PAC-Bayesian Theory for Transductive Learning

Luc Bégin[1], Pascal Germain[2], François Laviolette[2], Jean-Francis Roy[2]

[1] Campus d'Edmundston
Université de Moncton, Nouveau-Brunswick, Canada

[2] Département d'informatique et de génie logiciel
Université Laval, Québec, Canada

Abstract: We propose a PAC-Bayesian analysis of the transductive learning setting by proposing a family of new bounds on the generalization error.

Inductive Learning

Training set We draw m examples ***i.i.d.*** from a distribution D on $\mathcal{X}\times\{-1,+1\}$:

$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\} \sim D^m \,.$$

Task of an inductive learner Using S, learn a *classifier* $h : \mathcal{X} \mapsto \{-1,+1\}$ that has a low generalization risk on new examples drawn according to D:

$$R_D(h) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(x,y)\sim D} I\Big[h(x) \neq y\Big] \,,$$

where $I(a) = 1$ if predicate a is true and 0 otherwise. The number of errors $mR_S(h)$ follows a **binomial distribution** with parameters m and $R_D(h)$.

Transductive Learning (Vapnik, 1998)

Training set We draw m examples ***without replacement*** from a *full sample* Z of N examples. The remaining examples form a set U of $N-m$ examples.

Task of a transductive learner Using S and $U_{\mathcal{X}} = \{x_{m+1}, x_{m+2}, \ldots, x_N\}$, learn a classifier $h : Z_{\mathcal{X}} \mapsto \{-1,+1\}$ that has a low risk on the examples from the set Z:

$$R_Z(h) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{(x,y)\in Z} I\Big[h(x) \neq y\Big] \,.$$

The number of errors $mR_S(h)$ follows a **hypergeometric distribution** of m draws among a population of size N containing $NR_Z(h)$ successes.

PAC-Bayesian Basics

Given a *hypothesis space* $\mathcal{H}$ of classifiers and a *training set* S, we consider a *prior* distribution P on $\mathcal{H}$ and obtain a *posterior* distribution Q on $\mathcal{H}$ by learning from S.

PAC-Bayesian (inductive) theory bounds the *Gibbs risk* $R_D(G_Q) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{h\sim Q} R_D(h)$ from its empirical value $R_S(G_Q) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{h\sim Q} R_S(h)$ and $\mathrm{KL}(Q\|P) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{h\sim Q} \ln \frac{Q(h)}{P(h)}$.

Theorem 1 (inductive case) and Theorem 5 (transductive case) below are generic tools to derive various PAC-Bayesian bounds using any convex function $\mathcal{D} : [0,1]\times[0,1] \to \mathbb{R}$.

Inductive PAC-Bayesian Theory

Theorem 1 For any distribution D, for any set $\mathcal{H}$ of classifiers, for any prior distribution P on $\mathcal{H}$, for any $\delta \in (0,1]$, and for any convex function $\mathcal{D}$, with probability at least $1-\delta$ over the choice of $S \sim D^m$, we have

$$\forall Q \text{ on } \mathcal{H}: \quad \mathcal{D}\big(R_S(G_Q), R_D(G_Q)\big) \leq \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln \frac{\mathcal{I}_{\mathcal{D}}(m)}{\delta}\right],$$

where

$$\mathcal{I}_{\mathcal{D}}(m) \stackrel{\text{def}}{=} \sup_{r\in[0,1]} \left[\sum_{k=0}^{m} \tbinom{m}{k} r^k (1-r)^{m-k} e^{m\mathcal{D}(\frac{k}{m}, r)}\right].$$

To express a computable bound, one needs to calculate the value of $\mathcal{I}_{\mathcal{D}}(m)$. A common choice is $\mathcal{D} = \mathcal{D}_{\mathrm{KL}}$.

Kullback-Leibler divergence between two Bernoulli distributions

$$\mathcal{D}_{\mathrm{KL}}(q,p) \stackrel{\text{def}}{=} q \ln \frac{q}{p} + (1-q) \ln \frac{1-q}{1-p} = H(q,p) - H(q) \,,$$

where $H(q) \stackrel{\text{def}}{=} -q \ln q - (1-q)\ln(1-q)$ and $H(q,p) \stackrel{\text{def}}{=} -q \ln p - (1-q)\ln(1-p)$.

With these definitions, the r's cancel out in each term of the inner sum of $\mathcal{I}_{\mathcal{D}_{\mathrm{KL}}}(m)$:

$$\mathcal{I}_{\mathcal{D}_{\mathrm{KL}}}(m) = \sup_{r\in[0,1]} \left[\sum_{k=0}^{m} \tbinom{m}{k} e^{-mH(\frac{k}{m})}\right] = \sum_{k=0}^{m} \tbinom{m}{k} \left(\tfrac{k}{m}\right)^m \left(1-\tfrac{k}{m}\right)^{m-k} = \sum_{k=0}^{m} \alpha(k,m) \,.$$

Corollary 4 With probability at least $1-\delta$ over the choice of $S \sim D^m$, we have

$$\forall Q \text{ on } \mathcal{H}: \quad \text{a) } \mathcal{D}_{\mathrm{KL}}\big(R_S(G_Q), R_D(G_Q)\big) \leq \tfrac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln \tfrac{2\sqrt{m}}{\delta}\right],$$

$$\text{b) } R_D(G_Q) \leq R_S(G_Q) + \sqrt{\tfrac{1}{2m}\left[\mathrm{KL}(Q\|P) + \ln \tfrac{2\sqrt{m}}{\delta}\right]}.$$

Transductive PAC-Bayesian Theory

Theorem 5 For any set Z of N examples, for any set $\mathcal{H}$ of classifiers, for any prior distribution P on $\mathcal{H}$, for any $\delta \in (0,1]$, and for any convex function $\mathcal{D}$, with probability at least $1-\delta$ over the choice S of m examples among Z, we have

$$\forall Q \text{ on } \mathcal{H}: \quad \mathcal{D}\big(R_S(G_Q), R_Z(G_Q)\big) \leq \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln \frac{\mathcal{T}_{\mathcal{D}}(m,N)}{\delta}\right],$$

where

$$\mathcal{T}_{\mathcal{D}}(m,N) \stackrel{\text{def}}{=} \max_{K=0\ldots N} \left[\sum_{k\in\mathcal{K}_{mNK}} \frac{\tbinom{K}{k}\tbinom{N-K}{m-k}}{\tbinom{N}{m}} e^{m\mathcal{D}(\frac{k}{m}, \frac{K}{N})}\right],$$

and $\mathcal{K}_{mNK} \stackrel{\text{def}}{=} \{\max[0, K+m-N], \ldots, \min[m, K]\}$. One can compute the value of this bound for **any $\mathcal{D}$-function** if m and N are not unreasonably large.

A $\mathcal{D}$-function for the Transductive Case In the inductive setting, we express $\mathcal{I}_{\mathcal{D}_{\mathrm{KL}}}(m)$ by a sum of terms $\alpha(k,m)$, To recover the same phenomenon, we suggest

$$\mathcal{D}^{\star}_{\beta}(q,p) \stackrel{\text{def}}{=} \frac{1}{\beta}\Big[H(\beta) - pH(\beta\tfrac{q}{p}) - (1-p)H(\beta\tfrac{1-q}{1-p})\Big] = \mathcal{D}_{\mathrm{KL}}(q,p) + \tfrac{1-\beta}{\beta}\,\mathcal{D}_{\mathrm{KL}}\Big(\tfrac{p-\beta q}{1-\beta}, p\Big).$$

Theorem 6 Let m and N be any integers such that $20 \leq m \leq N-20$, we have

$$\mathcal{T}_{\mathcal{D}^{\star}_{m/N}}(m,N) = \max_{K=0\ldots N}\left[\sum_{k\in\mathcal{K}_{mNK}} \frac{\alpha(k,K)\,\alpha(m-k,N-K)}{\alpha(m,N)}\right] \leq 3\ln(m)\sqrt{m(1-\tfrac{m}{N})}\,.$$

Corollary 7 With probability at least $1-\delta$ over the choice S of m examples among Z (such that $20 \leq m \leq N-20$), we have

$$\forall Q \text{ on } \mathcal{H}: \quad \text{a) } \mathcal{D}^{\star}_{m/N}\big(R_S(G_Q), R_Z(G_Q)\big) \leq \tfrac{1}{m}\Big[\mathrm{KL}(Q\|P) + \ln \tfrac{3\ln(m)\sqrt{m(1-\frac{m}{N})}}{\delta}\Big],$$

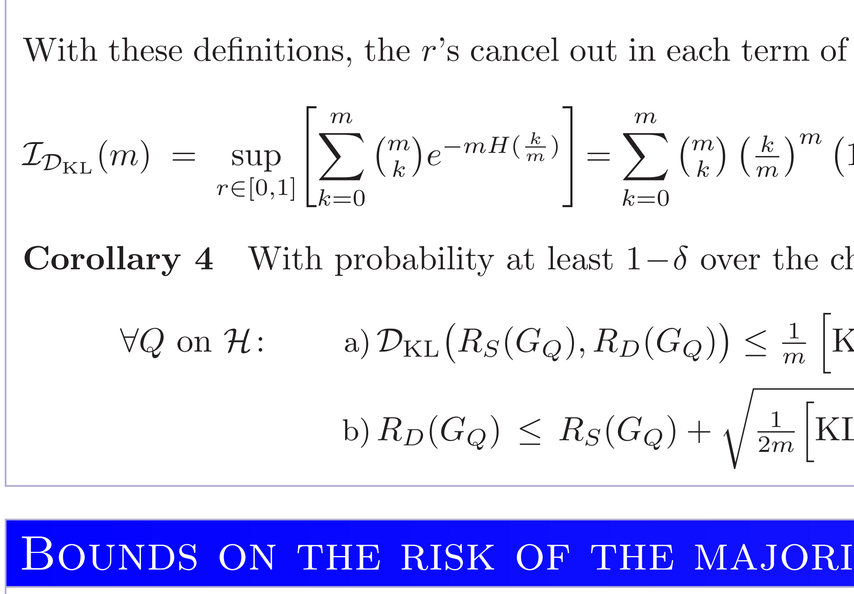
$$\text{b) } R_Z(G_Q) \leq R_S(G_Q) + \sqrt{\tfrac{1-\frac{m}{N}}{2m}\Big[\mathrm{KL}(Q\|P) + \ln \tfrac{3\ln(m)\sqrt{m(1-\frac{m}{N})}}{\delta}\Big]}\,.$$

Bounds on the risk of the majority vote classifier

To bound the risk of $B_Q(x) \stackrel{\text{def}}{=} \mathop{\mathrm{argmax}}_{c\in\{-1,+1\}} \Big[\mathop{\mathbf{E}}_{h\sim Q} I(h(x) = c)\Big]$, we use the *factor two* $R_Z(B_Q) \leq 2\,R_Z(G_Q)$ or the *$\mathcal{C}$-bound* $R_Z(B_Q) \leq 1 - \dfrac{(1-2\,R_Z(G_Q))^2}{1-2\,d^z_Q}$.

In the transductive setting, exact value of the *expected disagreement* $d^z_Q \stackrel{\text{def}}{=} \dfrac{1}{|Z|}\displaystyle\sum_{x\in Z_{\mathcal{X}}} \mathop{\mathbf{E}}_{h_1\sim Q}\mathop{\mathbf{E}}_{h_2\sim Q} I\Big[h_1(x) \neq h_2(x)\Big]$ is computed on the full sample!

Experiments with $\mathcal{D}$-functions



Experiments on real data

code: http://graal.ift.ulaval.ca/aistats2014/

Dataset information			Gibbs Classifier						Majority Vote Classifier			
			Observed Risk		Bounds of $R_Z(G_Q)$				Observed Risk		Bounds of $R_Z(B_Q)$	
Dataset	N	m/N	$R_S(G_Q)$	$R_Z(G_Q)$	Cor 7-(b)	Derbeko	Thm 5-$\mathcal{D}_{\mathrm{KL}}$	Thm 5-$\mathcal{D}^{\star}_{m/N}$	$R_S(B_Q)$	$R_Z(B_Q)$	2-$\mathcal{D}^{\star}_{m/N}$	$\mathcal{C}$-$\mathcal{D}^{\star}_{m/N}$
car	1728	0.1	0.193	0.194	**0.555**	0.793	**0.527**	0.546	0.105	0.159	1.092	-
car	1728	0.5	0.179	0.181	**0.418**	0.496	0.418	**0.415**	0.115	0.125	0.830	**0.819**
letter_AB	1555	0.1	0.146	0.149	**0.469**	0.718	**0.437**	0.457	0.000	0.017	**0.914**	0.961
letter_AB	1555	0.5	0.171	0.171	**0.402**	0.485	0.401	**0.399**	0.000	0.001	0.797	**0.626**
mushroom	8124	0.1	0.202	0.202	**0.486**	0.609	**0.471**	0.482	0.000	0.000	**0.964**	0.966
mushroom	8124	0.5	0.205	0.205	**0.439**	0.479	0.438	0.438	0.000	0.000	0.875	**0.546**
nursery	12959	0.1	0.169	0.168	**0.404**	0.504	**0.389**	0.399	0.009	0.016	0.798	**0.692**
nursery	12959	0.5	0.167	0.168	**0.357**	0.391	0.356	0.356	0.010	0.012	0.711	**0.379**
optdigits	3823	0.1	0.208	0.213	**0.533**	0.703	**0.513**	0.527	0.000	0.077	1.055	-
optdigits	3823	0.5	0.210	0.211	**0.460**	0.516	0.460	**0.458**	0.026	0.042	0.917	**0.793**
pageblock	5473	0.1	0.199	0.201	**0.495**	0.642	**0.476**	0.490	0.048	0.063	**0.979**	0.992
pageblock	5473	0.5	0.208	0.208	**0.448**	0.497	0.448	**0.447**	0.057	0.059	0.894	**0.697**
pendigits	7494	0.1	0.209	0.210	**0.499**	0.629	**0.481**	0.495	0.023	0.051	**0.989**	0.997
pendigits	7494	0.5	0.215	0.215	**0.457**	0.500	**0.455**	0.456	0.041	0.045	0.912	**0.706**
segment	2310	0.1	0.206	0.207	**0.558**	0.769	**0.533**	0.550	0.000	0.059	1.101	-
segment	2310	0.5	0.206	0.206	**0.462**	0.532	0.462	**0.460**	0.014	0.016	0.920	**0.834**
spambase	4601	0.1	0.222	0.227	**0.553**	0.708	**0.535**	0.548	0.115	0.161	1.096	-
spambase	4601	0.5	0.225	0.226	**0.488**	0.539	0.489	**0.486**	0.137	0.143	0.973	**0.961**