Transductive Bounds for the Multi-class Majority Vote Classifier

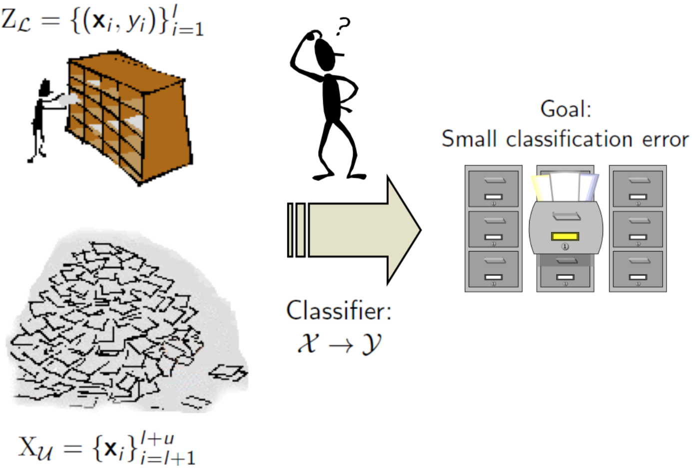
Vasilii Feofanov, Emilie Devijver, Massih-Reza Amini

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

PFIA 2019: France@International
July 5, 2019

Introduction

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



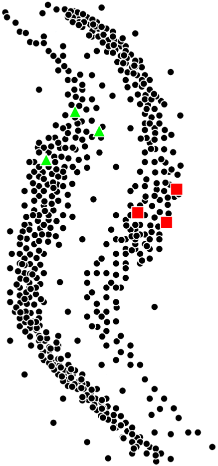
- **Supervised Learning:**
  Labeled data $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$.

$$\Downarrow$$

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

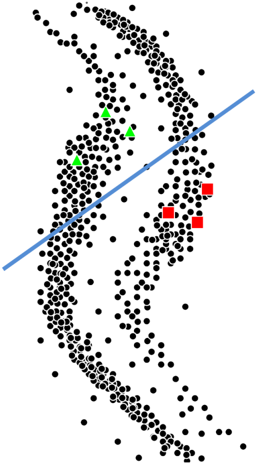
$$\Uparrow$$

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



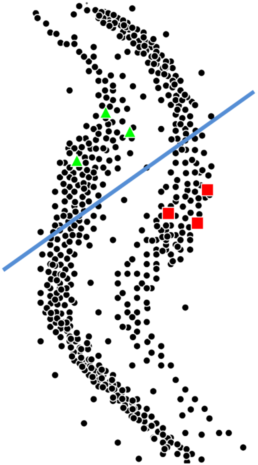
Example of partially labeled data

Problem: Supervised learning is not efficient to use.

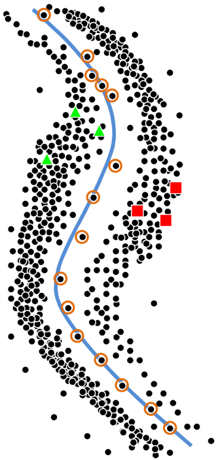


(a) Supervised classifier

Solution: Classifier that pass through the low density regions of both labeled and unlabeled examples.



(a) Supervised classifier



(b) Semi-supervised classifier

- We consider the transductive inference. The self-learning algorithm (SLA) is based on this paradigm. In [Amini et al., 2008] it was proposed to find a threshold for the **binary** SLA dynamically using a risk bound.

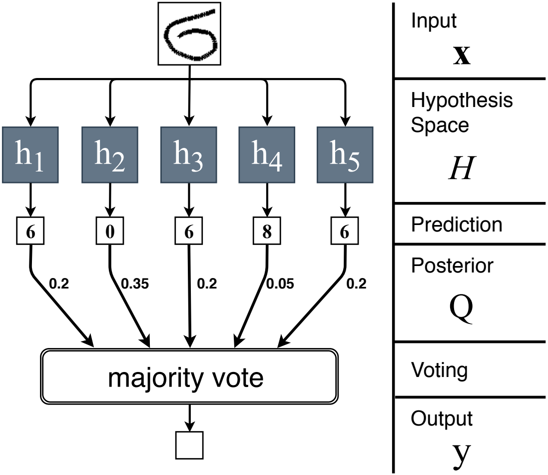
- PAC-Bayesian theorems [McAllester, 1999] bound risk of Gibbs and Bayes classifiers. Most of study is devoted to the binary framework. [Morvant et al., 2012] considers the multi-class case in the **supervised** setting.

Contribution

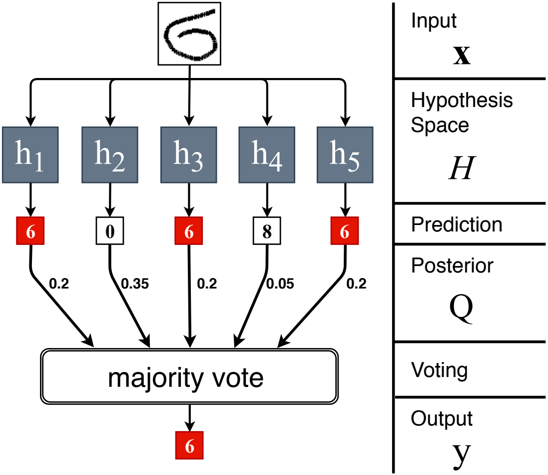
In this work, we propose:

1. **Transductive** bounds of the Bayes classifier,
2. A **multi-class** extension of the self-learning algorithm.

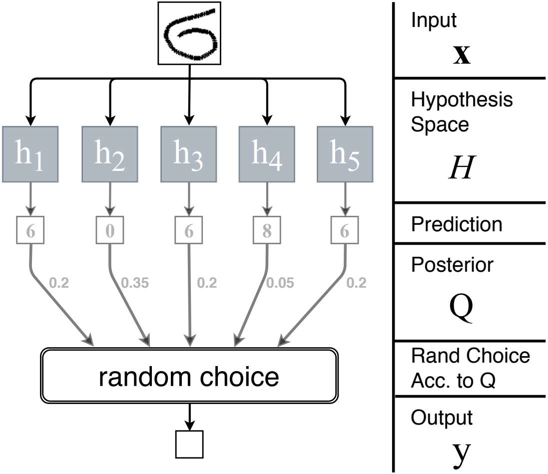
$$B_Q(\mathbf{x}) := \operatorname{argmax}_{c \in \mathcal{Y}} \left[\mathbb{E}_{h \sim Q} \mathbb{I}(h(\mathbf{x}) = c) \right]$$



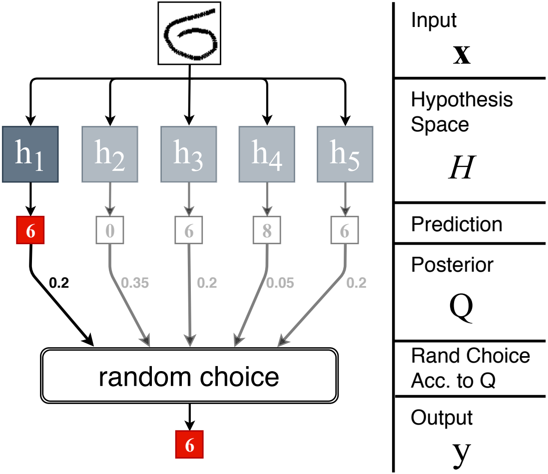
$$B_Q(\mathbf{x}) := \text{argmax}_{c \in \mathcal{Y}} \left[\mathbb{E}_{h \sim Q} \mathbb{I}(h(\mathbf{x}) = c) \right]$$



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

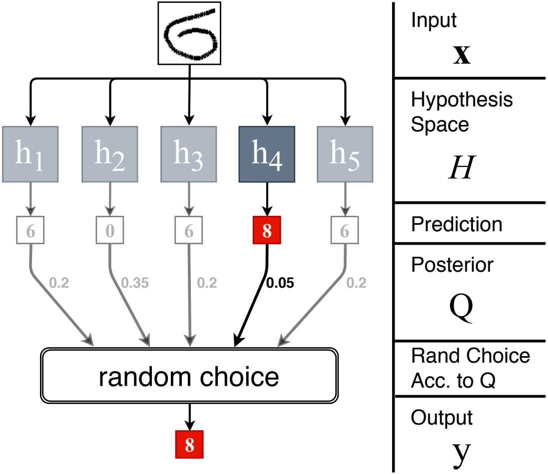


$G_Q(\mathbf{x}) := \mathrm{rand}_{h\sim Q} h(\mathbf{x})$

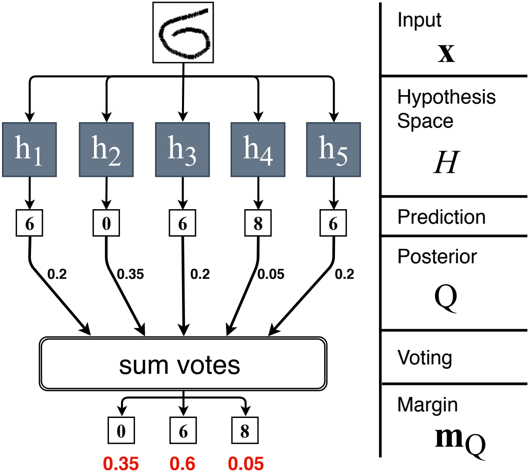


Gibbs Classifier

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



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



Error Measures

Conditional risk:

- $R_{\mathcal{U}}(B_Q, i, j) := \frac{1}{u_i} \sum_{\mathbf{x}' \in X_{\mathcal{U}}} \mathbb{I}(B_Q(\mathbf{x}') = j)\mathbb{I}(y' = i)$,
- $R_{\mathcal{U}}(G_Q, i, j) := \frac{1}{u_i} \sum_{\mathbf{x}' \in X_{\mathcal{U}}} \mathbb{E}_{h \sim Q}\mathbb{I}(h(\mathbf{x}') = j)\mathbb{I}(y' = i)$,
  The error to predict j given class i.

Error Measures

Conditional risk:

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

Error rate:

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

Confusion matrix:

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

Conditional risk:

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

Error rate:

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

Confusion matrix:

- $\mathbf{C}_h^{\mathcal{U}} := (R_{\mathcal{U}}(h, i, j))_{\substack{i,j=\{1,\dots,K\}^2 \\ i \neq j}}$,

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

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 $1-\delta$:

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

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.*

Theorem

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

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

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,*
- $I_{i,j}^{(\leq,<)}(\theta_j, \gamma)$ *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***.*

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

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:

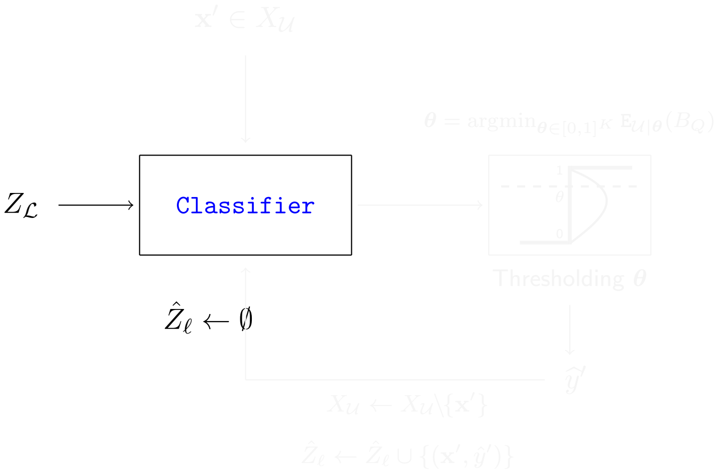
$$\mathrm{E}_{\mathcal{U}|\boldsymbol{\theta}}(B_Q) := \frac{\mathrm{E}_{\mathcal{U}\wedge\boldsymbol{\theta}}(B_Q)}{\pi(m_Q(\mathbf{x}', B_Q(\mathbf{x}')) \geq \theta_{B_Q(\mathbf{x}')})}.$$

A **trade-off** between:

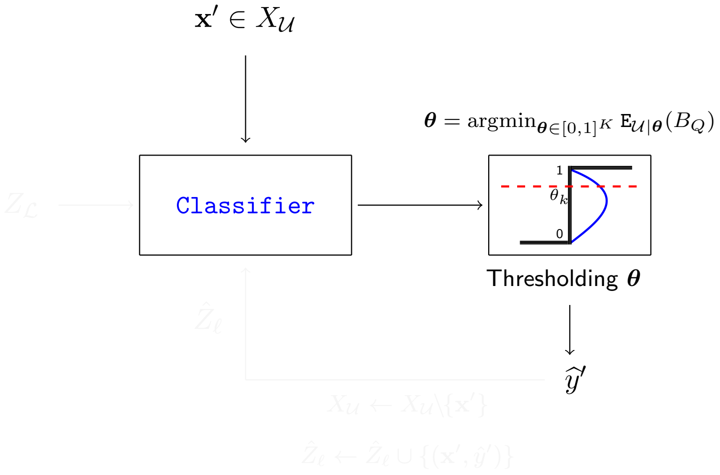
- Transductive error on pseudo-labeled examples (estimated using **Theorem**),
- Proportion of pseudo-labeled examples in $\mathrm{X}_{\mathcal{U}}$.

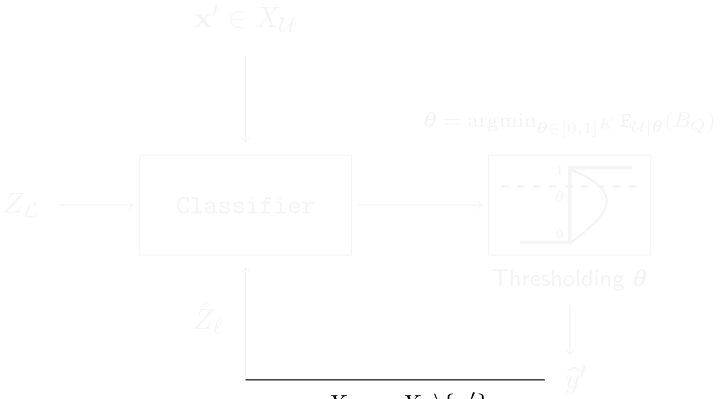
Multi-class Self-learning Algorithm

Multi-class Self-learning Algorithm



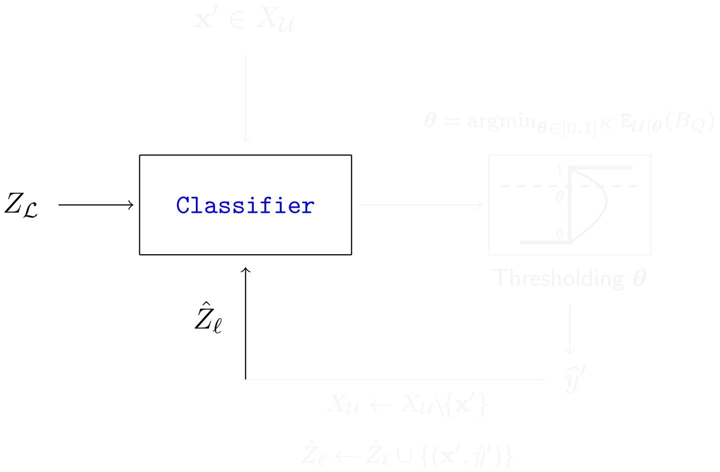
Multi-class Self-learning Algorithm





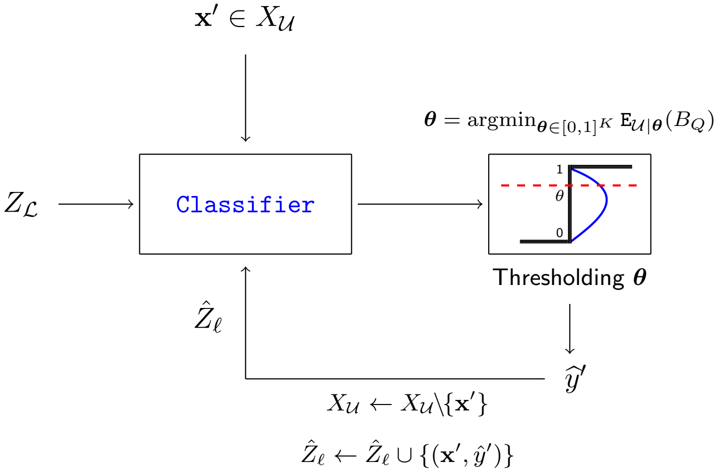
$$X_{\mathcal{U}} \leftarrow X_{\mathcal{U}} \backslash \{\mathbf{x}'\}$$

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



$\mathbf{x}' \in X_{\mathcal{U}}$
$\boldsymbol{\theta} = \mathrm{argmin}_{\boldsymbol{\theta} \in [0,1]^K} \mathrm{E}_{\mathcal{U}|\boldsymbol{\theta}}(B_Q)$
$Z_{\mathcal{L}}$
Classifier
1
θ
0
Thresholding $\boldsymbol{\theta}$
$\hat{Z}_{\ell}$
$\hat{y}'$
$X_{\mathcal{U}} \leftarrow X_{\mathcal{U}} \backslash \{\mathbf{x}'\}$
$\hat{Z}_{\ell} \leftarrow \hat{Z}_{\ell} \cup \{(\mathbf{x}', \hat{y}')\}$

Multi-class Self-learning Algorithm



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.

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.

The source code:

`github.com/vfeofanov/trans-bounds-maj-vote`

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.

The source code:

`github.com/vfeofanov/trans-bounds-maj-vote`

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, UK.