Learning from Multiple Partially Observed Views – an Application to Multilingual Text Categorization Supplementary material

Massih R. Amini
Interactive Language Technologies Group
National Research Council Canada
Massih-Reza.Amini@cnrc-nrc.gc.ca

Nicolas Usunier
Laboratoire d'Informatique de Paris 6
Université Pierre et Marie Curie, France
Nicolas.Usunier@lip6.fr

Cyril Goutte
Interactive Language Technologies Group
National Research Council Canada
Cyril.Goutte@cnrc-nrc.gc.ca

Appendix: Proof of Theorem 1

For clarity, we recall the theorem:

Theorem 1 *Let $\mathcal{D}$ be a distribution over $\mathcal{X}$, satisfying $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left(|\{v : x^v \neq \perp\}| \neq 1\right) = 0$. Let $\mathcal{S} = ((\mathbf{x}_i, y_i))_{i=1}^m$ be a dataset of m examples drawn i.i.d. according to $\mathcal{D}$. Let e be the $0/1$ loss, and let $(\mathcal{H}_v)_{v=1}^V$ be the view-specific deterministic classifier sets. For each view v, denote $e \circ \mathcal{H}_v \stackrel{def}{=} \{(x^v, y) \mapsto e(h, (x^v, y)) | h \in \mathcal{H}_v\}$, and denote , for any sequence $\mathcal{S}^v \in (\mathcal{X}_v \times \mathcal{Y})^{m_v}$ of size m_v, $\hat{\mathcal{R}}_{m_v}(e \circ \mathcal{H}_v, \mathcal{S}^v)$ the empirical Rademacher complexity of $e \circ \mathcal{H}_v$ on $\mathcal{S}^v$. Then, we have:*

Baseline setting *for all $1 > \delta > 0$, with probability at least $1 - \delta$ over $\mathcal{S}$:*

$$\epsilon(c^b_{h_1,\ldots,h_V}) \leq \inf_{h'_v \in \mathcal{H}_v} \left[\epsilon(c^b_{h'_1,\ldots,h'_V})\right] + 2\sum_{v=1}^V \frac{m_v}{m} \hat{\mathcal{R}}_{m_v}(e \circ \mathcal{H}_v, \mathcal{S}^v) + 6\sqrt{\frac{\ln(2/\delta)}{2m}}$$

where, for all v, $\mathcal{S}^v \stackrel{def}{=} \{(x_i^v, y_i) | i = 1..m \text{ and } x_i^v \neq \perp\}$, $m_v = |\mathcal{S}^v|$ and $h_v \in \mathcal{H}_v$ is the empirical risk minimizer on $\mathcal{S}^v$.

Multi-view Gibbs classification setting *for all $1 > \delta > 0$, with probability at least $1 - \delta$ over $\mathcal{S}$:*

$$\epsilon(c^{mg}_{h_1,\ldots,h_V}) \leq \inf_{h'_v \in \mathcal{H}_v} \left[\epsilon(c^b_{h'_1,\ldots,h'_V})\right] + \frac{2}{V}\sum_{v=1}^V \hat{\mathcal{R}}_m(e \circ \mathcal{H}_v, \underline{\mathcal{S}}^v) + 6\sqrt{\frac{\ln(2/\delta)}{2m}} + \eta$$

where, for all v, $\underline{\mathcal{S}}^v \stackrel{def}{=} \{(\underline{x}_i^v, y_i) | i = 1..m\}$, and $h_v \in \mathcal{H}_v$ is the empirical risk minimizer on $\mathcal{S}^v$.

$$\eta = \inf_{h'_v \in \mathcal{H}_v} \left[\epsilon(c^{mg}_{h'_1,\ldots,h'_V})\right] - \inf_{h'_v \in \mathcal{H}_v} \left[\epsilon(c^b_{h'_1,\ldots,h'_V})\right] \quad (1)$$

Proof for the Baseline Setting

We start by proving the baseline setting. Fix a dataset $\mathcal{S}$, and let $h_1, ..., h_V$ be the empirical risk minizers on each view, trained with the examples for which the corresponding view is observed. That is:

$$\forall v, h_v \in \arg\min_{h \in \mathcal{H}_v} \sum_{(\mathbf{x},y)\in S : x^v \neq \perp} e(h, (x^v, y)) \quad (2)$$

Considering that $c^b_{h_1,\ldots,h_V}$ classifies an instance according to the classifier of the observed view, we can notice that $c^b_{h_1,\ldots,h_V}$ is exactly the empirical risk minimizer over $\mathcal{S}$ for the set of classifiers $\mathcal{C}^b = \left\{c^b_{h'_1,\ldots,h'_V} | \forall v, h'_v \in \mathcal{H}_\mathcal{V}\right\}$ (recall that exactly one view is observed for each example) with $h_v, v \in \mathcal{V}$ defined as in Equation (2), we have:

$$c^b_{h_1,\ldots,h_V} \in \underset{c^b_{h'_1,\ldots,h'_V} \in \mathcal{C}^b}{\arg\min} \sum_{(\mathbf{x},y)\in\mathcal{S}} e(c^b_{h'_1,\ldots,h'_V}, (x^v, y))$$

We then start by using the well-known inequality that bounds the generalization error between the empirical risk minimizer and the error of the best-in class (see e.g. lemma 1.1 of [2]):

$$\epsilon(c^b_{h_1,\ldots,h_V}) - \inf_{c^b_{h'_1,\ldots,h'_V} \in \mathcal{C}^b} \left[\epsilon(c^b_{h'_1,\ldots,h'_V})\right] \leq 2 \sup_{c^b_{h'_1,\ldots,h'_V} \in \mathcal{C}^b} \left|\hat{\epsilon}\left(c^b_{h'_1,\ldots,h'_V}, \mathcal{S}\right) - \epsilon(c^b_{h'_1,\ldots,h'_V})\right| \quad (3)$$

where $\hat{\epsilon}\left(c^b_{h'_1,\ldots,h'_V}, \mathcal{S}\right) = \frac{1}{|\mathcal{S}|}\sum_{(\mathbf{x},y)\in\mathcal{S}} e\left(c^b_{h'_1,\ldots,h'_V}, (\mathbf{x}, y)\right)$ is the empirical risk, on the set $\mathcal{S}$, of $c^b_{h'_1,\ldots,h'_V}$.

Since e is the $0/1$ loss, we have $0 \leq e\left(c^b_{h'_1,\ldots,h'_V}, (\mathbf{x}, y)\right) \leq 1$ for any $(\mathbf{x}, y)$. We can thus use the standard Rademacher complexity analysis to obtain a data-dependent bound on the right-hand term of Equation (3), so we can apply McDiarmid's theorem [4] to the function $\mathcal{S} \mapsto \sup_{c^b_{h'_1,\ldots,h'_V} \in \mathcal{C}^b} \left|\hat{\epsilon}\left(c^b_{h'_1,\ldots,h'_V}, \mathcal{S}\right) - \epsilon(c^b_{h'_1,\ldots,h'_V})\right|$, which can not change by more than $1/m$ when a pair $(\mathbf{x}, y)$ changes. We then have, with probability at least $1 - \delta/2$:

$$\sup_{c^b_{h'_1,\ldots,h'_V} \in \mathcal{C}^b} \left|\hat{\epsilon}\left(c^b_{h'_1,\ldots,h'_V}, \mathcal{S}\right) - \epsilon(c^b_{h'_1,\ldots,h'_V})\right| \leq \mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m} \sup_{c^b_{h'_1,\ldots,h'_V} \in \mathcal{C}^b} \left|\hat{\epsilon}\left(c^b_{h'_1,\ldots,h'_V}, \mathcal{S}'\right) - \epsilon(c^b_{h'_1,\ldots,h'_V})\right| + \sqrt{\frac{\ln(2/\delta)}{2m}} \quad (4)$$

where $m = |\mathcal{S}|$.

We will now use the classical definition of the Rademacher complexity (see e.g. [1]) for some class of function $\mathcal{F}$ defined on the input space $\mathcal{X} \times \mathcal{Y}$:

$$\mathcal{R}_m(\mathcal{F}) \stackrel{def}{=} \mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m} \hat{\mathcal{R}}_m(\mathcal{F}, \mathcal{S}')$$

where $\hat{\mathcal{R}}_m(\mathcal{F}, \mathcal{S}')$ is the empirical Rademacher complexity of $\mathcal{F}$ on the dataset $\mathcal{S}'$, defined as:

$$\hat{\mathcal{R}}_m(\mathcal{F}, \mathcal{S}') \stackrel{def}{=} \mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_m} \sup_{f\in\mathcal{F}} \left|\frac{2}{m}\sum_{i=1}^{m} \sigma_i f((\mathbf{x}'_i, y'_i))\right|$$

where σ_i are independent random variables such that $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 1/2$. With standard arguments of the Rademacher complexity analysis, we have, with probability at least $1 - \delta/2$:

$$\mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^m} \sup_{c^b_{h'_1,\ldots,h'_V} \in \mathcal{C}^b} \left|\hat{\epsilon}\left(c^b_{h'_1,\ldots,h'_V}, \mathcal{S}'\right) - \epsilon(c^b_{h'_1,\ldots,h'_V})\right| \leq \hat{\mathcal{R}}_m(e \circ \mathcal{C}^b, \mathcal{S}) + \sqrt{\frac{2\ln(2/\delta)}{m}} \quad (5)$$

where $e \circ \mathcal{C}^b = \left\{(\mathbf{x}, y) \mapsto e(c^b_{h'_1,\ldots,h'_V}, (\mathbf{x}, y)) | c^b_{h'_1,\ldots,h'_V} \in \mathcal{C}^b\right\}$. Plugging Equation (5) into Equations (4) and (3), we obtain, with probability at least $1 - \delta$:

$$\epsilon(c^b_{h_1,\ldots,h_V}) \leq \inf_{c^b_{h'_1,\ldots,h'_V} \in \mathcal{C}^b} \left[\epsilon(c^b_{h'_1,\ldots,h'_V})\right] + 2\hat{\mathcal{R}}_m(e \circ \mathcal{C}^b, \mathcal{S}) + 6\sqrt{\frac{\ln(2/\delta)}{2m}} \quad (6)$$

The final step consists in re-writing the empirical Rademacher complexity $\hat{\mathcal{R}}_m(e \circ \mathcal{C}^b, \mathcal{S})$ depending on the Rademacher complexity of the view-specific classifier sets. Given the dataset $\mathcal{S}$, we define,

for each view v, the partial dataset $\mathcal{S}^v \stackrel{def}{=} \{(x_i^v, y_i) | i = 1..m \text{ and } x_i^v \neq \perp\}$. We use a specific index notation for the examples in $\mathcal{S}^v$: $\mathcal{S}^v = \left\{(x_{i_k^v}^v, y_{i_k^v}), k = 1..m_v\right\}$

Starting from the definition, we have:

$$\begin{aligned}
\hat{\mathcal{R}}_m(e \circ \mathcal{C}^b, \mathcal{S}) &= \mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_m} \sup_{c^b_{h'_1,\ldots,h'_V} \in \mathcal{C}^b} \left| \frac{2}{m} \sum_{i=1}^{m} \sigma_i e\left(c^b_{h'_1,\ldots,h'_V}, (\mathbf{x}_i, y_i)\right) \right| \\
&= \mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_m} \sup_{c^b_{h'_1,\ldots,h'_V} \in \mathcal{C}^b} \left| \frac{2}{m} \sum_{v=1}^{V} \sum_{i:x_i^v \neq \perp} \sigma_i e\left(h'_v, (x_i^v, y_i)\right) \right| \\
&\leq \sum_{v=1}^{V} \frac{m_v}{m} \mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_{m_v}} \sup_{h'_v \in \mathcal{H}_v} \left| \frac{2}{m_v} \sum_{k=1}^{m_v} \sigma_k e\left(h'_v, (x_{i_k^v}^v, y_{i_k^v})\right) \right| \\
&= \sum_{v=1}^{V} \frac{m_v}{m} \hat{\mathcal{R}}_{m_v}(e \circ \mathcal{H}_v, \mathcal{S}^v)
\end{aligned}$$

Together with Equation (6) gives the desired result.

Proof for the Multi-view Gibbs Classification Setting

The proof follows the same steps, replacing $\mathcal{C}^b$ by $\mathcal{C}^{mg} \stackrel{def}{=} \left\{c^{mg}_{h'_1,\ldots,h'_V} | \forall v, h'_v \in \mathcal{H}_\mathcal{V}\right\}$. Only the last step (the calculation of the empirical Rademacher complexity) has to be modified. Remind that the true and empirical risks of the multi-view Gibbs classifier are the average of empirical and true risks of the view-specific classifiers, as the multi-view Gibbs classifier is supposed to be drawn from a uniform posterior distribution [3].

Re-starting from the definition of the ampirical Rademacher complexity, we have:

$$\begin{aligned}
\hat{\mathcal{R}}_m(e \circ \mathcal{C}^{mg}, \mathcal{S}) &= \mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_m} \sup_{c^{mg}_{h'_1,\ldots,h'_V} \in \mathcal{C}^b} \left| \frac{2}{m} \sum_{i=1}^{m} \sigma_i e\left(c^{mg}_{h'_1,\ldots,h'_V}, (\mathbf{x}_i, y_i)\right) \right| \\
&= \mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_m} \sup_{c^{mg}_{h'_1,\ldots,h'_V} \in \mathcal{C}^b} \left| \frac{2}{mV} \sum_{v=1}^{V} \sum_{i=1}^{m} \sigma_i e\left(h'_v, (\underline{x}_i^v, y_i)\right) \right| \\
&\leq \frac{1}{V} \sum_{v=1}^{V} \mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_m} \sup_{h'_v \in \mathcal{H}_v} \left| \frac{2}{m} \sum_{i=1}^{m} \sigma_i e\left(h'_v, (\underline{x}_i^v, y_i)\right) \right| \\
&= \frac{1}{V} \sum_{v=1}^{V} \hat{\mathcal{R}}_m(e \circ \mathcal{H}_v, \underline{\mathcal{S}}^v)
\end{aligned}$$

where $\underline{\mathcal{S}}^v \stackrel{def}{=} \{(\underline{x}_i^v, y_i) | i = 1..m\}$

References

[1] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2003.

[2] G. Lugosi. Pattern classification and learning theory. In L. Gyorfi, editor, *Principles of Nonparametric Learning*, pages 1–56. Springer, 2002.

[3] D. A. McAllester. Pac-bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003.

[4] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics 1989*, pages 148–188. London Mathematical Society Lecture Notes, 1989.