Chebyshev-Cantelli PAC-Bayes-Bennett Inequality for the Weighted Majority Vote

Yi-Shan Wu, Andrés R. Masegosa, Stephan S. Lorenzen, Christian Igel, Yevgeny Seldin

yswu@di.ku.dk, arma@cs.aau.dk, {lorenzen, igel, seldin}@di.ku.dk

KØBENHAVNS UNIVERSITET

AALBORG UNIVERSITET

Weighted Majority Vote

- A central technique to combine predictions of multiple classifiers
- Used in random forest, boosting, bagging, etc
- Wins most ML competitions

Prediction Rule

$$\mathrm{MV}_\rho(X) = \arg\max_{y \in \mathcal{Y}} \mathbb{E}_{h\sim\rho}[\mathbb{1}(h(X) = y)].$$

Key Power: Cancellation of errors effect

Previous Work

Standard Analysis

If a majority vote makes an error, at least a ρ-weighted half of the classifiers have made an error:

$$L(\mathrm{MV}_\rho) \le \mathbb{P}\left(\underbrace{\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]}_{Z} \ge 0.5\right)$$

First Order Oracle Bound

By Markov's inequality $\mathbb{P}(Z \ge \varepsilon) \le \mathbb{E}[Z]/\varepsilon$:

$$L(\mathrm{MV}_\rho) \le 2\mathbb{E}_D[\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]] = 2\mathbb{E}_\rho[L(h)]$$

Issues: Ignores correlation

C-bounds [1]

By Chebyshev-Cantelli inequality, if $\mathbb{E}[Z] < 0.5$,

$$\mathbb{P}(Z \ge 0.5) \le= \frac{\mathbb{E}[Z^2] - \mathbb{E}[Z]^2}{0.25 - \mathbb{E}[Z] + \mathbb{E}[Z^2]}$$

Issues: Difficult to estimate and optimize

Tandem Bound (TND)[2]

Tandem loss $\ell(h,h') := \mathbb{1}(h(X) \neq Y \wedge h'(X) \neq Y)$.

By second order Markov's inequality $\mathbb{P}(Z \ge \varepsilon) \le \mathbb{E}[Z^2]/\varepsilon^2$,

$$L(\mathrm{MV}_\rho) \le 4\mathbb{E}_{\rho^2}[L(h,h')].$$

Issues: The second order Markov's inequality is not as tight as the Chebyshev-Cantelli inequality

Our Contributions

Theorem 1 (Parametrized Chebyshev-Cantelli inequality). For any $\varepsilon > 0$ and all $\mu < \varepsilon$

$$\mathbb{P}(Z \ge \varepsilon) \le \frac{\mathbb{E}\left[(Z-\mu)^2\right]}{(\varepsilon-\mu)^2} = \frac{\mathbb{E}\left[Z^2\right] - 2\mu\mathbb{E}[Z] + \mu^2}{(\varepsilon-\mu)^2}.$$

- Taking $\mu^* = \mathbb{E}[Z] - \frac{\mathbb{V}[Z]}{\varepsilon - \mathbb{E}[Z]}$ recovers Chebyshev-Cantelli inequality
- Taking $\mu = 0$ recovers second order Markov's inequality

Main Advantage:

No distribution dependent quantities in the denominator

$\Rightarrow$ Easy to optimize & estimate

Chebyshev-Cantelli bound with tandem loss estimate

Oracle Bound: In multiclass classification, if $\mu < 0.5$,

$$L(\mathrm{MV}_\rho) \le \frac{\mathbb{E}_{\rho^2}[L(h,h')] - 2\mu\mathbb{E}_\rho[L(h)] + \mu^2}{(0.5-\mu)^2}.$$

From Oracle to Empirical: By PAC-Bayes-kl inequality [3],

$$\mathrm{kl}\left(\mathbb{E}_\rho[\hat{L}(h,S)] \middle\| \mathbb{E}_\rho[L(h)]\right) \le \frac{\mathrm{KL}(\rho\|\pi) + \ln(2\sqrt{n}/\delta)}{n}$$

Chebyshev-Cantelli bound with μ-tandem loss estimate

μ-tandem loss $\ell_\mu(h,h') := (\mathbb{1}(h(X) \neq Y) - \mu)(\mathbb{1}(h'(X) \neq Y) - \mu)$

Oracle Bound: In multiclass classification, if $\mu < 0.5$,

$$L(\mathrm{MV}_\rho) \le \frac{\mathbb{E}_{\rho^2}[L_\mu(h,h')]}{(0.5-\mu)^2}.$$

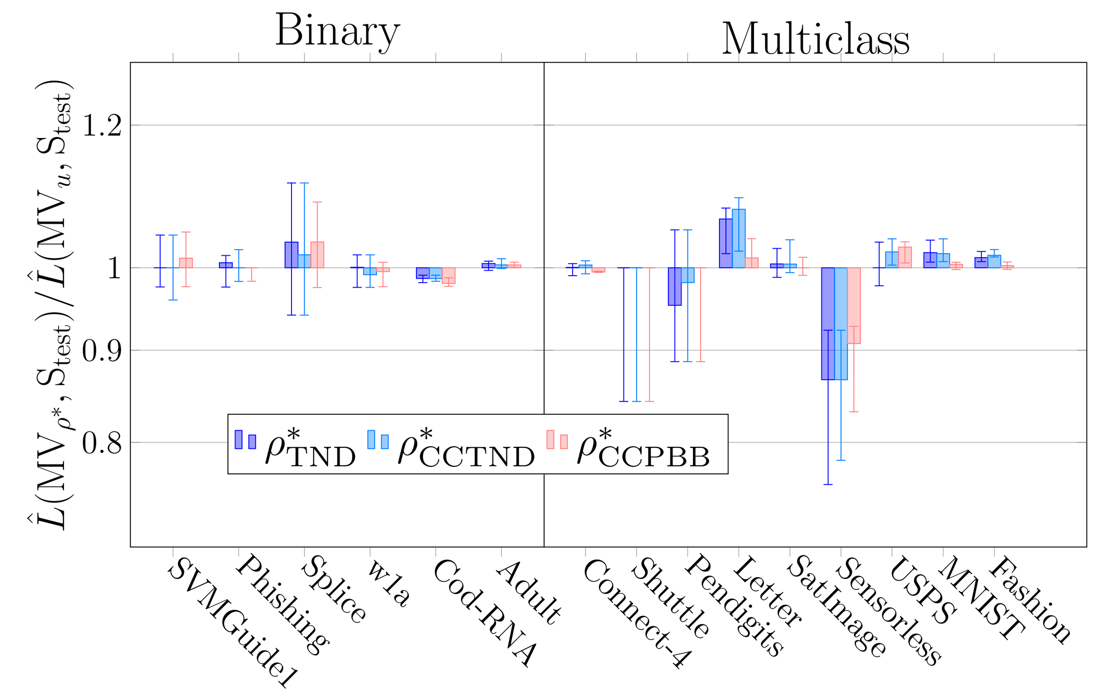
From Oracle to Empirical:

Theorem 2 (PAC-Bayes-Bennett). Assume $\tilde{\ell}(\cdot,\cdot) \le b$ and the corresponding variance is finite. Let $\phi(x) = e^x - x - 1$. Then for $\gamma > 0$,

$$\mathbb{E}_\rho[\tilde{L}(h)] \le \mathbb{E}_\rho[\hat{\tilde{L}}(h,S)] + \frac{\phi(\gamma b)}{\gamma b^2}\mathbb{E}_\rho[\tilde{\mathbb{V}}(h)] + \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{\gamma n}.$$

- Improves on the PAC-Bayes-Bernstein inequality by [4]
- The oracle variance $\mathbb{E}_\rho[\tilde{\mathbb{V}}(h)]$ can be bounded by Eq.(15) of [5]

Experiments



- CCTND: The empirical Chebyshev-Cantelli bound with tandem loss estimate
- CCPBB: The empirical Chebyshev-Cantelli bound with μ-tandem loss estimate

Summary

- Parametric form of Chebyshev-Cantelli inequality
  - No variance in the denominator and as tight as the original bound
  - Enables efficient minimization and empirical estimation
- New second order oracle bounds for weighted majority vote
  - Resulting empirical bounds are amenable to efficient minimization
- PAC-Bayes-Bennett inequality
  - Improves on the PAC-Bayes-Bernstein inequality by [4]
  - Can be used for bounding the μ-tandem loss

References

[1] A. Lacasse, F. Laviolette, M. Marchand, P. Germain, and N. Usunier, "PAC-Bayes bounds for the risk of the majority vote and the variance of the Gibbs classifier," in Advances in Neural Information Processing Systems (NeurIPS), 2007.

[2] A. R. Masegosa, S. S. Lorenzen, C. Igel, and Y. Seldin, "Second order PAC-Bayesian bounds for the weighted majority vote," in Advances in Neural Information Processing Systems (NeurIPS), 2020.

[3] M. Seeger, "PAC-Bayesian generalization error bounds for Gaussian process classification," Journal of Machine Learning Research, vol. 3, 2002.

[4] Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer, "PAC-Bayesian inequalities for martingales," IEEE Transactions on Information Theory, vol. 58, 2012.

[5] I. Tolstikhin and Y. Seldin, "PAC-Bayes-Empirical-Bernstein inequality," in Advances in Neural Information Processing Systems (NeurIPS), 2013.