Transductive Bounds for the Multi-class Majority Vote Classifier

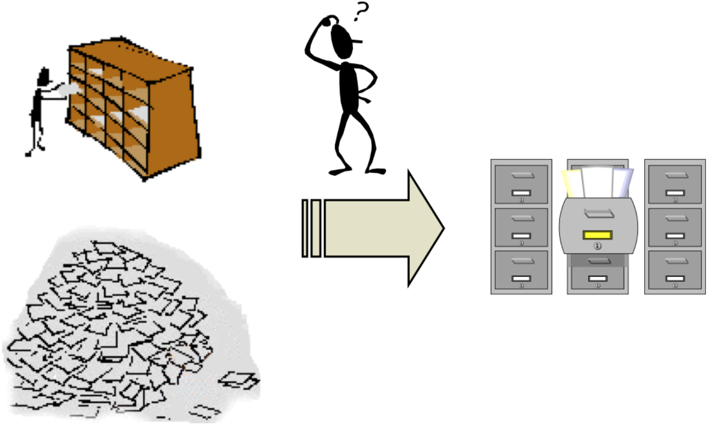
Vasilii Feofanov, Emilie Devijver, Massih-Reza Amini

University Grenoble Alpes, Grenoble INP
LIG, CNRS, Grenoble 38000, France
(firstname.lastname@univ-grenoble-alpes.fr)

AAAI, 2019

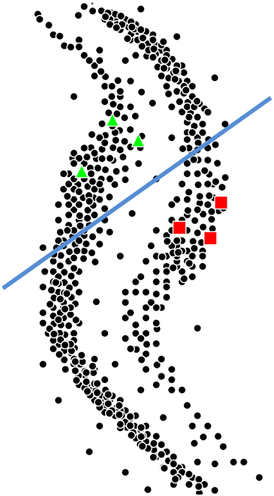
Introduction

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



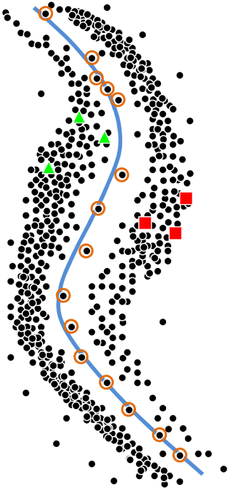
Introduction

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



Introduction

In many applications, labeling examples is prohibitive while huge number of unlabeled data are available. Taking into account the margin to find the low density regions of labeled and unlabeled examples constitutes the basis of many semi-supervised learning algorithms.



Related Work

- To increase performance on the unlabelled set we can infer a model in the transductive way.
- The PAC-Bayesian theory [McAllester, 1999] proposes transductive risk bounds for the Gibbs and the Bayes classifiers.
- Based on this, the self-learning algorithm has been proposed for the binary classification [Amini et al., 2008]. It iteratively pseudo-labels unlabelled examples depending on the prediction scores.

Vasilii Feofanov, Emilie Devijver, Massih-Reza Amini

Motivation

1. Multi-class PAC-Bayesian Theory concerns only the **inductive** case by now.
   - No **transductive** bound of the Bayes classifier has been proposed yet.
2. Few **multi-class** SSL approaches.
3. No self-learning algorithm for the multi-class case yet.

Problem statement

- Input space $\mathcal{X} \subset \mathbb{R}^d$, output $\mathcal{Y} = \{1, \ldots, K\}$ space.
- Labelled examples $\mathrm{Z}_{\mathcal{L}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$.
- Unlabelled observations $\mathrm{X}_{\mathcal{U}} = \{\mathbf{x}_i\}_{i=l+1}^{l+u}$.
- $(\mathbf{x}_i, y_i) \in \mathrm{Z}_{\mathcal{L}} \sim \mathcal{D}$ i.i.d.
- Marginal distribution P_X defined over $\mathcal{X}$.
- Assumption: $\forall\ \mathbf{x}_i \in \mathrm{X}_{\mathcal{U}}$, there is exactly one possible label.
- Hypothesis space $\mathcal{H}$.
- Prior P and posterior Q distributions over $\mathcal{H}$.

Goal: accurate classification of the unlabelled set.
Context: $l << u$.

Definitions

- $B_Q(\mathbf{x}) := \text{argmax}_{c\in\mathcal{Y}} \left[\mathbb{E}_{h\sim Q}\mathbb{1}_{h(\mathbf{x})=c}\right]$, – Bayes majority vote classifier
- $G_Q(\mathbf{x}) := \text{rand}_{h\sim Q} h(\mathbf{x})$, – Gibbs stochastic classifier

Definitions

- $B_Q(\mathbf{x}) := \mathrm{argmax}_{c \in \mathcal{Y}} \left[\mathbb{E}_{h \sim Q} \mathbb{1}_{h(\mathbf{x})=c} \right]$,
- $G_Q(\mathbf{x}) := \mathrm{rand}_{h \sim Q} h(\mathbf{x})$,
- $R_{\mathcal{U}}(B_Q, i, j) := \frac{1}{u_i} \sum_{\mathbf{x}' \in \mathrm{X}_{\mathcal{U}}} \mathbb{1}_{B_Q(\mathbf{x}')=j} \mathbb{1}_{y'=i}$,
- $R_{\mathcal{U}}(G_Q, i, j) := \frac{1}{u_i} \sum_{\mathbf{x}' \in \mathrm{X}_{\mathcal{U}}} \mathbb{E}_{h \sim Q} \mathbb{1}_{h(\mathbf{x}')=j} \mathbb{1}_{y'=i}$,

The error to predict j given class i.

Definitions

- $B_Q(\mathbf{x}) := \operatorname{argmax}_{c\in\mathcal{Y}} \left[\mathbb{E}_{h\sim Q}\mathbb{1}_{h(\mathbf{x})=c}\right]$,
- $G_Q(\mathbf{x}) := \operatorname{rand}_{h\sim Q} h(\mathbf{x})$,
- $R_\mathcal{U}(B_Q, i, j) := \frac{1}{u_i}\sum_{\mathbf{x}'\in\mathrm{X}_\mathcal{U}} \mathbb{1}_{B_Q(\mathbf{x}')=j}\mathbb{1}_{y'=i}$,
- $R_\mathcal{U}(G_Q, i, j) := \frac{1}{u_i}\sum_{\mathbf{x}'\in\mathrm{X}_\mathcal{U}} \mathbb{E}_{h\sim Q}\mathbb{1}_{h(\mathbf{x}')=j}\mathbb{1}_{y'=i}$,
- $\mathrm{E}_\mathcal{U}(h) := \frac{1}{u}\sum_{\mathbf{x}'\in\mathrm{X}_\mathcal{U}} \mathbb{1}_{h(\mathbf{x}')\neq y'}$, – error rate
- $\mathbf{C}_h^\mathcal{U} := (R_\mathcal{U}(h, i, j))_{\substack{i,j=\{1,\dots,K\}^2 \\ i\neq j}}$, – confusion matrix[1]

[1] [Morvant et al., 2012]

Definitions

- $B_Q(\mathbf{x}) := \mathrm{argmax}_{c\in\mathcal{Y}} \left[\mathbb{E}_{h\sim Q}\mathbb{1}_{h(\mathbf{x})=c}\right]$,
- $G_Q(\mathbf{x}) := \mathrm{rand}_{h\sim Q} h(\mathbf{x})$,
- $R_{\mathcal{U}}(B_Q, i, j) := \frac{1}{u_i}\sum_{\mathbf{x}'\in \mathrm{X}_{\mathcal{U}}} \mathbb{1}_{B_Q(\mathbf{x}')=j}\mathbb{1}_{y'=i}$,
- $R_{\mathcal{U}}(G_Q, i, j) := \frac{1}{u_i}\sum_{\mathbf{x}'\in \mathrm{X}_{\mathcal{U}}} \mathbb{E}_{h\sim Q}\mathbb{1}_{h(\mathbf{x}')=j}\mathbb{1}_{y'=i}$,
- $\mathrm{E}_{\mathcal{U}}(h) := \frac{1}{u}\sum_{\mathbf{x}'\in \mathrm{X}_{\mathcal{U}}} \mathbb{1}_{h(\mathbf{x}')\neq y'}$,
- $\mathbf{C}_h^{\mathcal{U}} := \left(R_{\mathcal{U}}(h, i, j)\right)_{\substack{i,j=\{1,\dots,K\}^2 \\ i\neq j}}$,
- $m_Q(\mathbf{x}, c) = \mathbb{E}_{h\sim Q}\mathbb{1}_{h(\mathbf{x})=c}$, – margin: indicator of confidence

Definitions

- $B_Q(\mathbf{x}) := \text{argmax}_{c\in\mathcal{Y}}\left[\mathbb{E}_{h\sim Q}\mathbb{1}_{h(\mathbf{x})=c}\right]$,
- $G_Q(\mathbf{x}) := \text{rand}_{h\sim Q} h(\mathbf{x})$,
- $R_{\mathcal{U}}(B_Q, i, j) := \frac{1}{u_i}\sum_{\mathbf{x}'\in \mathrm{X}_{\mathcal{U}}} \mathbb{1}_{B_Q(\mathbf{x}')=j}\mathbb{1}_{y'=i}$,
- $R_{\mathcal{U}}(G_Q, i, j) := \frac{1}{u_i}\sum_{\mathbf{x}'\in \mathrm{X}_{\mathcal{U}}} \mathbb{E}_{h\sim Q}\mathbb{1}_{h(\mathbf{x}')=j}\mathbb{1}_{y'=i}$,
- $\mathrm{E}_{\mathcal{U}}(h) := \frac{1}{u}\sum_{\mathbf{x}'\in \mathrm{X}_{\mathcal{U}}} \mathbb{1}_{h(\mathbf{x}')\neq y'}$,
- $\mathbf{C}_h^{\mathcal{U}} := \left(R_{\mathcal{U}}(h, i, j)\right)_{\substack{i,j=\{1,\dots,K\}^2 \\ i\neq j}}$,
- $m_Q(\mathbf{x}, c) = \mathbb{E}_{h\sim Q}\mathbb{1}_{h(\mathbf{x})=c}$,
- $R_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q, i, j) := \frac{1}{u_i}\sum_{\mathbf{x}'\in \mathrm{X}_{\mathcal{U}}} \mathbb{1}_{B_Q(\mathbf{x}')=j}\mathbb{1}_{y'=i}\mathbb{1}_{m_Q(\mathbf{x}',j)\geq\theta_j}$,
  – risk to have the conditional error **and** the margin above θ_j

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* $1-\delta$*:*

$$R_{\mathcal{U}}(B_Q, i, j) \le \inf_{\gamma \in [0,1]} \left\{ I_{i,j}^{(\le,<)}(0,\gamma) + \frac{1}{\gamma} \left\lfloor \left(K_{i,j}^{\delta} - M_{i,j}^{<}(\gamma) \right) \right\rfloor_+ \right\},$$

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

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,*
- $I_{i,j}^{(\lhd_1,\lhd_2)}(t,s)$ *is the proportion of class* i *examples with the margin between* t *and* s*,*
- $M_{i,j}^{<}(t)$ *is the average of* j*-margins in class* i *that less than* t*.*

Proof of Theorem

- Lemma 1: connection between the Gibbs conditional risk and the joint Bayes one.
- Bound derived from a solution of a linear program where the error is maximized.
- Lemma 2: the solution of the linear program is the maximal feasible solution in the lexicographical order.

Tightness

Proposition

Ass.: $\forall\ \mathbf{x}' \in \mathrm{X}_{\mathcal{U}}\ \exists\ C \in [0,1]$ *s.t.* $\forall\ (i,j) \in \mathcal{Y}^2,\ \forall\ \gamma > 0$ *if* B_Q *makes cond. mistakes on examples with the margin* $\gamma \Rightarrow$ *the proportion of cond. misclassified examples with margin* $< \gamma$ *is lower bounded by* C*. Then, with prob. at least* $1 - \delta$*:*

$$F_{i,j}^{\delta} - R_{\mathcal{U}}(B_Q, i, j) \leq \frac{1-C}{C} R_{\mathcal{U}}(B_Q, i, j) + \frac{R_u^{\delta}(G_Q, i, j) - R_{\mathcal{U}}(G_Q, i, j)}{\gamma^*},$$

where $F_{i,j}^{\delta}$ *is the proposed bound, and* γ^* *is max margin on which the Bayes classifier makes a cond. mistake.*

Tightness

Proposition

Ass.: $\forall\ \mathbf{x}' \in \mathrm{X}_{\mathcal{U}}\ \exists\ C \in [0,1]$ *s.t.* $\forall\ (i,j) \in \mathcal{Y}^2,\ \forall\ \gamma > 0$ *if* B_Q *makes cond. mistakes on examples with the margin* $\gamma \Rightarrow$ *the proportion of cond. misclassified examples with margin* $< \gamma$ *is lower bounded by* C. *Then, with prob. at least* $1-\delta$:

$$F_{i,j}^{\delta} - R_{\mathcal{U}}(B_Q,i,j) \leq \frac{1-C}{C} R_{\mathcal{U}}(B_Q,i,j) + \frac{R_u^{\delta}(G_Q,i,j) - R_{\mathcal{U}}(G_Q,i,j)}{\gamma^*},$$

where $F_{i,j}^{\delta}$ *is the proposed bound, and* γ^* *is max margin on which the Bayes classifier makes a cond. mistake.*

Additional assumptions:

- The Gibbs conditional risk bound is tight,

Tightness

Proposition

Ass.: $\forall\ \mathbf{x}' \in \mathrm{X}_{\mathcal{U}}\ \exists\ C \in [0,1]$ *s.t.* $\forall\ (i,j) \in \mathcal{Y}^2,\ \forall\ \gamma > 0$ *if* B_Q *makes cond. mistakes on examples with the margin* $\gamma \Rightarrow$ *the proportion of cond. misclassified examples with margin* $< \gamma$ *is lower bounded by* C*.*
Then, with prob. at least $1-\delta$*:*

$$F^{\delta}_{i,j} - R_{\mathcal{U}}(B_Q,i,j) \leq \frac{1-C}{C} R_{\mathcal{U}}(B_Q,i,j) + \frac{R^{\delta}_{u}(G_Q,i,j) - R_{\mathcal{U}}(G_Q,i,j)}{\gamma^*},$$

where $F^{\delta}_{i,j}$ *is the proposed bound, and* γ^* *is max margin on which the Bayes classifier makes a cond. mistake.*

Additional assumptions:

- The Gibbs conditional risk bound is tight,
- The Bayes classifier makes its mistakes mostly on examples with low margins $\Rightarrow C$ is close to 1.

Tightness

Proposition

Ass.: $\forall\ \mathbf{x}' \in X_{\mathcal{U}}\ \exists\ C \in [0,1]$ *s.t.* $\forall\ (i,j) \in \mathcal{Y}^2$, $\forall\ \gamma > 0$ *if* B_Q *makes cond. mistakes on examples with the margin* $\gamma \Rightarrow$ *the proportion of cond. misclassified examples with margin* $< \gamma$ *is lower bounded by* C*. Then, with prob. at least* $1-\delta$*:*

$$F^{\delta}_{i,j} - R_{\mathcal{U}}(B_Q,i,j) \leq \frac{1-C}{C} R_{\mathcal{U}}(B_Q,i,j) + \frac{R^{\delta}_{u}(G_Q,i,j) - R_{\mathcal{U}}(G_Q,i,j)}{\gamma^*},$$

where $F^{\delta}_{i,j}$ *is the proposed bound, and* γ^* *is max margin on which the Bayes classifier makes a cond. mistake.*

Additional assumptions:

- The Gibbs conditional risk bound is tight,
- The Bayes classifier makes its mistakes mostly on examples with low margins $\Rightarrow C$ is close to 1.

Hence, the bound is tight!

A Transductive Bound of the Error Rate

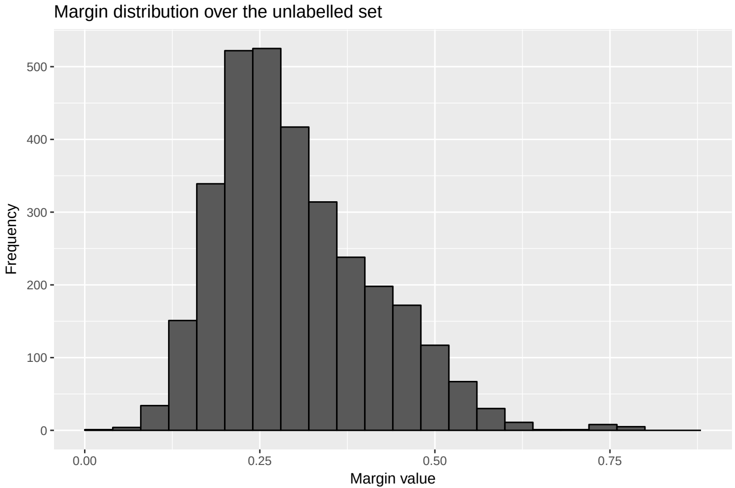
Corollary

$U^{\delta}_{i,j}(\boldsymbol{\theta})$ is a bound of $R_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q, i, j)$ holding with prob. at least $1-\delta$,
$\mathbf{U}^{\delta}_{\boldsymbol{\theta}}$ the corresponding confusion matrix.
Then, we have:

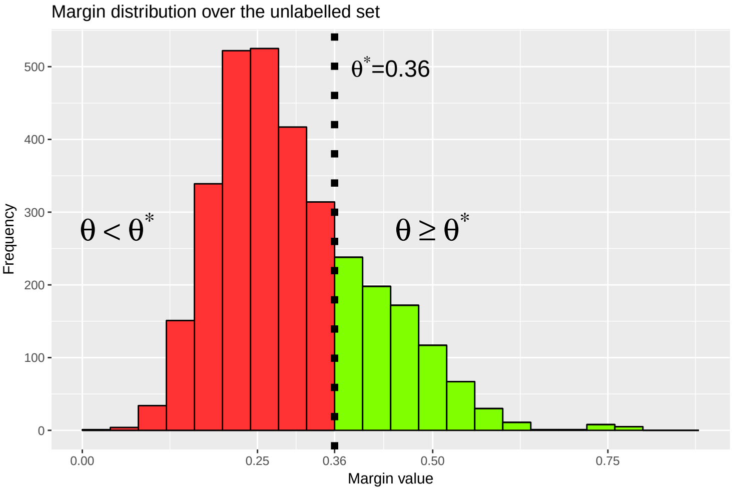
$$\mathrm{E}_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q) \leq \left\| \left(\mathbf{U}^{\delta}_{\boldsymbol{\theta}}\right)^{\mathsf{T}} \mathbf{p} \right\|_1,$$
$$\mathrm{E}_{\mathcal{U}}(B_Q) \leq \left\| \left(\mathbf{U}^{\delta}_{\mathbf{0}_K}\right)^{\mathsf{T}} \mathbf{p} \right\|_1,$$

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

Automatic Threshold Finding



Automatic Threshold Finding



Conditional Bayes Error

Definition

Conditional Bayes error rate $\mathrm{E}_{\mathcal{U}|\boldsymbol{\theta}}(B_Q)$*:*

$$\mathrm{E}_{\mathcal{U}|\boldsymbol{\theta}}(B_Q) := \frac{\mathrm{E}_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q)}{\frac{1}{u}\sum_{\mathbf{x}'\in \mathrm{X}_{\mathcal{U}}} \mathbb{1}_{m_Q(\mathbf{x}',B_Q(\mathbf{x}'))\geq \theta_{B_Q(\mathbf{x}')}}},$$

Trade-off between:

Conditional Bayes Error

Definition

Conditional Bayes error rate $\mathrm{E}_{\mathcal{U}|\boldsymbol{\theta}}(B_Q)$*:*

$$\mathrm{E}_{\mathcal{U}|\boldsymbol{\theta}}(B_Q) := \frac{\mathrm{E}_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q)}{\frac{1}{u}\sum_{\mathbf{x}'\in \mathrm{X}_{\mathcal{U}}} \mathbb{1}_{m_Q(\mathbf{x}',B_Q(\mathbf{x}'))\geq\theta_{B_Q(\mathbf{x}')}}},$$

Trade-off between:

- Error on unlabeled examples with margin above $\theta_{B_Q(\mathbf{x}')}$,

Conditional Bayes Error

Vasilii Feofanov, Emilie Devijver, Massih-Reza Amini

Definition

Conditional Bayes error rate $\mathrm{E}_{\mathcal{U}|\boldsymbol{\theta}}(B_Q)$*:*

$$\mathrm{E}_{\mathcal{U}|\boldsymbol{\theta}}(B_Q) := \frac{\mathrm{E}_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q)}{\frac{1}{u}\sum_{\mathbf{x}'\in \mathrm{X}_{\mathcal{U}}} \mathbb{1}_{m_Q(\mathbf{x}',B_Q(\mathbf{x}'))\geq\theta_{B_Q(\mathbf{x}')}}},$$

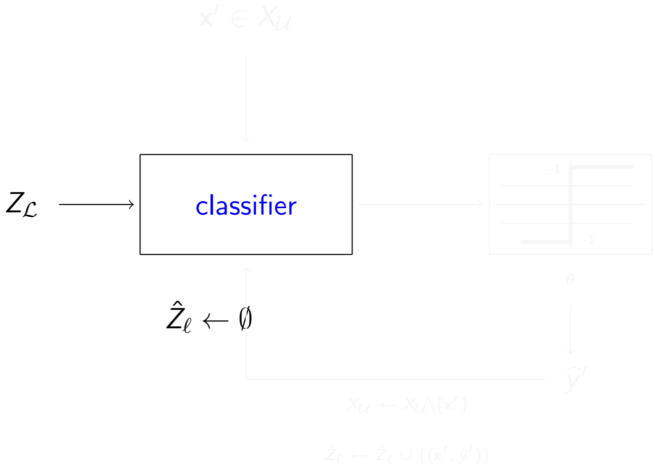
Trade-off between:

- Error on unlabeled examples with margin above $\theta_{B_Q(\mathbf{x}')}$,
- Fraction of pseudo-labeled examples in $\mathrm{X}_{\mathcal{U}}$.

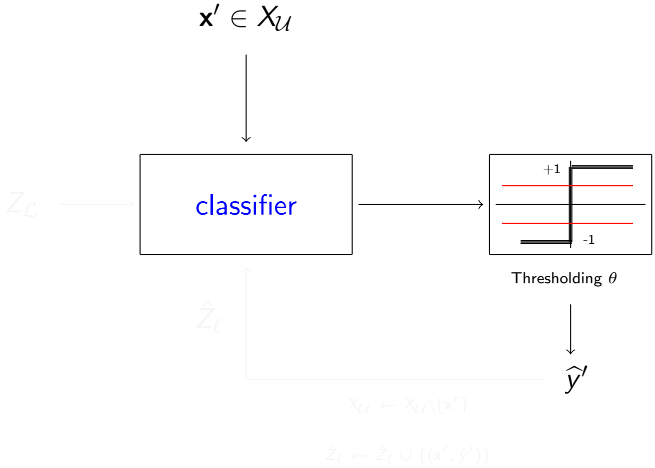
Multi-class Self-learning Algorithm

Vasilii Feofanov, Emilie Devijver, Massih-Reza Amini

Multi-class Self-learning Algorithm



Multi-class Self-learning Algorithm

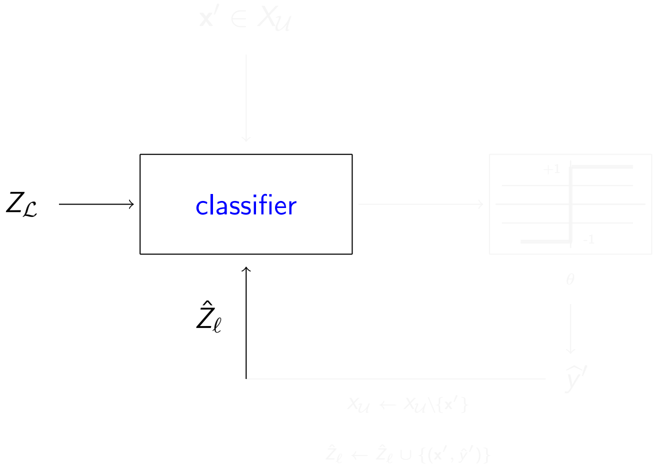


Multi-class Self-learning Algorithm

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

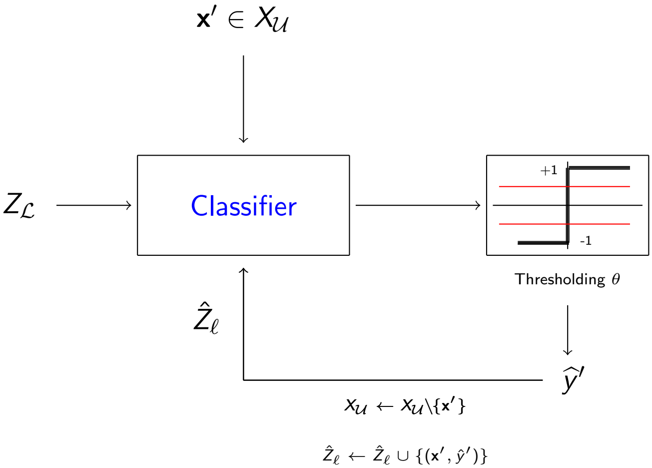
$$\hat{Z}_{\ell} \leftarrow \hat{Z}_{\ell} \cup \{(\mathbf{x}', \hat{y}')\}$$

Multi-class Self-learning Algorithm



Vasilii Feofanov, Emilie Devijver, Massih-Reza Amini

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.

Classification Performance with respect to l.

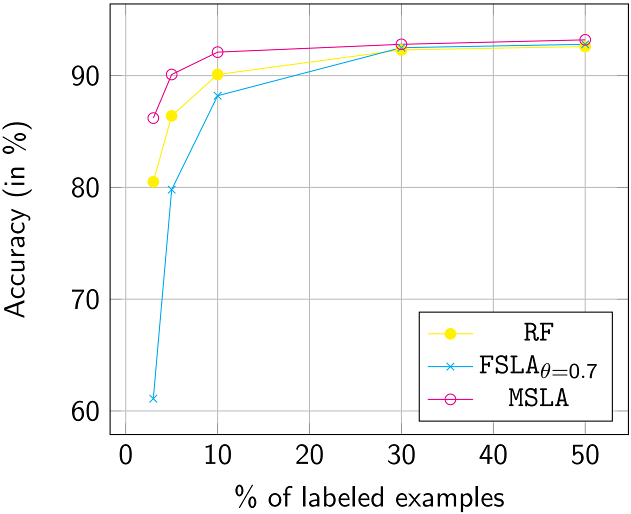


Figure: Accuracy on 3500 examples randomly chosen from the MNIST dataset.

Conclusion and Perspectives

- Proposed transductive bounds for the Bayes classifier, which are tight under certain conditions.
- Self-learning with automatic threshold finding shows promising results for semi-supervised tasks.

Conclusion and Perspectives

- Proposed transductive bounds for the Bayes classifier, which are tight under certain conditions.
- Self-learning with automatic threshold finding shows promising results for semi-supervised tasks.
- Future perspective: self-learning with semi-supervised feature selection.

Conclusion and Perspectives

- Proposed transductive bounds for the Bayes classifier, which are tight under certain conditions.
- Self-learning with automatic threshold finding shows promising results for semi-supervised tasks.
- Future perspective: self-learning with semi-supervised feature selection.

References

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.

McAllester, D. A. (1999).
PAC-bayesian model averaging.
In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, COLT '99, pages 164–170, New York, NY, USA. ACM.

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, United Kingdom.
Arxiv: http://arxiv.org/abs/1202.6228, Accepted at ICML 2012.