Learning a Multiview Weighted Majority Vote Classifier:
Using PAC-Bayesian Theory and Boosting

Anil Goyal

Laboratoire Hubert Curien UMR CNRS, Université Jean Monnet, Saint-Etienne, France

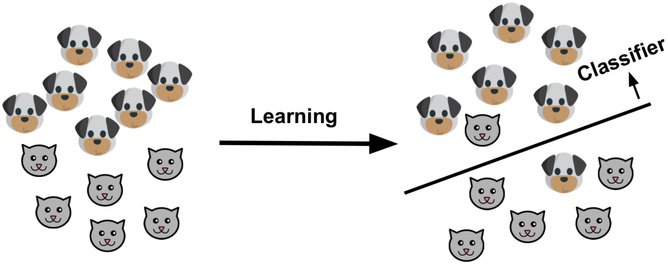
Univ. Grenoble Alps, Laboratoire d'Informatique de Grenoble, AMA, Grenoble, France

Soutenance de thèse : October 23, 2018

Devant le jury composé de:

Rapporteurs : Jean-Christophe Janodet, Cécile Capponi
Examinateur : Amaury Habrard
Directeur : Massih-Reza Amini
Co-directrice : Emilie Morvant

Supervised Learning



Find a classifier which performs well on new unseen data

$\Rightarrow$ Minimize the empirical error on training data

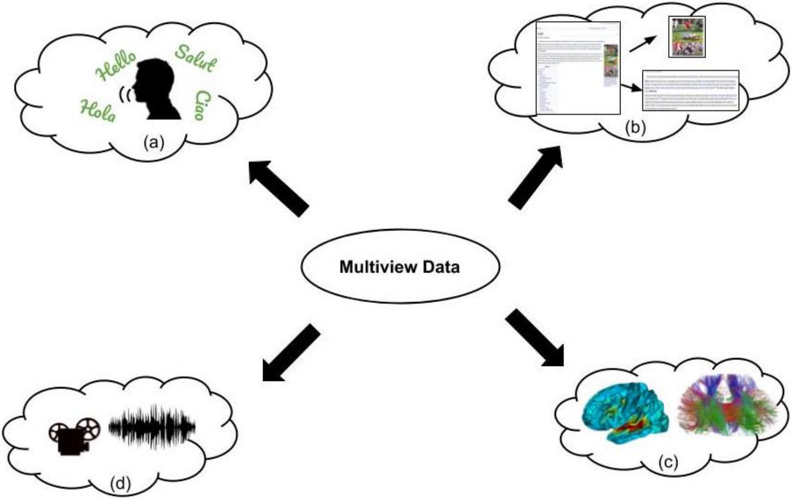
$\Rightarrow$ Require generalization guarantees

Generalization bound

$$\text{True Error} \leq \text{Empirical Error} + f\left(\textit{complexity}, \frac{1}{\text{number of examples}}\right)$$

Observations with Multiple Views

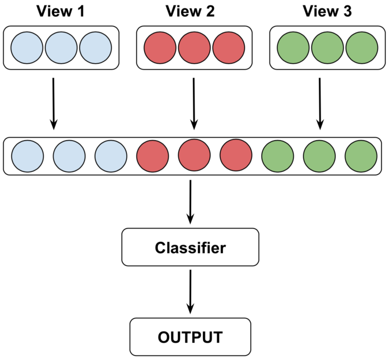
The corpus is described by different features called views



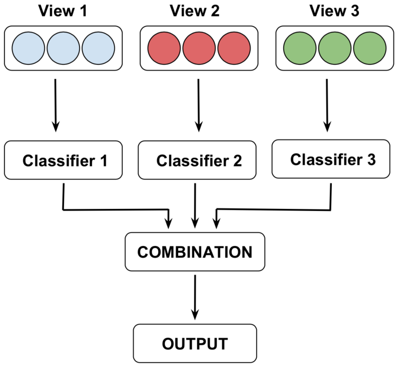
The corpus is described by different features called views

Objective

Take advantage of multiple views of data to make better prediction

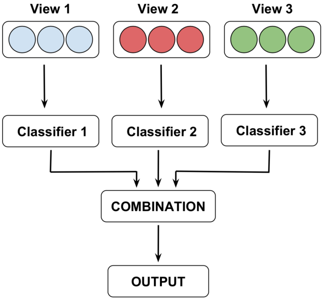


Early Fusion



Late Fusion

Multiview Learning



Objective

- Consider more than two views
- Derive generalization guarantees for multiview learning

Our Solution

Combination $=$ Weighted majority vote over the classifiers

$\Rightarrow$ Exploiting PAC-Bayesian Theory and Boosting

Thesis Contributions

Theoretical Contributions

- A new PAC-Bayesian Theorem as an Expected Risk Bound (CAp'17, ECML-PKDD'17)
- PAC-Bayesian Analysis of Multiview Learning (CAp'16, CAp'17, ECML-PKDD'17)

Algorithmic Contributions

- Two-step multiview learning algorithm based on late fusion approach (CAp'17, ECML-PKDD'17)
- One-step boosting based multiview learning algorithm (Submitted to Neurocomputing)
- Multiview Learning as Bregman Divergence Minimization (CAp'18, IDA'18)

Outline

PAC-Bayesian Setting

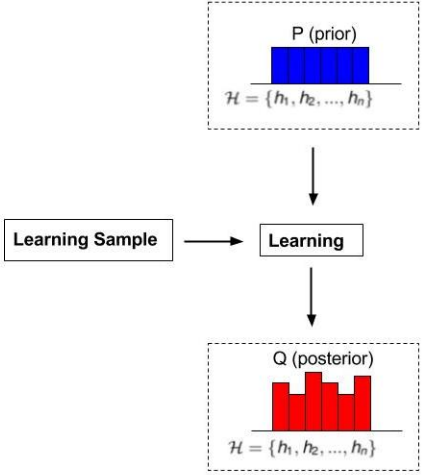
- $\mathcal{X} \subseteq \mathbb{R}^d$ input space, $\mathcal{Y} = \{-1, +1\}$
- $\mathcal{D}$ is unknown distribution on $\mathcal{X} \times \mathcal{Y}$
- Learning Sample:
  $S = \{(x_i, y_i)\}_{i=1}^m \overset{iid}{\sim} (\mathcal{D})^m$
- A set of classifiers $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}$

PAC-Bayesian classification (1-view)

Goal:
Finding the ***Q*-weighted majority vote** B_Q over $\mathcal{H}$ which minimizes $R_\mathcal{D}(B_Q)$

$$\underbrace{R_\mathcal{D}(B_Q) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \mathbb{1}_{[B_Q(x) \neq y]}}_{\text{True Risk}}$$

where $B_Q(\mathbf{x}) = \text{sign}\left[\mathop{\mathbb{E}}_{h\sim Q} h(\mathbf{x})\right]$



The stochastic Gibbs classifier

- The PAC-Bayesian approach does **not** directly focus on $R_{\mathcal{D}}(B_Q)$
- but on the error of the stochastic **Gibbs classifier** G_Q

  which labels a new example $\mathbf{x} \in X$ by

  - picking one h according to Q
  - returning $h(\mathbf{x})$

The stochastic Gibbs classifier

- The PAC-Bayesian approach does **not** directly focus on $R_{\mathcal{D}}(B_Q)$
- but on the error of the stochastic **Gibbs classifier** G_Q
  which labels a new example $\mathbf{x} \in X$ by
  - picking one h according to Q
  - returning $h(\mathbf{x})$

IMPORTANT — the risk of G_Q is the expectation of the risks on $\mathcal{H}$ according to Q

$$R_{\mathcal{D}}(G_Q) = \mathop{\mathbb{E}}_{h \sim Q} R_{\mathcal{D}}(h)$$

We can prove

i) $R_{\mathcal{D}}(B_Q) \leq 2 R_{\mathcal{D}}(G_Q)$

ii) $\mathcal{C}$-Bound:

$$R_{\mathcal{D}}(B_Q) \leq 1 - \frac{\left(1 - 2R_{\mathcal{D}}(G_Q)\right)^2}{1 - 2d_{\mathcal{D}}(Q)}$$

where $d_{\mathcal{D}}(Q) = \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} \mathop{\mathbb{E}}_{h,h' \sim Q^2} \mathbb{1}_{[h(x) \neq h'(x)]}$ is the **expected disagreement**

Outline

Monoview PAC-Bayesian Bound

General Form of Probabilistic Generalization Bound:

$$\underset{S\sim(\mathcal{D})^m}{\textbf{Prob}}\left(\forall h \in \mathcal{H}, \underbrace{R_{\mathcal{D}}(h)}_{\text{True Error}} \leq \underbrace{R_S(h)}_{\text{Empirical Error}} + f\left(\mathit{complexity}(h), \frac{1}{m}, \delta\right)\right) \geq 1-\delta$$

General Form of Probabilistic Generalization Bound:

$$\underset{S\sim(\mathcal{D})^m}{\textbf{Prob}}\left(\forall h \in \mathcal{H},\ \underbrace{R_{\mathcal{D}}(h)}_{\text{True Error}} \leq \underbrace{R_S(h)}_{\text{Empirical Error}} + f\left(complexity(h), \frac{1}{m}, \delta\right)\right) \geq 1-\delta$$

Theorem (McAllester 2003, Germain et al. 2015)

For any $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, for any $\mathcal{H}$, for any prior P over $\mathcal{H}$, for any $\delta \in (0,1]$, we have

$$\underset{S\sim(\mathcal{D})^m}{\textbf{Prob}}\left(\forall Q \text{ on } \mathcal{H},\ \underbrace{\mathop{\mathbb{E}}_{h\sim Q} R_{\mathcal{D}}(h)}_{R_{\mathcal{D}}(G_Q)} \leq \underbrace{\mathop{\mathbb{E}}_{h\sim Q} R_S(h)}_{R_S(G_Q)} + \sqrt{\frac{\mathrm{KL}(Q\|P)+\ln\frac{2\sqrt{m}}{\delta}}{2m}}\right) \geq 1-\delta$$

where, $R_{\mathcal{D}}(h)$ and $R_S(h)$ are respectively the true and the empirical risks of individual voters and $\mathrm{KL}(Q\|P) = \mathop{\mathbb{E}}_{h\sim Q} \ln\frac{Q(h)}{P(h)}$

For any $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, for any $\mathcal{H}$, for any prior P over $\mathcal{H}$, for any $\delta \in (0,1]$, we have

$$\operatorname*{\mathbf{Prob}}_{S\sim(\mathcal{D})^m}\left(\forall Q \text{ on } \mathcal{H},\ \underbrace{\mathop{\mathbb{E}}_{h\sim Q} R_{\mathcal{D}}(h)}_{R_{\mathcal{D}}(G_Q)} \le \underbrace{\mathop{\mathbb{E}}_{h\sim Q} R_S(h)}_{R_S(G_Q)} + \sqrt{\frac{\mathrm{KL}(Q\|P) + \ln\frac{2\sqrt{m}}{\delta}}{2m}}\right) \ge 1-\delta$$

First contribution (CAp'17, ECML-PKDD'17)

$\Rightarrow$ Risk Bound **in expectation** over all learning samples $S \overset{iid}{\sim} (\mathcal{D})^m$

For any $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, for any $\mathcal{H}$, for any prior P on $\mathcal{H}$, for any convex function $D : [0,1] \times [0,1] \to \mathbb{R}$

$$\mathop{\mathbb{E}}_{S\sim(\mathcal{D})^m} R_{\mathcal{D}}(G_{Q_S}) \le \mathop{\mathbb{E}}_{S\sim(\mathcal{D})^m} R_S(G_{Q_S}) + \sqrt{\frac{\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \mathrm{KL}(Q_S\|P) + \ln 2\sqrt{m}}{2m}}$$

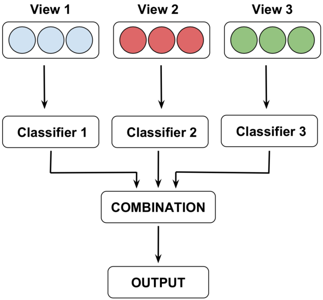
Take Home Message

Expected Risk bound for PAC-Bayesian theory

- General bound for single view learning
- Expressed as expectation over all the possible learning samples
- Extension of this bound to multiview learning

Outline

Multiview Learning Setting



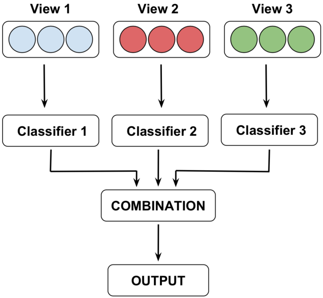
Objective

- Take advantage of $\mathbf{V} \geq \mathbf{2}$ views of data to make better prediction
- Control the trade-off **accuracy and diversity** between the views

From Multiview to PAC-Bayes. . .

Hierarchy of distribution over all the view-specific classifiers

Multiview Learning Setting

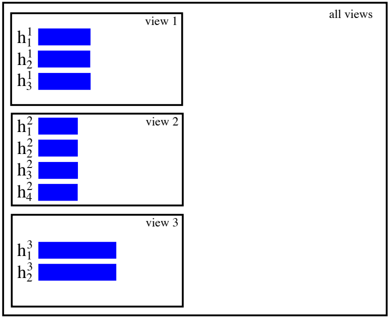


Formally,

- $V \geq 2$ different input spaces $\mathcal{X}_v \subseteq \mathbb{R}^{d_v}$
- **Joint input space**: $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_V$, **output space**: $\mathcal{Y} = \{-1, +1\}$
- **An example**: $(\mathbf{x}, y) = ((x^1, \ldots, x^V), y) \in \mathcal{X} \times \mathcal{Y}$
- $\mathcal{D}$ unknown **distribution** on $\mathcal{X} \times \mathcal{Y}$
- Given multiview **learning sample** $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \sim (\mathcal{D})^m$
- $\forall v \in \mathcal{V}$, we have $\mathcal{H}_v$ a set of view-specific classifiers s.t. $\forall h_v \in \mathcal{H}_v, h_v : \mathcal{X}_v \to \mathcal{Y}$

Hierarchy of distributions for PAC-Bayes

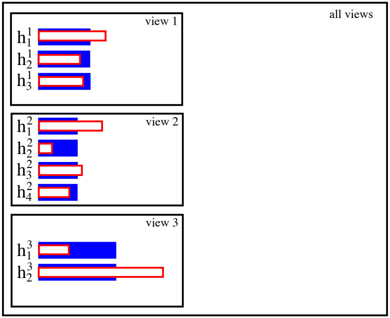
For **each** view $v \in \mathcal{V}$, P_v prior distribution on $\mathcal{H}_v$



Hierarchy of distributions for PAC-Bayes

For **each** view $v \in \mathcal{V}$, P_v prior distribution on $\mathcal{H}_v$

$\Longrightarrow$ finding a posterior distribution Q_v over $\mathcal{H}_v$

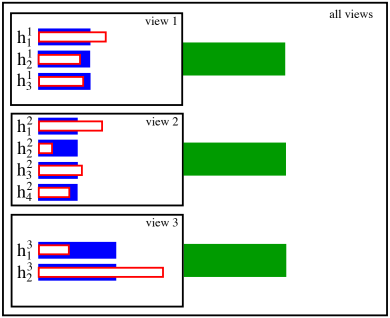


Hierarchy of distributions for PAC-Bayes

For **each** view $v \in \mathcal{V}$, P_v prior distribution on $\mathcal{H}_v$

$\Longrightarrow$ finding a posterior distribution Q_v over $\mathcal{H}_v$

π hyper-prior distribution over **all** the views $\mathcal{V}$



Hierarchy of distributions for PAC-Bayes

For **each** view $v \in \mathcal{V}$, P_v prior distribution on $\mathcal{H}_v$

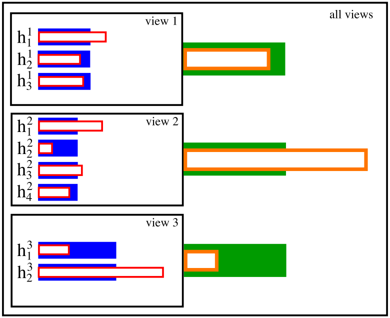
$\Longrightarrow$ finding a posterior distribution Q_v over $\mathcal{H}_v$

π hyper-prior distribution over **all** the views $\mathcal{V}$

$\Longrightarrow$ finding a hyper-posterior distribution ρ on $\mathcal{V}$

such that they minimize the true risk $R_{\mathcal{D}}(B_\rho^{MV})$ of the majority vote B_ρ^{MV}

$$B_\rho^{MV}(\mathbf{x}) = \operatorname{sign}\left[\mathop{\mathbb{E}}_{v\sim\rho}\ \mathop{\mathbb{E}}_{h\sim Q_v} h(x^v)\right]$$



The Multiview Gibbs classifier

True risk of the Multiview Gibbs classifier

$$R_{\mathcal{D}}(G_{\rho}^{MV}) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathop{\mathbb{E}}_{v\sim\rho} \mathop{\mathbb{E}}_{h\sim Q_v} R_{\mathcal{D}}(h(x^v)) = \frac{1}{2} \underbrace{d_{\mathcal{D}}^{\text{MV}}(\rho)}_{\text{disagreement}} + \underbrace{e_{\mathcal{D}}^{\text{MV}}(\rho)}_{\text{joint error}}$$

We can prove

(i) $R_{\mathcal{D}}(B_{\rho}^{MV}) \leq 2R_{\mathcal{D}}(G_{\rho}^{MV})$

(ii) The multiview $\mathcal{C}$-Bound
$\hookrightarrow$ Controls the trade-off between **accuracy** and **diversity**

$$R_{\mathcal{D}}(B_{\rho}^{MV}) \leq 1 - \frac{\left(1 - 2R_{\mathcal{D}}(G_{\rho}^{MV})\right)^2}{1 - 2d_{\mathcal{D}}^{\text{MV}}(\rho)} \leq 1 - \frac{\left(1 - 2\,\mathbb{E}_{v\sim\rho} R_{\mathcal{D}}(G_{Q_v})\right)^2}{1 - 2\,\mathbb{E}_{v\sim\rho} d_{\mathcal{D}}(Q_v)}$$

where
$$d_{\mathcal{D}}^{\text{MV}}(\rho) = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}} \mathop{\mathbb{E}}_{v\sim\rho} \mathop{\mathbb{E}}_{v'\sim\rho} \mathop{\mathbb{E}}_{h\sim Q_v} \mathop{\mathbb{E}}_{h'\sim Q_{v'}} \mathbb{1}_{[h(x^v)\neq h'(x^{v'})]}$$
$$e_{\mathcal{D}}^{\text{MV}}(\rho) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathop{\mathbb{E}}_{v\sim\rho} \mathop{\mathbb{E}}_{v'\sim\rho} \mathop{\mathbb{E}}_{h\sim Q_v} \mathop{\mathbb{E}}_{h'\sim Q_{v'}} \mathbb{1}_{[h(x^v)\neq y]}\mathbb{1}_{[h'(x^{v'})\neq y]}$$

Outline

Non-probabilistic Multiview PAC-Bayes Bound

(CAp'16, CAp'17, ECML-PKDD'17)

For any $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, for any set of priors $\{P_v\}_{v=1}^{V}$, for any hyper-priors π over $\mathcal{V}$, we have

$$\mathbb{E}_{S\sim\mathcal{D}^m} R_{\mathcal{D}}(G^{MV}_{\rho_S}) \leq \underbrace{\frac{1}{2}\mathbb{E}_{S\sim\mathcal{D}^m} d^{MV}_S(\rho_S) + \mathbb{E}_{S\sim\mathcal{D}^m} e^{MV}_S(\rho_S)\}}_{\mathbb{E}_{S\sim\mathcal{D}^m} R_S(G^{MV}_{\rho_S})}$$

$$+ \sqrt{\frac{\mathbb{E}_{S\sim\mathcal{D}^m} \mathrm{KL}(\rho_S \| \pi) + \ln 2\sqrt{m}}{2m}}$$

Non-probabilistic Multiview PAC-Bayes Bound
(CAp'16, CAp'17, ECML-PKDD'17)

For any $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, for any set of priors $\{P_v\}_{v=1}^{V}$, for any hyper-priors π over $\mathcal{V}$, we have

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} R_{\mathcal{D}}(G_{\rho_S}^{MV}) \leq \underbrace{\frac{1}{2} \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} d_S^{\text{MV}}(\rho_S) + \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} e_S^{\text{MV}}(\rho_S)\}}_{\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} R_S(G_{\rho_S}^{MV})}$$

$$+ \sqrt{\frac{\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{v\sim\rho_S} \text{KL}(Q_{v,S}\|P_v) + \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \text{KL}(\rho_S\|\pi) + \ln 2\sqrt{m}}{2m}}$$

Trade-off:

- Empirical disagreement and joint error
- Expectation of view-specific KL divergences over all the views
- KL divergence between hyper-posterior and hyper-prior

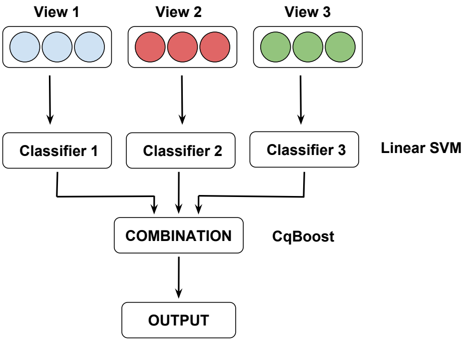
Take Home Message

Instantiation of the **PAC-Bayesian theory** to multiview learning

- with **more than 2 views**
- by taking into account trade-off between **accuracy** and **diversity** between views and view-specific classifiers
- by considering a **non-uniform distribution** over the views
- Derived Multiview $\mathcal{C}$-Bound controlling the trade-off between **accuracy** and **diversity**

Outline

Algorithm 1: $\texttt{Fusion}^{\texttt{all}}_{\texttt{Cq}}$ (CAp'17, ECML-PKDD'17)



1.) **First Level**

↪ Learned with a `Linear SVM` from 60% **of the learning sample**

↪ This step is done **without cross-validation** with different C parameter values

2.) **Second Level**

↪ Learned with `CqBoost` [Roy et al., 2016] from 40% **of the learning sample**

↪ `CqBoost` is PAC-Bayes algorithm based on monoview $\mathcal{C}$-Bound

Algorithm 2: PB-MVBoost (Boosting based algorithm)
(Submitted to Neurocomputing)

Given: $S = \{(\mathbf{x}_i, y_i), \ldots, (\mathbf{x}_m, y_m)\}$, where $\mathbf{x}_i = (x_i^1, x_i^2, \ldots, x_i^V)$ and $y_i \in \{-1, 1\}$.

Initialize: $\mathcal{D}_1(\mathbf{x}_i) \leftarrow 1/m,\ \rho_v^1 \leftarrow 1/V,$ and $H_v \leftarrow \phi$

For $t = 1, \ldots, T$:

1. For each view, learn a weak classifier $h_v^t : \mathcal{X}_v \rightarrow \{-1, 1\}$ w.r.t. distribution $\mathcal{D}_t$
2. Compute classifier's weight: $\forall v \in \mathcal{V}, Q_v^t$
3. $\forall v \in \mathcal{V}, H_v \leftarrow H_v \cup \{h_v^t\}$
4. Update the weights over views (ρ) by optimizing multview C-Bound.
5. Update

$$\mathcal{D}_{t+1}(\mathbf{x}_i) \leftarrow \frac{\mathcal{D}_t(\mathbf{x}_i) \exp(-y_i \sum_{v=1}^{V} \rho_v^t (Q_v^t h_v^t(x_i^v)))}{\sum_{j=1}^{m} \mathcal{D}_t(\mathbf{x}_j) \exp(-y_j \sum_{v=1}^{V} \rho_v^t (Q_v^t h_v^t(x_j^v)))}$$

Output the multiview majority vote classifier:

$$B_\rho^{MV}(\mathbf{x}) = \text{sign}\left[\mathop{\mathbb{E}}_{v \sim \rho} \mathop{\mathbb{E}}_{h \sim Q_v} h(x^v) \right]$$

Algorithm 2: PB−MVBoost (Submitted to Neurocomputing)

Learning the **weights over view-specific classifiers** (view-specific informations):

$$\forall v \in \mathcal{V}, Q_v^t \leftarrow \frac{1}{2}\left[\ln\left(\frac{1-\epsilon_v^t}{\epsilon_v^t}\right)\right]$$

$$\text{where }, \epsilon_v^t \leftarrow \underset{(\mathbf{x}_i, y_i)\sim\mathcal{D}_t}{\mathbb{E}}\left[\mathbb{1}_{[h_v^t(x_i^v)\neq y_i]}\right]$$

Learning the **weights over views** (accuracy and diversity between views):

$$\max_{\rho} \quad \frac{\left(1-2\,\mathbb{E}_{v\sim\rho}\, R_{\mathcal{D}}(G_{Q_v})\right)^2}{1-2\,\mathbb{E}_{v\sim\rho}\, d_{\mathcal{D}}(Q_v)}$$

$$s.t. \quad \sum_{v=1}^{V}\rho_v^t = 1, \quad \rho_v^t \geq 0 \quad \forall v \in \{1, ..., V\}$$

Outline

Datasets (MNIST)

↪ Images of handwritten digits (70K images)

↪ Distributed over 10 classes

↪ Generated 2 four-view datasets where each view is a vector of $\mathbb{R}^{14\times 14}$

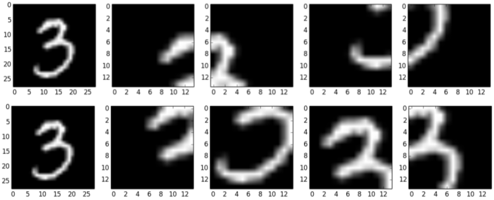
⇒ $\texttt{MNIST}_1$: 4 quarters of image as 4 views

⇒ $\texttt{MNIST}_2$: 4 overlapping views around centre of image

↪ 10K of documents as test samples

$\texttt{MNIST}_1$

$\texttt{MNIST}_2$



Datasets (Reuters RCV1/RCV2)

$\hookrightarrow$ Multilingual text classification corpus (110K documents)

$\hookrightarrow$ Documents written in 5 languages (views / representations)

$\hookrightarrow$ Documents are distributed over 6 classes

$\hookrightarrow$ 30% of documents as test samples

Experimental Protocol

⇒ $\texttt{Fusion}^{\texttt{all}}_{\texttt{Cq}}$: Linear SVM at first level and Cqboost at second level

⇒ PB-MVBoost: Decision Trees as weak learner with $T = 100$ iterations

Baseline Approaches:

⇒ Mono: Learn view-specific model on each view (Decision Trees)

⇒ Concat : One single Decision Trees model (Early Fusion)

⇒ $\texttt{Fusion}_{\texttt{dt}}$: Late fusion approach using Decision Trees at both levels

⇒ MV-MV [Amini et al., 2009]: Multiview uniform majority vote using Decision Trees

⇒ rBoost.SH [Peng et al., 2011]: Boosting based multiview learning algorithm

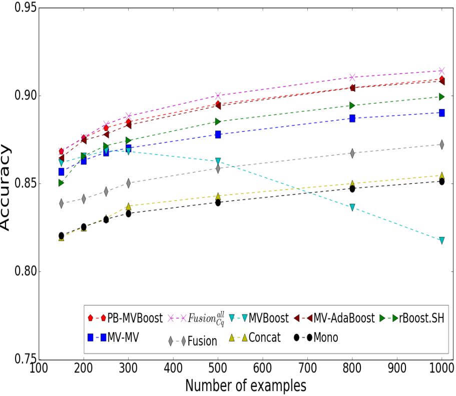
⇒ MV-AdaBoost : Multiview uniform majority vote using Adaboost

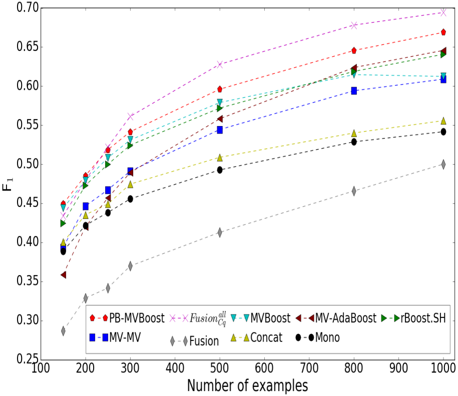
⇒ MV-Boost : Variant of our algorithm PB-MVBoost but without learning weights over views by optimizing multiview $\mathcal{C}$-Bound

Accuracy and F_1-score of different approaches averaged over all the classes and over 20 random sets of $m = 500$ labeled examples per training set.

Strategy	MNIST$_1$		MNIST$_2$		Reuters	
	Accuracy	F_1	Accuracy	F_1	Accuracy	F_1
Mono	.9034±.001	.5353±.006	.9164±.001	.5987±.007	.8420±.002	.5051±.007
Concat	.9224±.002	.6168±.011	.9214±.002	.6142±.013	.8431±.004	.5088±.012
Fusion$_{dt}$	.9320±.001	.5451±.019	.9366±.001	.5937±.020	.8587±.003	.4128±.017
MV-MV	.9402±.001	.6321±.009	.9450±.001	.6849±.008	.8780±.002	.5443±.012
rBoost.SH	.9256±.001	.5315±.009	.9545±.0007	.7258±.005	.8853±.002	.5718±.011
MV-AdaBoost	*.9514*±.001	.6510±.012	*.9641*±.0009	.7776±.007	.8942±.006	.5581±.013
MV-Boost	.9494±.003	*.7733*±.009	.9555±.002	*.7910*±.006	.8627±.007	.5789±.012
Fusion$_{Cq}^{all}$	.9418±.002	.6120±.040	.9548±.003	.7217±.041	**.9001** ±.003	**.6279** ±.019
PB-MVBoost	**.9661**±.0009	**.8066**±.005	**.9674**±.0009	**.8166**±.006	*.8953*±.002	*.5960*±.015

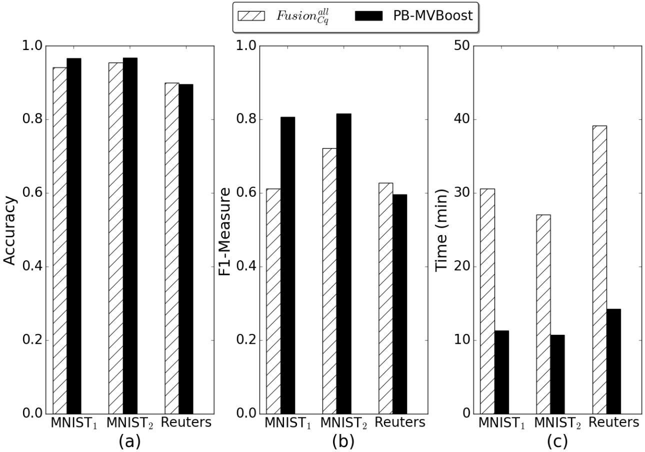
Evolution of Accuracy and F_1 w.r.t. the size of labeled training set





Results

Comparison between $\texttt{Fusion}_{\texttt{Cq}}^{\texttt{all}}$ and $\texttt{PB-MVBoost}$



One-step algorithm $\texttt{PB-MVBoost}$ is more stable and more effective

Results (PB-MVBoost vs. MV-Boost vs. MV-AdaBoost)

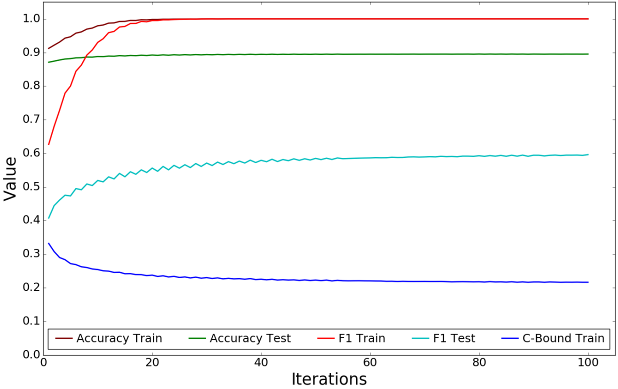
Accuracy and F_1-score of different approaches averaged over all the classes and over 20 random sets of $m = 500$ labeled examples per training set.

Strategy	MNIST$_1$		MNIST$_2$		Reuters	
	Accuracy	F_1	Accuracy	F_1	Accuracy	F_1
Mono	.9034±.001	.5353±.006	.9164±.001	.5987±.007	.8420±.002	.5051±.007
Concat	.9224±.002	.6168±.011	.9214±.002	.6142±.013	.8431±.004	.5088±.012
Fusion$_{dt}$	.9320±.001	.5451±.019	.9366±.001	.5937±.020	.8587±.003	.4128±.017
MV-MV	.9402±.001	.6321±.009	.9450±.001	.6849±.008	.8780±.002	.5443±.012
rBoost.SH	.9256±.001	.5315±.009	.9545±.0007	.7258±.005	.8853±.002	.5718±.011
MV-AdaBoost	*.9514*±.001	.6510±.012	*.9641*±.0009	.7776±.007	.8942±.006	.5581±.013
MV-Boost	.9494±.003	*.7733*±.009	.9555±.002	*.7910*±.006	.8627±.007	.5789±.012
Fusion$^{all}_{Cq}$	.9418±.002	.6120±.040	.9548±.003	.7217±.041	**.9001** ±.003	**.6279** ±.019
PB-MVBoost	**.9661**±.0009	**.8066**±.005	**.9674**±.0009	**.8166**±.006	*.8953*±.002	*.5960*±.015

Two-level hierarchical strategy in a PAC-Bayesian manner is an effective way

Results

Behaviour of `PB-MVBoost` over $T = 100$ iterations for `Reuters` (m = 500) dataset



$\hookrightarrow$ The empirical multiview $\mathcal{C}$-Bound keeps on decreasing over the iterations
$\hookrightarrow$ Control of trade-off between accuracy and diversity between the views

Take Home Message

Designed two multiview learning algorithms based on PAC-Bayesian Theory

- $\texttt{Fusion}_{\texttt{Cq}}^{\texttt{all}}$: Late fusion based algorithm
- `PB-MVBoost`: One-step boosting based algorithm
- `PB-MVBoost` is more stable and effective algorithm for multiview learning

Outline

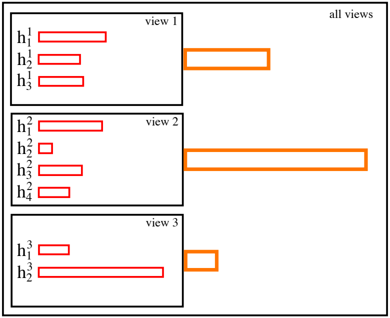
For **each** view $v \in \mathcal{V}$, $\mathcal{H}_v$ is a set of n_v classifiers

$\Longrightarrow$ find weights $\boldsymbol{Q} = (Q_v)_{1 \leq v \leq V}$ over $\mathcal{H}_v$

$\Longrightarrow$ find weights over views $\boldsymbol{\rho} = (\rho_v)_{1 \leq v \leq V}$

Majority Vote: $B^{MV}_{\boldsymbol{\rho}}(\mathbf{x}) = \underset{v \sim \boldsymbol{\rho}}{\mathbb{E}} \; \underset{h_v \sim Q_v}{\mathbb{E}} \; h_v(x^v)$

such that $B^{MV}_{\boldsymbol{\rho}}(\mathbf{x})$ has smallest generalization error on $\mathcal{D}$



Following ERM principle,

Aim $\Longrightarrow$ Minimize $0/1$-loss over training sample S:

$$R_S(B_\rho^{MV}) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{[y_i \neq B_\rho^{MV}(\mathbf{x}_i)]} \leq \frac{1}{m}\sum_{i=1}^{m} \ln\left(1 + \exp\left(-y_i B_\rho^{MV}(\mathbf{x}_i)\right)\right)$$

which is equivalent to the **minimization of a bregman divergence**:

$$D_F\left(\mathbf{0}\middle|\middle| L_F\left(\frac{1}{2}\mathbf{1}_m, \sum_{v=1}^{V} \rho_v \mathbf{M}_v Q_v\right)\right) = \sum_{i=1}^{m} \ln\left(1 + \exp\left(-y_i \mathop{\mathbb{E}}_{v\sim\rho} \mathop{\mathbb{E}}_{h_v \sim Q_v} h_v(x^v)\right)\right)$$

where, $D_F(\mathbf{p}||\mathbf{q}) = \sum_{i=1}^{m} p_i \ln\left(\frac{p_i}{q_i}\right) + (1-p_i)\ln\left(\frac{1-p_i}{1-q_i}\right)$ and $L_F\left(\frac{1}{2}\mathbf{1}_m, \mathbf{z}\right)_i = \frac{1}{(1+e^{z_i})}$

Outline

Given: Training set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$, where $\mathbf{x}_i = (x_i^1, \ldots, x_i^V)$ and $y_i \in \{-1, 1\}$

Initialize: $\boldsymbol{\rho}^{(1)} \leftarrow \frac{1}{V}\mathbf{1}_V$ and $\forall v$, $\mathbf{Q}_v^{(1)} \leftarrow \frac{1}{n_V}\mathbf{1}_{n_v}$
Train the weak classifiers $(\mathcal{H}_v)_{1 \leq v \leq V}$ over S
For $v \in \mathcal{V}$ set the $m \times n_v$ matrix $\mathbf{M}_v$ such that $(\mathbf{M}_v)_{ij} = y_i h_v^j(x_i^v)$

For $t = 1, \ldots, T$:

1. Update weights over examples:

$$\forall i \in \{1, \ldots, m\}, q_i^{(t)} = \sigma\left(y_i \sum_{v=1}^{V} \rho_v^{(t)} \sum_{j=1}^{n_v} Q_v^{j\ (t)}\ h_v^j(x_i^v)\right)$$

2. Update weights $\boldsymbol{Q}$ over the view-specific classifiers
3. Update weights $\boldsymbol{\rho}$ over the views.

Output the weighted multiview majority vote classifier:

$$B_{\boldsymbol{\rho}}^{MV}(\mathbf{x}) = \mathop{\mathbb{E}}_{v\sim\boldsymbol{\rho}}\ \mathop{\mathbb{E}}_{h_v\sim Q_v} h_v(x^v)$$

For each view v, update **weights** $Q_v^{(t+1)}$ **over the view-specific classifiers**:

$$W_{v,j}^{(t)+} = \sum_{i:\text{sign}((\mathbf{M}_v)_{ij})=+1} q_i^{(t)} |(\mathbf{M}_v)_{ij}|$$

$$W_{v,j}^{(t)-} = \sum_{i:\text{sign}((\mathbf{M}_v)_{ij})=-1} q_i^{(t)} |(\mathbf{M}_v)_{ij}|$$

$$Q_v^{j\ (t+1)} = Q_v^{j\ (t)} + \frac{1}{2} \ln \left(\frac{W_{v,j}^{(t)+}}{W_{v,j}^{(t)-}} \right)$$

Update **weights** $\rho^{(t+1)}$ **over the views**:

$$\min_{\rho} \quad -\sum_{v=1}^{V} \rho_v \sum_{j=1}^{n_v} \left(\sqrt{W_{v,j}^{(t)+}} - \sqrt{W_{v,j}^{(t)-}} \right)^2$$

$$\text{s.t.} \quad \sum_{v=1}^{V} \rho_v = 1, \quad \rho_v \geq 0 \quad \forall v \in \mathcal{V}$$

Outline

Datasets

MNIST:

$\hookrightarrow$ Images of handwritten digits (70K images)

$\hookrightarrow$ Distributed over 10 classes

$\hookrightarrow$ Generated 2 four-view datasets where each view is a vector of $\mathbb{R}^{14\times 14}$

$\Rightarrow$ $\texttt{MNIST}_1$: 4 quarters of image as 4 views

$\Rightarrow$ $\texttt{MNIST}_2$: 4 overlapping views around centre of image

$\hookrightarrow$ 10K of documents as test samples

Reuters RCV1/RCV2:

$\hookrightarrow$ Multilingual text classification corpus (110K documents)

$\hookrightarrow$ Documents written in 5 languages (views / representations)

$\hookrightarrow$ Documents are distributed over 6 classes

$\hookrightarrow$ 30% of documents as test samples

Note: Reduced the imbalance between positive and negative examples by subsampling in the training sets

Experimental Protocol

⇒ `MωMvC`^{2}: `Decision Trees` (1 to $\max_d - 2$) as weak learners with $T = 2$ iterations

Baseline Approaches:

⇒ `Mono` : Learn view-specific model on each view (`Decision Trees`)

⇒ `Concat` : One single `Decision Trees` model (Early Fusion)

⇒ `Fusion`: Late fusion approach using `Decision Trees` at both levels

⇒ `MVMLsp` [Huusari et al., 2018] : Multiview metric learning approach.

⇒ `MV-MV` [Amini et al., 2009]: Multiview uniform majority vote using `Decision Trees`

⇒ `rBoost.SH` [Peng et al., 2011]: Boosting based multiview learning algorithm ($T = 100$ iterations)

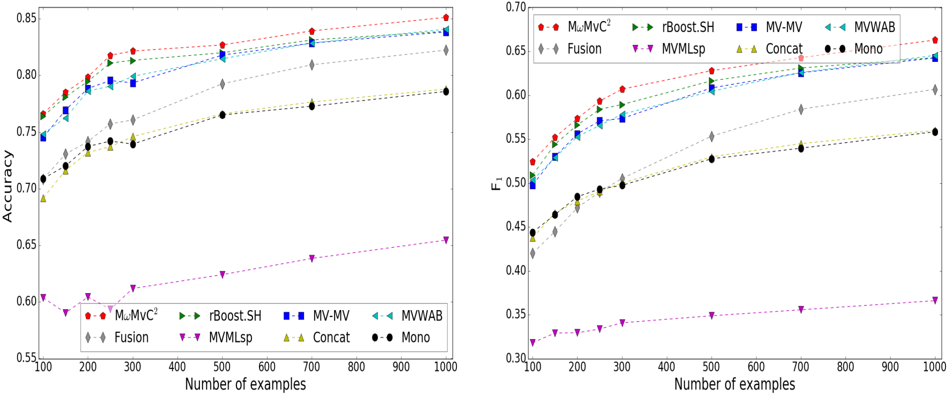
⇒ `MVWAB` [Xiao et al., 2012] : Multiview Weighted Voting AdaBoost algorithm ($T = 100$ iterations)

Accuracy and F_1-score of different approaches averaged over all the classes and over 20 random sets of $m = 500$ labeled examples per training set

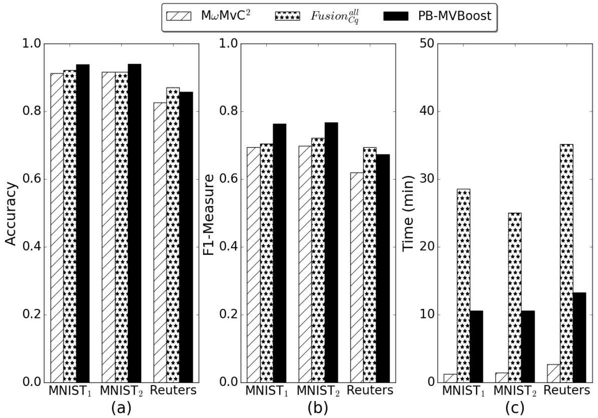
Strategy	MNIST$_1$		MNIST$_2$		Reuters	
	Accuracy	F_1	Accuracy	F_1	Accuracy	F_1
Mono	.8369 ± .002	.5206 ± .003	.8540 ± .003	.5523 ± .004	.7651 ± .005	.5276 ± .005
Concat	.8708 ± .005	.5851 ± .011	.8719 ± .004	.5866 ± .010	.7661 ± .009	.5298 ± .008
Fusion	.8708 ± .005	.5851 ± .010	.9029 ± .009	.6559 ± .018	.7926 ± .013	.5533 ± .015
MVMLsp	.7783 ± .041	.4185 ± .051	.7766 ± .062	.4813 ± .067	.6241 ± .032	.3488 ± .045
MV-MV	.8956 ± .003	.6404 ± .005	.9045 ± .004	.6627 ± .009	.8179 ± .007	.6083 ± .007
MVWAB	.9175 ± .003	.7011 ± .009	.9038 ± .003	.6838 ± .008	.8147 ± .007	.6045 ± .009
rBoost.SH	.7950 ± .006	.4652 ± .006	.8762 ± .004	.6089 ± .007	.8200 ± .007	.6164 ± .007
MωMvC^2	**.9260** ± .004	**.7122** ± .010	**.9169** ± .005	**.6977** ± .012	**.8269** ± .013	**.6280** ± .010

Two-level hierarchical strategy is an effective way to handle multiview learning

Evolution of Accuracy and F_1 w.r.t. the size of labeled training set



Comparison (MωMvC^2 vs. PB-MVBoost vs. $\text{Fusion}_{\text{Cq}}^{\text{all}}$)



- ↪ MωMvC^2 is faster than PB-MVBoost and $\text{Fusion}_{\text{Cq}}^{\text{all}}$
- ↪ PB-MVBoost: $O\Big(T\big(V\,d_v\,m.log(m) + V^3\big)\Big)$ and MωMvC^2: $O\Big(V\,d_v\,m.log(m) + T\,V^3\Big)$
- ↪ PB-MVBoost can handle the imbalance between classes
- ↪ PB-MVBoost controls the trade-off between accuracy and diversity between the views

Take Home Message

Minimization of the multiview classification error is equivalent to the minimization of Bregman divergences

- parallel-update optimization boosting-like algorithm ($\mathtt{M}\omega\mathtt{MvC}^2$)
- Computationally faster than $\mathtt{Fusion}^{\mathtt{all}}_{\mathtt{Cq}}$ and $\mathtt{PB\text{-}MVBoost}$

Outline

Conclusion

Theoretical point of view:

- A non-probabilistic PAC-Bayesian generalization bound
- Instantiation of PAC-Bayesian theory to multiview learning with more than 2 views

  $\hookrightarrow$ Considering hierarchy of distributions over the view-specific classifiers

Algorithmic point of view:

- Late fusion based two-step multiview learning algorithm $\texttt{Fusion}_{\texttt{Cq}}^{\texttt{all}}$
- One-step boosting based multiview learning algorithm $\texttt{PB-MVBoost}$

  $\hookrightarrow$ Optimizes multiview $\mathcal{C}$-Bound

  $\hookrightarrow$ Controls the accuracy and diversity between views
- Multiview Learning as Bregman Divergence Minimization

  $\hookrightarrow$ Parallel update boosting like multiview learning algorithm $\texttt{M}\omega\texttt{MvC}^2$

Perspectives

- Specialize our PAC-Bayesian generalization bounds to linear classifiers
- Suitable **stopping criteria** for PB-MVBoost

  $\hookrightarrow$ Analyze the margins of training examples
- Extension of our algorithms to **semi-supervised** multiview learning

  $\hookrightarrow$ Learn view-specific classifiers using pseudo-labels (for unlabeled data) generated from other view-specific classifiers

  $\hookrightarrow$ For PB-MVBoost, use unlabeled data while computing view-specific disagreement for optimizing multiview $\mathcal{C}$-Bound
- Extension of our algorithms to the case of **missing views or incomplete views**

  $\hookrightarrow$ For PB-MVBoost, learn view-specific classifiers using available training examples and adapt the distribution over learning sample accordingly

  $\hookrightarrow$ For $\mathtt{M\omega MvC}^2$, adapt the definition of the input matrix $\mathbf{M}_v$

Thank you for your attention

List of Publications

- Anil Goyal, Emilie Morvant, Pascal Germain, Massih-Reza Amini
  Multiview Boosting by Controlling the Diversity and the Accuracy of View-specific Voters
  Neurocomputing (Submitted)
- Anil Goyal, Emilie Morvant, Massih-Reza Amini
  Multiview Learning of Weighted Majority Vote by Bregman Divergence Minimization
  Intelligent Data Analysis (IDA), 2018
- Anil Goyal, Emilie Morvant, Massih-Reza Amini
  Apprentissage d'un vote de majorité hiérarchique pour l'apprentissage multivue
  Conférence sur l'Apprentissage Automatique (CAp), 2018
- Anil Goyal, Emilie Morvant, Pascal Germain, Massih-Reza Amini
  PAC-Bayesian Analysis for a two-step Hierarchical Mutliview Learning Approach
  European Conference on Machine Learning & Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD), 2017
- Anil Goyal, Emilie Morvant, Pascal Germain
  Une borne PAC-Bayésienne en espérance et son extension à l'apprentissage multivues
  Conférence sur l'Apprentissage Automatique(CAp), 2017
- Anil Goyal, Emilie Morvant, Pascal Germain, Massih-Reza Amini
  Théorèmes PAC-Bayésiens pour l'apprentissage multivues Conférence sur l'Apprentissage Automatique (CAp), 2016

References

Massih-Reza Amini, Nicolas Usunier, and Cyril Goutte.
Learning from Multiple Partially Observed Views - an Application to Multilingual Text Categorization.
In *NIPS*, pages 28–36, 2009.

Pascal Germain, Alexandre Lacasse, François Laviolette, Mario Marchand, and Jean-Francis Roy.
Risk bounds for the majority vote: from a PAC-Bayesian analysis to a learning algorithm.
JMLR, 16:787–860, 2015.

R. Huusari, H. Kadri, and C. Capponi.
Multi-view Metric Learning in Vector-valued Kernel Spaces.
In *AISTATS*, 2018.

David A. McAllester.
PAC-Bayesian stochastic model selection.
In *Machine Learning*, pages 5–21, 2003.

J. Peng, A. J. Aved, G. Seetharaman, and K. Palaniappan.
Multiview boosting with information propagation for classification.
IEEE Transactions on Neural Networks and Learning Systems, PP(99):1–13, 2017.

Jing Peng, Costin Barbu, Guna Seetharaman, Wei Fan, Xian Wu, and Kannappan Palaniappan.
Shareboost: Boosting for multi-view learning with performance guarantees.
In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part II*, pages 597–612, 2011.

Jean-Francis Roy, Mario Marchand, and FranÃ§ois Laviolette.
A column generation bound minimization approach with PAC-Bayesian generalization guarantees.
In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 1241–1249, 2016.

Min Xiao and Yuhong Guo.
Multi-view adaboost for multilingual subjectivity analysis.
In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 2851–2866, 2012.

Multiview probabilistic PAC-Bayes Bound

Monoview bound

$$\mathop{\mathbf{Prob}}_{S\sim\mathcal{D}^m}\left(D\Big(R_{\mathcal{D}}(G_Q), R_S(G_Q)\Big) \leq \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\left(\mathop{\mathbb{E}}_{h\sim P} e^{m\,D(R_S(h), R_{\mathcal{D}}(h))}\right)\right]\right) \geq 1-\delta$$

Proposed Bound for Multiview

For any $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$,for any $\delta\in(0,1]$, with a probability at least $1-\delta$ over the random choice of $S\sim(\mathcal{D})^m$, for all posterior $\{Q_v\}_{v=1}^{V}$ and hyper-posterior ρ distributions, for any convex function $D : [0,1]\times[0,1]\to\mathbb{R}$, we have

$$D\Big(R_{\mathcal{D}}(G_\rho^{MV}), \underbrace{\frac{1}{2}d_S^{\mathrm{MV}}(\rho) + e_S^{\mathrm{MV}}(\rho)}_{R_S(G_\rho^{MV})}\Big)$$

$$\leq \frac{1}{m}\left[\mathop{\mathbb{E}}_{v\sim\rho}\mathrm{KL}(Q_v\|P_v) + \mathrm{KL}(\rho\|\pi) + \ln\left(\frac{1}{\delta}\mathop{\mathbb{E}}_{v\sim\pi}\mathop{\mathbb{E}}_{h\sim P_v} e^{mD(R_S(h), R_{\mathcal{D}}(h))}\right)\right]$$

Multiview Non-probabilistic PAC-Bayes Bound

Monoview bound

$$D\left(\mathbb{E}_{S\sim\mathcal{D}^m} R_{\mathcal{D}}(G_{Q_S}), \mathbb{E}_{S\sim\mathcal{D}^m} R_S(G_{Q_S})\right) \le \frac{1}{m}\left[\mathbb{E}_{S\sim\mathcal{D}^m} \mathrm{KL}(Q_S\|P) + \ln\left(\mathbb{E}_{S\sim\mathcal{D}^m}\mathbb{E}_{h\sim P} e^{m\,D(R_S(h),R_{\mathcal{D}}(h))}\right)\right]$$

Proposed Bound for Multiview

For any $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, for any set of priors $\{P_v\}_{v=1}^V$, for any hyper-priors π over $\mathcal{V}$, for any convex function $D : [0,1]\times[0,1]\to\mathbb{R}$, we have

$$D\Bigg(\mathbb{E}_{S\sim\mathcal{D}^m} R_{\mathcal{D}}(G^{MV}_{\rho_S}), \underbrace{\frac{1}{2}\mathbb{E}_{S\sim\mathcal{D}^m} d^{\mathrm{MV}}_S(\rho_S) + \mathbb{E}_{S\sim\mathcal{D}^m} e^{\mathrm{MV}}_S(\rho_S)}_{R_S(G^{MV}_{\rho_S})}\Bigg)$$

$$\le \frac{1}{m}\left[\mathbb{E}_{S\sim\mathcal{D}^m}\mathbb{E}_{v\sim\rho_S} \mathrm{KL}(Q_{v,S}\|P_v) + \mathbb{E}_{S\sim\mathcal{D}^m} \mathrm{KL}(\rho_S\|\pi) + \ln\left(\mathbb{E}_{S\sim\mathcal{D}^m}\mathbb{E}_{v\sim\pi}\mathbb{E}_{h\sim P_v} e^{mD(R_S(h),R_{\mathcal{D}}(h))}\right)\right]$$

Square Root Bound

Obtained using $D(a,b) = 2(a-b)^2$

For any $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, for any set of priors $\{P_v\}_{v=1}^{V}$, for any hyper-priors π over $\mathcal{V}$, we have

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} R_{\mathcal{D}}(G_{\rho_S}^{MV}) \leq \underbrace{\frac{1}{2}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} d_S^{\mathrm{MV}}(\rho_S) + \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} e_S^{\mathrm{MV}}(\rho_S)\}}_{\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} R_S(G_{\rho_S}^{MV})}$$

$$+ \sqrt{\frac{\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{v\sim\rho_S} \mathrm{KL}(Q_{v,S}\|P_v) + \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \mathrm{KL}(\rho_S\|\pi) + \ln 2\sqrt{m}}{2m}}$$

Trade-off:

- Empirical disagreement and joint error
- Expectation of view-specific KL divergences over all the views
- KL divergence between hyper-posterior and hyper-prior

Links the true risk and the empirical risk by a linear relation

Parametrized Bound

Obtained using $D(a,b) = \mathcal{F}(b) - C\,a$

For any $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, for any set of priors $\{P_v\}_{v=1}^{V}$, for any hyper-priors π over $\mathcal{V}$, for all $C > 0$ we have

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} R_{\mathcal{D}}(G_{\rho_S}^{\mathrm{MV}}) \le \frac{1}{1-e^{-C}}\Bigg(1 - \exp\Bigg[-\Bigg[C \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} R_S(G_{\rho_S}^{MV}) + \frac{1}{m}\Big[\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{v\sim\rho_S} \mathrm{KL}(Q_{v,S}\|P_v) + \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathrm{KL}(\rho_S\|\pi)\Big]\Bigg]\Bigg]\Bigg)$$

Explicitly controls the trade-off between the empirical risk and the KL divergence terms using the hyperparameter C

Parametrized Bound

Restricting $C \in (0,2)$ and using $e^{-C} \leq 1 - C - \frac{1}{2}C^2$, we can obtain looser but simpler bound

For any $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, for any set of priors $\{P_v\}_{v=1}^{V}$, for any hyper-priors π over $\mathcal{V}$, for all $C > 0$ we have

$$\underset{S\sim\mathcal{D}^m}{\mathbb{E}} R_{\mathcal{D}}(G^{\mathrm{MV}}_{\rho_S}) \leq \frac{1}{1-\frac{1}{2}C}\left(\underset{S\sim\mathcal{D}^m}{\mathbb{E}} R_S(G^{MV}_{\rho_S}) + \frac{\underset{S\sim\mathcal{D}^m}{\mathbb{E}}\ \underset{v\sim\rho_S}{\mathbb{E}} \mathrm{KL}(Q_{v,S}\|P_v) + \underset{S\sim\mathcal{D}^m}{\mathbb{E}} \mathrm{KL}(\rho_S\|\pi)}{m \times C}\right)$$

Choosing $C = \frac{1}{\sqrt{m}}$ the bound converges to $1 \times \left[R_S(G^{MV}_{\rho_S}) + 0\right]$ as m grows

Small kl Bound

Obtained using $D(a,b) \;=\; \mathrm{kl}(a,b) \;=\; a\ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b}$

For any $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, for any set of priors $\{P_v\}_{v=1}^{V}$, for any hyper-priors π over $\mathcal{V}$, we have

$$\mathrm{kl}\left(\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} R_S(G^{MV}_{\rho_S}),\ \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} R_{\mathcal{D}}(G^{\mathrm{MV}}_{\rho_S})\right)$$
$$\le \frac{1}{m}\left[\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{v\sim\rho_S}\mathrm{KL}(Q_{v,S}\|P_v) + \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathrm{KL}(\rho_S\|\pi) + \ln\ 2\sqrt{m}\right]$$

For upper bound value, one needs to solve:

$$\begin{aligned}\max\quad & b\\ s.t.\quad & \mathrm{kl}\left(\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} R_S(G^{MV}_{\rho_S})\ \|\ b\right) = \frac{1}{m}\left[\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{v\sim\rho_S}\mathrm{KL}(Q_{v,S}\|P_v) + \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathrm{KL}(\rho_S\|\pi) + \ln\ 2\sqrt{m}\right]\\ & 0\le b\le 1.\end{aligned}$$

Parametrized bound and Small kl bound

Proposition (Germain et al. , 2009)

For $0 \leq \underset{S\sim\mathcal{D}^m}{\mathbb{E}} R_S(G^{MV}_{\rho_S}) \leq \underset{S\sim\mathcal{D}^m}{\mathbb{E}} R_{\mathcal{D}}(G^{MV}_{\rho_S}) \leq 1$, we have

$$\max_{C\geq 0}\left\{-\ln\left(1-\underset{S\sim\mathcal{D}^m}{\mathbb{E}} R_{\mathcal{D}}(G^{MV}_{\rho_S})\left[1-e^{-C}\right]\right)-C\underset{S\sim\mathcal{D}^m}{\mathbb{E}} R_S(G^{MV}_{\rho_S})\right\} = \mathrm{kl}\left(\underset{S\sim\mathcal{D}^m}{\mathbb{E}} R_S(G^{MV}_{\rho_S}), \underset{S\sim\mathcal{D}^m}{\mathbb{E}} R_{\mathcal{D}}(G^{MV}_{\rho_S})\right)$$

$\hookrightarrow$ Small kl bound is tighter or equal to Parametrized bound

$\hookrightarrow$ There always exists values of C for which Parametrized bound is tighter than Small kl bound

A Generalization Bound for the Multiview C-Bound

Let $V \geq 2$ be the number of views. For any distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, for any set of prior distributions $\{P_v\}_{v=1}^{V}$, for any hyper-prior distributions π over views $\mathcal{V}$, and for any convex function $D : [0,1] \times [0,1] \to \mathbb{R}$, with a probability at least $1-\delta$ over the random choice of $S \sim (D)^m$ for all posterior $\{Q_v\}_{v=1}^{v}$ and hyper-posterior ρ distributions, we have:

$$R_{\mathcal{D}}(B_{\rho}^{MV}) \leq 1 - \frac{\left(1 - 2 \mathop{\mathbb{E}}_{v\sim\rho} \sup \left(\mathbf{r}_{Q_v,\mathcal{S}}^{\delta/2}\right)\right)^2}{1 - 2 \mathop{\mathbb{E}}_{v\sim\rho} \inf \mathbf{d}_{Q_v,\mathcal{S}}^{\delta/2}},$$

where

$$\mathbf{r}_{Q_v,\mathcal{S}}^{\delta/2} = \left\{ r : \mathrm{kl}(R_S(G_{Q_v}) \| r) \leq \frac{1}{n}\left[\mathrm{KL}(Q_v \| P_v) + \ln \frac{4\sqrt{m}}{\delta}\right] \text{ and } r \leq \frac{1}{2} \right\}$$

$$\text{and} \quad \mathbf{d}_{Q_v,\mathcal{S}}^{\delta/2} = \left\{ d : \mathrm{kl}(d_{Q_v}^{S} \| d) \leq \frac{1}{n}\left[2.\,\mathrm{KL}(Q_v \| P_v) + \ln \frac{4\sqrt{m}}{\delta}\right] \right\}$$

Bregman-Divergence

Let $\Omega \subseteq \mathbb{R}^m$ and $F : \Omega \to \mathbb{R}$ be a continuously differentiable and strictly convex real-valued function. The Bregman divergence D_F associated to F is defined for all $(\mathbf{p}, \mathbf{q}) \in \Omega \times \Omega$ as

$$D_F(\mathbf{p}||\mathbf{q}) = F(\mathbf{p}) - F(\mathbf{q}) - \langle \nabla F(\mathbf{q}), (\mathbf{p} - \mathbf{q}) \rangle \, ,$$

where $\nabla F(\mathbf{q})$ is the gradient of F estimated at $\mathbf{q}$, and the operator $\langle \cdot, \cdot \rangle$ is the dot product function.

For our multiview learning setting, we consider

$$F(\mathbf{p}) = \sum_{i=1}^{m} p_i \ln(p_i) + (1 - p_i) \ln(1 - p_i)$$

Bregman-divergence is defined as

$$D_F(\mathbf{p}||\mathbf{q}) = \sum_{i=1}^{m} p_i \ln\left(\frac{p_i}{q_i}\right) + (1 - p_i) \ln\left(\frac{1 - p_i}{1 - q_i}\right)$$

Bregman-divergence optimization

Find a vector $\mathbf{p}^\star \in \Omega$—that is the closest to a given vector $\mathbf{q}_0 \in \Omega$—under the set $\mathcal{P}$ of V linear constraints such that

$$\operatorname*{argmin}_{p\in\mathcal{P}} \; D_F(\mathbf{p}||\mathbf{q}_0)$$

$$\text{s.t.} \quad \mathcal{P} = \{\mathbf{p} \in \Omega | \forall v \in [V],\; \rho_v \mathbf{p}^\top \mathbf{M}_v = \rho_v \tilde{\mathbf{p}}^\top \mathbf{M}_v\}$$

Solving above optimization problem using the Langrangian multipliers, we have

$$K = D_F(\mathbf{p}||\mathbf{q}_0) + \sum_{v=1}^{V} \left(\rho_v \mathbf{p}^\top \mathbf{M}_v - \rho_v \tilde{\mathbf{p}}^\top \mathbf{M}_v\right) Q_v$$

Differentiating K w.r.t. $\mathbf{p}$ and Q_v, the original optimization reduced to minimization of

$$D_F\left(\mathbf{0}\middle|\middle| L_F\left(\frac{1}{2}\mathbf{1}_m, \sum_{v=1}^{V} \rho_v \mathbf{M}_v Q_v\right)\right) = \sum_{i=1}^{m} \ln\left(1 + \exp\left(-y_i \sum_{v=1}^{V} \rho_v \sum_{j=1}^{n_v} Q_v^j h_v^j(x_i^v)\right)\right)$$

$$\text{where, } L_F\left(\frac{1}{2}\mathbf{1}_m, \mathbf{z}\right)_i = \frac{1}{(1 + e^{z_i})}$$

Multiview Parallel Update Algorithm

Given: Training set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, where $\mathbf{x}_i = (x_i^1, \ldots, x_i^V)$ and $y_i \in \{-1, 1\}$

Initialize: $\boldsymbol{\rho}^{(1)} \leftarrow \frac{1}{V}\mathbf{1}_V$ and $\forall v$, $\mathbf{Q}_v^{(1)} \leftarrow \frac{1}{n_V}\mathbf{1}_{n_v}$
Train the weak classifiers $(\mathcal{H}_v)_{1 \le v \le V}$ over S
For $v \in \mathcal{V}$ set the $m \times n_v$ matrix $\mathbf{M}_v$ such that $(\mathbf{M}_v)_{ij} = y_i h_v^j(x_i^v)$

Using the current parameters $\boldsymbol{\rho}^{(t)}$, $\boldsymbol{Q}^{(t)}$ and $\mathbf{q}^{(t)} \in \mathcal{Q}_0$, we update

$$\mathbf{q}^{(t+1)} = L_F\left(\frac{1}{2}\mathbf{1}_m, \sum_{v=1}^{V} \rho_v^{(t+1)} \mathbf{M}_v (Q_v^{(t)} + \boldsymbol{\delta}_v^{(t)})\right),$$

such that $D_F(0||\mathbf{q}^{(t+1)}) \le D_F(0||\mathbf{q}^{(t)})$.

At each iteration of algorithm, following inequality holds:

$$D_F(\mathbf{0}||\mathbf{q}^{(t+1)}) - D_F(\mathbf{0}||\mathbf{q}^{(t)}) \le -\sum_{v=1}^{V} \rho_v^{(t+1)} \sum_{j=1}^{n_v} \left(\sqrt{W_{v,j}^{(t)+}} - \sqrt{W_{v,j}^{(t)-}}\right)^2$$