq y')$,

Conditional risk:

- $R_{\mathcal{U}}(B_Q, i, j) := \frac{1}{u_i} \sum_{\mathbf{x}' \in \mathrm{X}_{\mathcal{U}}} \mathbb{I}(B_Q(\mathbf{x}') = j)\mathbb{I}(y' = i)$,
- $R_{\mathcal{U}}(G_Q, i, j) := \frac{1}{u_i} \sum_{\mathbf{x}' \in \mathrm{X}_{\mathcal{U}}} \mathbb{E}_{h \sim Q}\mathbb{I}(h(\mathbf{x}') = j)\mathbb{I}(y' = i)$,

Confusion matrix:

- $\mathbf{C}_h^{\mathcal{U}} := (c_{i,j})_{i,j=\{1,\dots,K\}^2}, \quad c_{i,j} = \begin{cases} 0 & i = j \\ R_{\mathcal{U}}(h, i, j) & i \neq j \end{cases}.$

Joint conditional risk:

- $R_{\mathcal{U} \wedge \boldsymbol{\theta}}(B_Q, i, j) :=$ $\frac{1}{u_i} \sum_{\mathbf{x}' \in \mathrm{X}_{\mathcal{U}}} \mathbb{I}(B_Q(\mathbf{x}') = j)\mathbb{I}(y' = i)\mathbb{I}(m_Q(\mathbf{x}', j) \geq \theta_j)$, – risk to have the conditional error **and** the margin above θ_j

Remark

The error rate and the confusion matrix are connected in the following way:

$$\mathrm{E}_{\mathcal{U}}(h) = \left\| (\mathbf{C}_h^{\mathcal{U}})^\intercal \mathbf{p} \right\|_1 ,$$

where $\mathbf{p} = \{u_i/u\}_{i=1}^{K}$.

A Transductive Bound for the Conditional Risk

Theorem

$\forall\ Q$ and $\forall \delta \in (0,1]$, $\forall \boldsymbol{\theta} \in [0,1]^K$ with prob. at least $1-\delta$:

$$R_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q, i, j) \leq \inf_{\gamma \in [\theta_j, 1]} \left\{ I_{i,j}^{(\leq,<)}(\theta_j, \gamma) + \frac{1}{\gamma} \Big\lfloor \left(K_{i,j}^{\delta} - M_{i,j}^{<}(\gamma) + M_{i,j}^{<}(\theta_j)\right) \Big\rfloor_+ \right\},$$

where

- $K_{i,j}^{\delta} = R_u^{\delta}(G_Q, i, j) - \varepsilon_{i,j}$,
- $R_u^{\delta}(G_Q, i, j)$ *is an upper bound that holds with prob. at least* $1-\delta$.
- $\varepsilon_{i,j}$ *is the average of j-margins in class i and class j is not predicted,*
- $I_{i,j}^{(\leq,<)}(\theta_j, \gamma)$ *is proportion of obs. from i with margin in interval $[\theta_j, \gamma)$,*
- $M_{i,j}^{<}(t)$ *is the average of j-margins in class i that less than t.*

A Transductive Bound for the Conditional Risk

Theorem

$\forall\ Q$ *and* $\forall \delta \in (0,1]$, $\forall \boldsymbol{\theta} \in [0,1]^K$ *with prob. at least* $1-\delta$*:*

$$R_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q, i, j) \leq \inf_{\gamma \in [\theta_j, 1]} \left\{ I_{i,j}^{(\leq,<)}(\theta_j, \gamma) + \frac{1}{\gamma} \Big\lfloor \left(K_{i,j}^{\delta} - M_{i,j}^{<}(\gamma) + M_{i,j}^{<}(\theta_j) \right) \Big\rfloor_+ \right\},$$

where

- $K_{i,j}^{\delta} = R_u^{\delta}(G_Q, i, j) - \varepsilon_{i,j}$,
- $R_u^{\delta}(G_Q, i, j)$ *is an upper bound that holds with prob. at least* $1-\delta$.
- $\varepsilon_{i,j}$ *is the average of* j*-margins in class* i *and class* j *is not predicted,*
- $I_{i,j}^{(\leq,<)}(\theta_j, \gamma)$ *is proportion of obs. from* i *with margin in interval* $[\theta_j, \gamma)$*,*
- $M_{i,j}^{<}(t)$ *is the average of* j*-margins in class* i *that less than* t.

Proof

- *Bound derived from a solution of a linear program where the error is maximized.*
- *Constraint: connection between* $R_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q, i, j)$ *and* $R_{\mathcal{U}}(G_Q, i, j)$.
- *The solution of linear program is explicit and is computed in practice.*

Theorem: Remarks

Proposition

Suppose

- *The Gibbs conditional risk bound is tight,*
- *The Bayes classifier makes its mistakes mostly on examples with low margins*

$\Rightarrow$ *the bound is* **tight**.

Theorem: Remarks

Proposition

Suppose

- *The Gibbs conditional risk bound is tight,*
- *The Bayes classifier makes its mistakes mostly on examples with low margins*

$\Rightarrow$ *the bound is* **tight**.

Corollary

Let $\mathbf{U}_{\boldsymbol{\theta}}^{\delta} := (R_{\mathcal{U}}^{\delta}(B_Q, i, j))_{\substack{i,j=\{1,\dots,K\}^2 \\ i\neq j}}$,

where $R_{\mathcal{U}}^{\delta}(B_Q, i, j)$ *is defined by Theorem. Then, we have:*

$$\mathbf{E}_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q) \leq \left\| \left(\mathbf{U}_{\boldsymbol{\theta}}^{\delta}\right)^{\intercal} \mathbf{p} \right\|_1,$$

where $\mathbf{p} = \{u_i/u\}_{i=1}^{K}$.

Conditional Bayes Error

We look for $\boldsymbol{\theta}$ that minimizes:

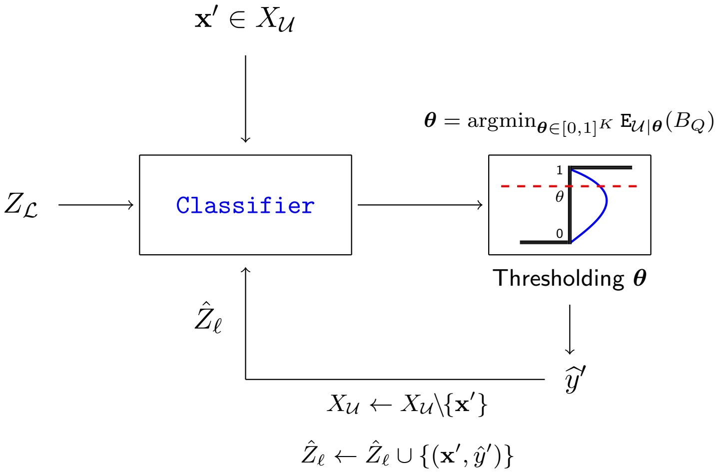
$$\mathrm{E}_{\mathcal{U}|\boldsymbol{\theta}}(B_Q) := \frac{\mathrm{E}_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q)}{\pi(m_Q(\mathbf{x}', B_Q(\mathbf{x}')) \geq \theta_{B_Q(\mathbf{x}')})}.$$

A **trade-off** between:

- Transductive error on pseudo-labeled examples (estimated using **Theorem**),
- Proportion of pseudo-labeled examples in $\mathrm{X}_{\mathcal{U}}$.

Outline

Multi-class Self-learning Algorithm



Experiment Results on Different Data Sets

Data set	Info	Score	RF	LP	OVA-TSVM	FSLA $_{\theta=0.7}$	MSLA
Vowel	$l = 99$ $u = 891$ $d = 10$ $K = 11$	ACC	$.583 \pm .026$	$.577 \pm .027$	NA	$.516^{\downarrow} \pm .043$	$\mathbf{.592} \pm .027$
		F1	$.572 \pm .028$	$.568 \pm .026$	NA	$.493^{\downarrow} \pm .046$	$\mathbf{.580} \pm .030$
DNA	$l = 31$ $u = 3155$ $d = 180$ $K = 3$	ACC	$.693^{\downarrow} \pm .072$	$.538^{\downarrow} \pm .039$	$\mathbf{.812} \pm .039$	$.516^{\downarrow} \pm .09$	$.706^{\downarrow} \pm .083$
		F1	$.65^{\downarrow} \pm .109$	$.535^{\downarrow} \pm .044$	$\mathbf{.812} \pm .038$	$.372^{\downarrow} \pm .096$	$.663^{\downarrow} \pm .118$
Pendigits	$l = 109$ $u = 10883$ $d = 16$ $K = 10$	ACC	$.864^{\downarrow} \pm .022$	$.777^{\downarrow} \pm .052$	$.667^{\downarrow} \pm .023$	$.847^{\downarrow} \pm .035$	$\mathbf{.887} \pm .019$
		F1	$.861^{\downarrow} \pm .025$	$.756^{\downarrow} \pm .069$	$.656^{\downarrow} \pm .021$	$.842^{\downarrow} \pm .042$	$\mathbf{.885} \pm .02$
MNIST	$l = 175$ $u = 69825$ $d = 900$ $K = 10$	ACC	$.865^{\downarrow} \pm .018$	NA	NA	$.8^{\downarrow} \pm .059$	$\mathbf{.909} \pm .018$
		F1	$.863^{\downarrow} \pm .019$	NA	NA	$.774^{\downarrow} \pm .077$	$\mathbf{.909} \pm .018$
SensIT	$l = 49$ $u = 98479$ $d = 100$ $K = 3$	ACC	$.67 \pm .0291$	NA	NA	$.619^{\downarrow} \pm .037$	$\mathbf{.675} \pm .029$
		F1	$.654 \pm .045$	NA	NA	$.578^{\downarrow} \pm .068$	$\mathbf{.66} \pm .042$

Table: Classification performance on 5 data sets.
$\downarrow$: the performance is statistically worse than the best result on the level 0.01 of significance.
NA: the algorithm does not converge.

Questions?

Amini, M., Laviolette, F., and Usunier, N. (2008).
A transductive bound for the voted classifier with an application to semi-supervised learning.
In *Advances in Neural Information Processing Systems (NIPS 21)*, pages 65–72.

Amini, M.-R. and Gallinari, P. (2002).
Semi-supervised logistic regression.
In *Proceedings of the 15th European Conference on Artificial Intelligence*, ECAI'02, pages 390–394, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Feofanov, V., Devijver, E., and Amini, M.-R. (2019).
Transductive bounds for the multi-class majority vote classifier.
In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, (AAAI-19), Honolulu, Hawaii, USA, January 27 - February 1, 2019*.

Joachims, T. (1999).
Transductive inference for text classification using support vector machines.
In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pages 200–209, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Kingma, D. P., Rezende, D. J., Mohamed, S., and Welling, M. (2014).
Semi-supervised learning with deep generative models.
In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3581–3589, Cambridge, MA, USA. MIT Press.

McLachlan, G. J. (1992).
Discriminant Analysis and Statistical Pattern Recognition.
Wiley-Interscience.

References II

Morvant, E., Koço, S., and Ralaivola, L. (2012).
PAC-Bayesian Generalization Bound on Confusion Matrix for Multi-Class Classification.
In *International Conference on Machine Learning (ICML)*, pages 815–822, Edinburgh, UK.

Tür, G., Hakkani-Tür, D. Z., and Schapire, R. E. (2005).
Combining active and semi-supervised learning for spoken language understanding.
Speech Communication, 45:171–186.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Schölkopf, B. (2004).
Learning with local and global consistency.
In *Advances in Neural Information Processing Systems 16*, pages 321–328. MIT Press.

Zhu, X. and Ghahramani, Z. (2002).
Learning from labeled and unlabeled data with label propagation.