Voting classifiers: generalization and optimization

Valentina Zantedeschi - ServiceNow Research

StatLearn 2023 - Montpellier

servicenow

What is a voting classifier?

given $\mathcal{Y}$ a set of K classes and a set of d voters $\{h_j : \mathcal{X} \to \mathcal{Y}\}_{j=1}^{d}$

Majority Vote classifier (MV)

$$f(x) = \underset{k \in \mathcal{Y}}{\operatorname{argmax}} \sum_{j=1}^{d} \mathbb{1}(h_j(x) = k)$$

where $\mathbb{1}(\cdot)$ is the indicator function.

What is a voting classifier?

given $\mathcal{Y}$ a set of K classes and a set of d voters $\{h_j : \mathcal{X} \to \mathcal{Y}\}_{j=1}^{d}$

Majority Vote classifier (MV)

$$f(x) = \underset{k \in \mathcal{Y}}{\operatorname{argmax}} \sum_{j=1}^{d} \mathbb{1}(h_j(x) = k)$$

where $\mathbb{1}(\cdot)$ is the indicator function.

To reach the right decision:

- how many voters?
- how good should they be?

Marquis de Condorcet

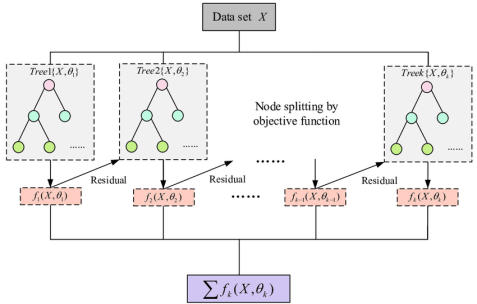
What is a voting classifier?

given $\mathcal{Y}$ a set of K classes and a set of d voters $\{h_j : \mathcal{X} \to \mathcal{Y}\}_{j=1}^{d}$

Majority Vote classifier (MV)

$$f(x) = \underset{k \in \mathcal{Y}}{\operatorname{argmax}} \sum_{j=1}^{d} \mathbb{1}(h_j(x) = k)$$

where $\mathbb{1}(\cdot)$ is the indicator function.



XGBOOST [GZW+20]

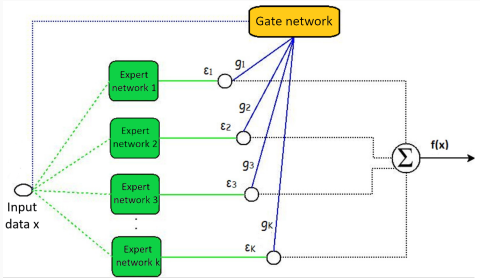
What is a voting classifier?

given $\mathcal{Y}$ a set of K classes and a set of d voters $\{h_j : \mathcal{X} \to \mathcal{Y}\}_{j=1}^{d}$

Majority Vote classifier (MV)

$$f(x) = \underset{k \in \mathcal{Y}}{\operatorname{argmax}} \sum_{j=1}^{d} \mathbb{1}(h_j(x) = k)$$

where $\mathbb{1}(\cdot)$ is the indicator function.



Mixture of Experts [PSCS19]

What is a voting classifier?

given $\mathcal{Y}$ a set of K classes and a set of d voters $\{h_j : \mathcal{X} \to \mathcal{Y}\}_{j=1}^{d}$

Majority Vote classifier (MV)

$$f(x) = \underset{k \in \mathcal{Y}}{\operatorname{argmax}} \sum_{j=1}^{d} \mathbb{1}(h_j(x) = k)$$

where $\mathbb{1}(\cdot)$ is the indicator function.

What is a voting classifier?

given $\mathcal{Y}$ a set of K classes and a set of d voters $\{h_j : \mathcal{X} \to \mathcal{Y}\}_{j=1}^{d}$

Majority Vote classifier (MV)

$$f(x) = \underset{k \in \mathcal{Y}}{\operatorname{argmax}} \sum_{j=1}^{d} \mathbb{1}(h_j(x) = k)$$

where $\mathbb{1}(\cdot)$ is the indicator function.

Weighted MV classifier

$$f_\theta(x) = \underset{k \in \mathcal{Y}}{\operatorname{argmax}} \sum_{j=1}^{d} \theta_j \mathbb{1}(h_j(x) = k)$$

where $\theta \in \Delta_d$ (positive and sum to 1).

What is a voting classifier?

given $\mathcal{Y}$ a set of K classes and a set of d voters $\{h_j : \mathcal{X} \to \mathcal{Y}\}_{j=1}^{d}$

Majority Vote classifier (MV)

$$f(x) = \underset{k \in \mathcal{Y}}{\operatorname{argmax}} \sum_{j=1}^{d} \mathbb{1}(h_j(x) = k)$$

where $\mathbb{1}(\cdot)$ is the indicator function.

Weighted MV classifier

$$f_\theta(x) = \underset{k \in \mathcal{Y}}{\operatorname{argmax}} \sum_{j=1}^{d} \theta_j \mathbb{1}(h_j(x) = k)$$

where $\theta \in \Delta_d$ (positive and sum to 1).

How to learn θ?

Which guarantees on accuracy?

Outline

PAC-Bayes meets Margin Theory

1. Define Stochastic MV class
2. Bound their error via PAC-Bayes
3. Use it to bound (*deterministic*) MV via Margin Loss

To obtain

1. the tightest generalization bounds for MV
2. that can be optimized by gradient descent

Majority Votes - Errors

Define $W_\theta(X, Y)$, weighted number of incorrect voters on (X, Y):

$$W_\theta(X, Y) = \sum_{j=1}^{d} \theta_j \mathbb{1}(h_j(X) \neq Y).$$

Majority Votes - Errors

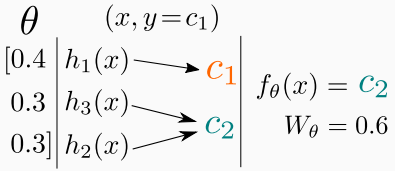
Define $W_\theta(X, Y)$, weighted number of incorrect voters on (X, Y):

$$W_\theta(X, Y) = \sum_{j=1}^{d} \theta_j \mathbb{1}(h_j(X) \neq Y).$$

In binary classification

$$\begin{aligned} R(\theta) &\overset{def}{=} \mathbb{E}_{\mathcal{P}}\, \mathbb{1}(W_\theta(X, Y) \geq 0.5) \\ &= \mathbb{P}(W_\theta \geq 0.5) \end{aligned}$$

$$\hat{R}(\theta) \overset{def}{=} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(W_\theta(x_i, y_i) \geq 0.5)$$



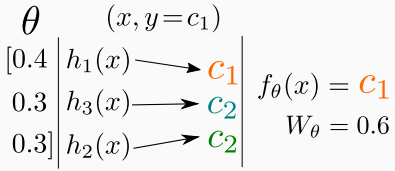
Majority Votes - Errors

Define $W_\theta(X, Y)$, weighted number of incorrect voters on (X, Y):

$$W_\theta(X, Y) = \sum_{j=1}^{d} \theta_j \mathbb{1}(h_j(X) \neq Y).$$

In multiclass classification:

$$R(\theta) \leq \mathbb{P}(W_\theta \geq 0.5).$$



PAC-Bayes Bounds

PAC-Bayes Framework

- Generalization bounds for several losses
- Randomised predictions using $h \sim Q$ from posterior Q
- Fix a data-independent prior P

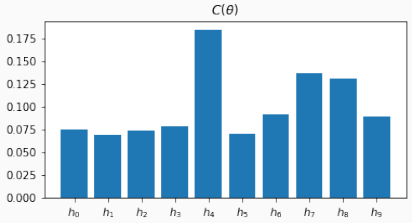
PAC generalization bound[See02, Mau04]
With probability $\geq 1-\delta$ over the sample and for all Q,

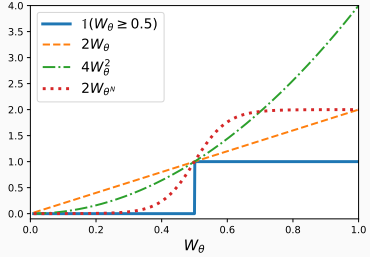
$$\mathbb{E}_{h\sim Q}\left[R(h) - \hat{R}(h)\right] \leq O\left(\sqrt{\frac{\mathrm{KL}(Q,P) + \log\frac{1}{\delta}}{\#\text{samples.}}}\right).$$

- bounds as objective functions to optimize the posterior

PAC-Bayes bounds for MV

- Using randomized classifier: $h \sim \mathcal{C}(\theta)$
- Upper oracle bounds to link error of randomized classifier to error of Majority Vote





PAC-Bayes bounds for MV

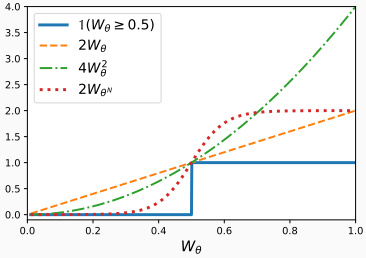
- Using randomized classifier: $h \sim \mathcal{C}(\theta)$
- Upper oracle bounds to link error of randomized classifier to error of Majority Vote

First Order bound [LS02]

Draw one base classifier

$$\mathbb{E}_{h\sim\mathcal{C}(\theta)}\mathbb{1}(h(X)\neq Y) \;=\; W_\theta$$

$$R(\theta) \leq 2\,\mathbb{E}_{\mathcal{P}}\; W_\theta$$



PAC-Bayes bounds for MV

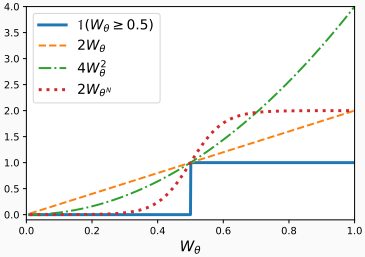
- Using randomized classifier: $h \sim \mathcal{C}(\theta)$
- Upper oracle bounds to link error of randomized classifier to error of Majority Vote

Second Order bound [MLIS20]

Draw two base classifiers

$$\mathbb{E}_{h\sim\mathcal{C}(\theta),h'\sim\mathcal{C}(\theta)}\mathbb{1}(h(X)\neq Y \wedge h'(X)\neq Y) \;=\; W_\theta^2$$

$$R(\theta) \leq 4\,\mathbb{E}_{\mathcal{P}}\; W_\theta^2$$



PAC-Bayes bounds for MV

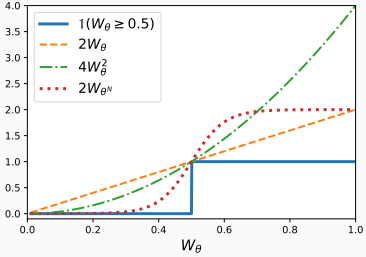
- Using randomized classifier: $h \sim \mathcal{C}(\theta)$
- Upper oracle bounds to link error of randomized classifier to error of Majority Vote

Binomial-law bound [SH09, LLMT10]

Draw N base classifiers, probability that at least $\frac{N}{2}$ make an error

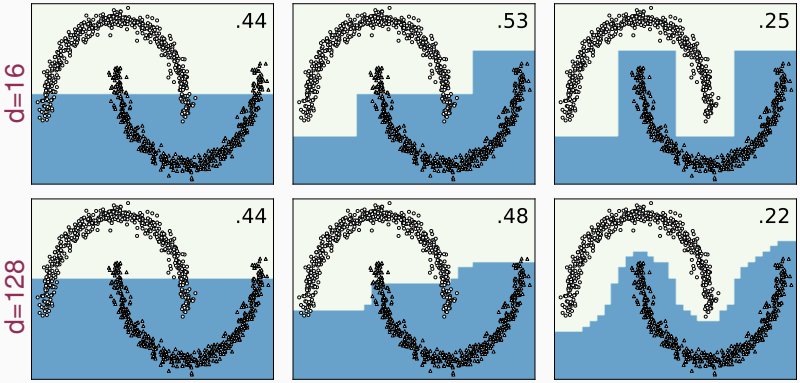
$$W_{\theta^N}(X, Y) \stackrel{def}{=} \sum_{k=\frac{N}{2}}^{N} \binom{N}{k} W_\theta^k (1 - W_\theta)^{(N-k)}$$

$$R(\theta) \leq 2\, \mathbb{E}_{\mathcal{P}} W_{\theta^N}$$



PAC-Bayes Self-bounding Algorithms on Two-Moons

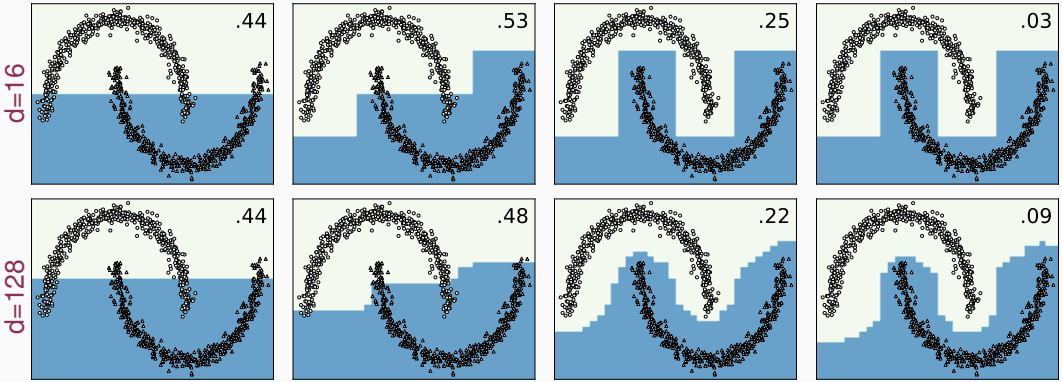
d decision stumps (axis-aligned, evenly distributed)



(a) First Order **(b)** Second Order **(c)** Binomial

PAC-Bayes Self-bounding Algorithms on Two-Moons

d decision stumps (axis-aligned, evenly distributed)



(a) First Order **(b)** Second Order **(c)** Binomial **(d)** ours

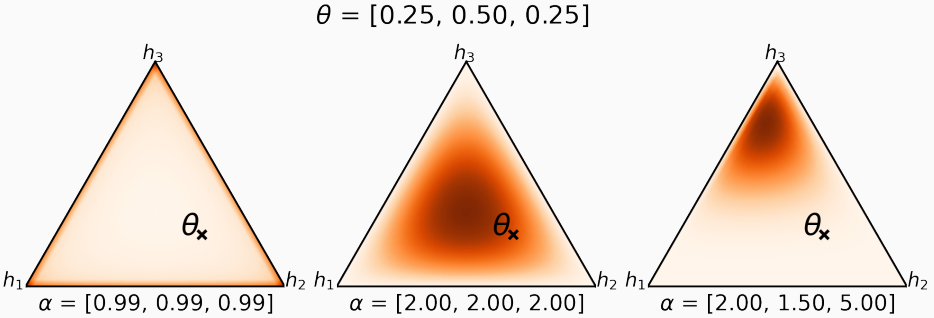
Stochastic Majority Votes

Majority Vote with hyper-priors

$\alpha = [\alpha_j \in \mathbb{R}^+]_{j=1}^M$, $B(\alpha)$ a normalization factor

Dirichlet:

$$\theta \sim \mathcal{D}(\alpha_1, \ldots, \alpha_M), \qquad \rho(\theta) = \frac{1}{B(\alpha)} \prod_{j=1}^{M} (\theta_j)^{\alpha_j - 1},$$



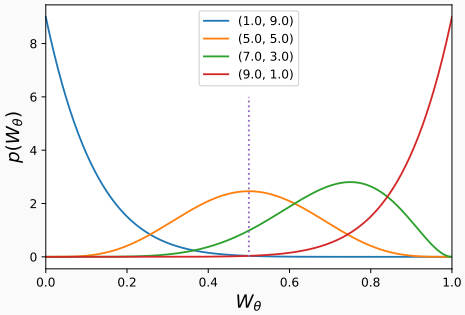
Error: closed-form solution with Dirichlet

For a given $(x, y) \sim \mathcal{P}$, let

- $w = \{j | h_j(x) \neq y\}$, the set of indices of the base classifiers that misclassify (x, y),
- $c = \{j | h_j(x) = y\}$, the set of indices of the base classifiers that correctly classify (x, y).

W_θ follows a Beta distribution (bi-variate Dirichlet distribution)

$$W_\theta \sim \mathcal{B}\left(\sum_{j \in w} \alpha_j, \sum_{j \in c} \alpha_j\right)$$



Error: closed-form solution with Dirichlet

For a given $(x, y) \sim \mathcal{P}$, let

- $w = \{j | h_j(x) \neq y\}$, the set of indices of the base classifiers that misclassify (x, y),
- $c = \{j | h_j(x) = y\}$, the set of indices of the base classifiers that correctly classify (x, y).

Lemma
The expected error (or 01*-loss) for* (x, y) *of the stochastic majority vote under* $\theta \sim \mathcal{D}(\alpha_1, \ldots, \alpha_M)$ *is equal to*

$$\int_{\Theta} \mathbb{1}(W_\theta(x, y) \geq 0.5)\rho(d\theta) = I_{0.5}\left(\sum_{j \in c} \alpha_j, \sum_{j \in w} \alpha_j\right),$$

with $I_{0.5}(\cdot)$ *the regularized incomplete beta function evaluated at* 0.5*.*

... and can be optimized by gradient descent

PAC-Bayes Bound for Stochastic MV

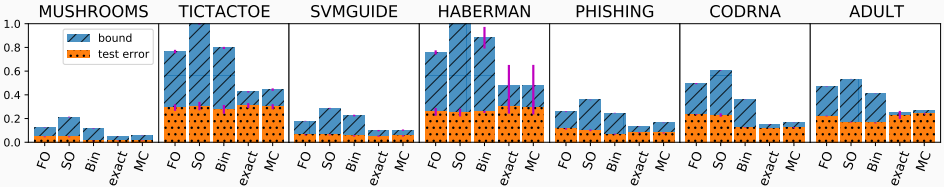
Theorem (Generalization bound)
with probability at least $1-\delta$ *over samples* $S = \{(x_i, y_i) \sim \mathcal{P}\}_{i=1}^{m}$ *of size* m *we have simultaneously for any posterior* ρ *over* Θ*:*

$$\int_{\Theta} |R(\theta) - \hat{R}(\theta)| \rho(d\theta) \leq O\left(\frac{\mathrm{KL}(\rho, \pi) + \log\left(\frac{2\sqrt{m}}{\delta}\right)}{m} \right)$$

... also tight bound with data-dependent prior $\rightarrow$ learning the base classifiers

Comparison with SOTA

Binary classification: $M = 10$ decision stumps per feature and per class



Multiclass classification: $M = 200$ decision trees

From stochastic to deterministic MV guarantees

How can we derandomize the bound?

PAC-Bayes Margin Bounds

Margin and Margin Loss

difference between correct and most heavily weighted other class

Margin

$$\sum_{j|h_i(x)=y} \theta_j - \max_{k \neq y} \sum_{j|h_i(x)=k} \theta_j$$

- positive margin
  $\rightarrow$ correct classification
- large margin
  $\rightarrow$ confident correct classification

Margin and Margin Loss

difference between correct and most heavily weighted other class

Margin

$$\sum_{j|h_i(x)=y} \theta_j - \max_{k \neq y} \sum_{j|h_i(x)=k} \theta_j$$

- positive margin
  $\rightarrow$ correct classification
- large margin
  $\rightarrow$ confident correct classification

Margin Loss
for margin parameter $\gamma \geq 0$

$$R_\gamma(\theta) \stackrel{def}{=} \mathbb{P}(W_\theta \geq 0.5 + \gamma)$$

$$\hat{R}_\gamma(\theta) \stackrel{def}{=} \frac{1}{m} \sum_{i=1}^{n} \mathbb{1}(W_\theta(x_i, y_i) \geq 0.5 + \gamma)$$

Derandomization via Margin Loss

Let $R(\theta)$ be the error of the MV we want to bound

Consider stochastic MV with Dirichlet posterior $\rho = \mathcal{D}(K\theta)$:

- θ is the mean of ρ
- K is its concentration parameter

Derandomization via Margin Loss

Let $R(\theta)$ be the error of the MV we want to bound

Consider stochastic MV with Dirichlet posterior $\rho = \mathcal{D}(K\theta)$:

- θ is the mean of ρ
- K is its concentration parameter

Lemma
For any $\gamma \geq 0$ *and* $K \geq 1$

$$|R(\theta) - \mathbb{E}_{\xi \sim \mathcal{D}(K\theta)} R_\gamma(\xi)| \leq \exp(-4(K+1)\gamma^2)$$

thanks to Dirichlet's aggregation property.

Derandomization via Margin Loss

Let $R(\theta)$ be the error of the MV we want to bound

Consider stochastic MV with Dirichlet posterior $\rho = \mathcal{D}(K\theta)$:

- θ is the mean of ρ
- K is its concentration parameter

Lemma
For any $\gamma \geq 0$ *and* $K \geq 1$

$$|R(\theta) - \mathbb{E}_{\xi \sim \mathcal{D}(K\theta)} R_\gamma(\xi)| \leq \exp(-4(K+1)\gamma^2)$$

thanks to Dirichlet's aggregation property.

...and dimension-free

PAC-Bayes Margin Bound for Majority Votes

Theorem (Derandomized bound)
with probability at least $1-\delta$ *over samples* $S = \{(x_i, y_i) \sim \mathcal{P}\}_{i=1}^{m}$ *of size* m *we have simultaneously for any* θ*:*

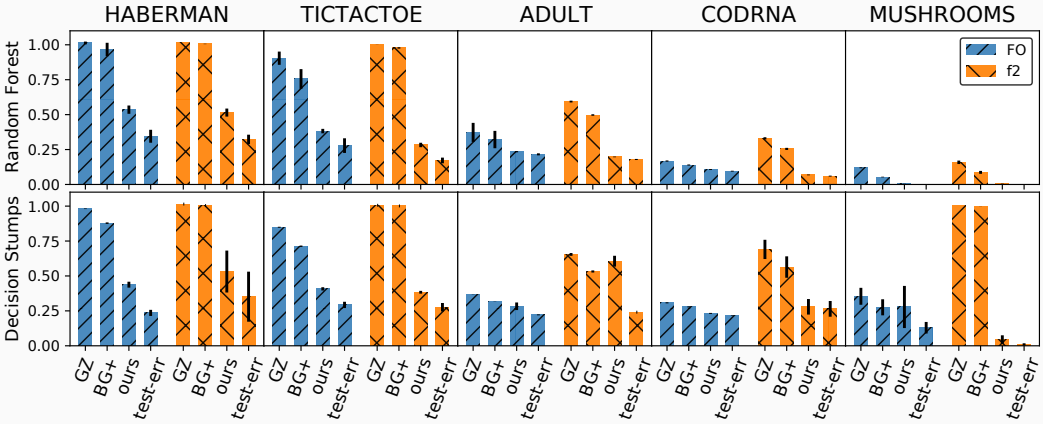
$$|R(\theta) - \mathbb{E}_{\xi \sim \mathcal{D}(K\theta)} \hat{R}_\gamma(\xi)| \leq O\left(\frac{\mathrm{KL}(\mathcal{D}(K\theta), \pi) + \log\left(\frac{2\sqrt{m}}{\delta}\right)}{m} + \exp(-4(K+1)\gamma^2) \right)$$

To tighten bound:

1. increase concentration K (but higher KL)
2. increase margin γ (but larger loss)

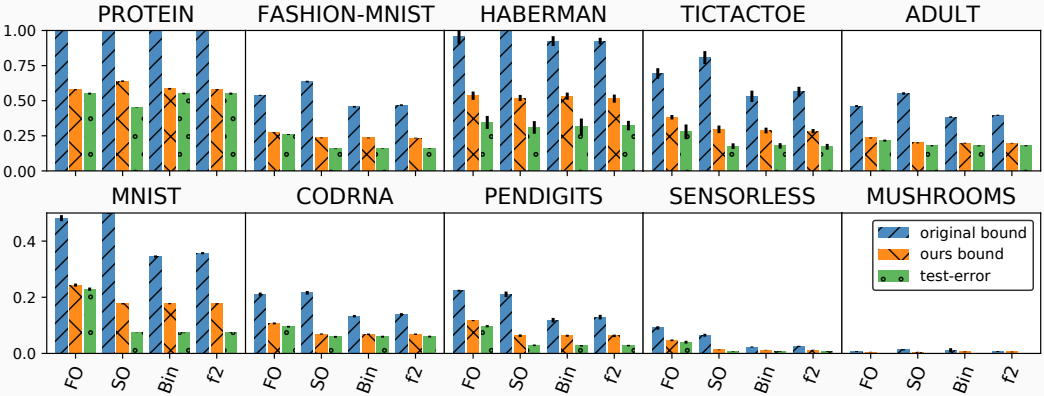
Results with Random Forests

Comparison with other margin bounds: **GZ** [GZ13], **BG+** [BG22]



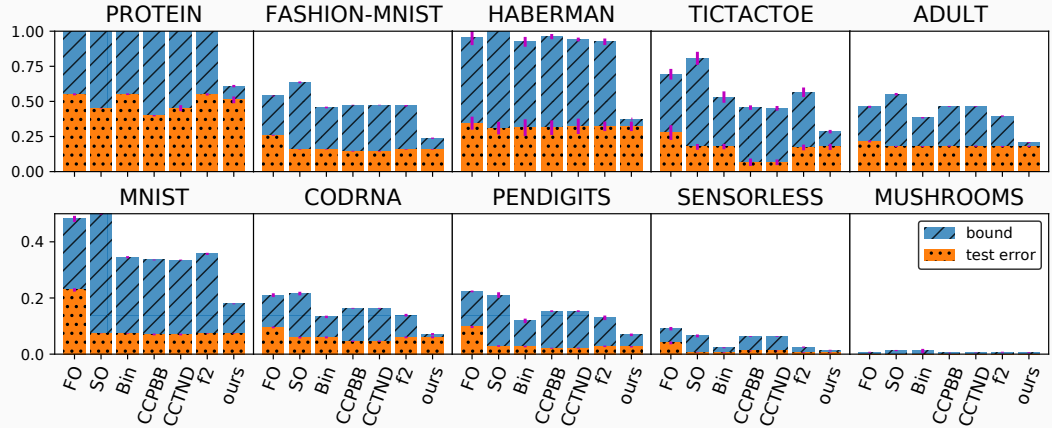
Results with Random Forests

Comparison with other PAC-Bayes bounds



Results with Random Forests

Comparison with other PAC-Bayes optimizations



Takeaways and Future Work

Randomize with Dirichlet and derandomize via margin

- + tightest bounds for MV (in most cases)
- + independent of number of voters
- + tractable training objective

Takeaways and Future Work

Randomize with Dirichlet and derandomize via margin

+ tightest bounds for MV (in most cases)
+ independent of number of voters
+ tractable training objective

Not solved yet:

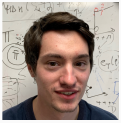
1. exact bounds for multiclass classification
2. data-conditioned weighting
3. learn jointly posterior and voters

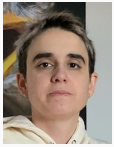
Warm thanks to my amazing collaborators

NeurIPS 2021 **Learning Stochastic Majority Votes by Minimizing a PAC-Bayes Generalization Bound**

NeurIPS 2022 **On Margins and Generalisation for Voting Classifiers**


Felix Biggs


Paul Viallard


Emilie Morvant


Remi Emonet


Amaury Habrard


Pascal Germain


Benjamin Guedj

References i

[BG22] Felix Biggs and Benjamin Guedj.
On margins and derandomisation in pac-bayes.
In *International Conference on Artificial Intelligence and Statistics*, pages 3709–3731. PMLR, 2022.

[BK15] Daniel Berend and Aryeh Kontorovich.
A finite sample analysis of the naive bayes classifier.
J. Mach. Learn. Res., 16(1):1519–1545, 2015.

[GZ13] Wei Gao and Zhi-Hua Zhou.
On the doubt about margin explanation of boosting.
Artificial Intelligence, 203:1–18, 2013.

[GZW+20] Rui Guo, Zhiqian Zhao, Tao Wang, Guangheng Liu, Jingyi Zhao, and Dianrong Gao.
Degradation state recognition of piston pump based on iceemdan and xgboost.
Applied Sciences, 10:6593, 09 2020.

[LLMT10] Alexandre Lacasse, François Laviolette, Mario Marchand, and Francis Turgeon-Boutin.
Learning with randomized majority votes.
In *ECML/PKDD (2)*. Springer, 2010.

References ii

[LS02] John Langford and John Shawe-Taylor.
PAC-Bayes & margins.
In *NIPS*. MIT Press, 2002.

[Mau04] Andreas Maurer.
A note on the PAC Bayesian theorem.
CoRR, cs.LG/0411099, 2004.

[MLIS20] Andrés R. Masegosa, Stephan Sloth Lorenzen, Christian Igel, and Yevgeny Seldin.
Second order PAC-Bayesian bounds for the weighted majority vote.
In *NeurIPS*, 2020.

[PSCS19] Billy Peralta, Ariel Saavedra, Luis Caro, and Alvaro Soto.
Mixture of experts with entropic regularization for data classification.
Entropy, 21(2):190, 2019.

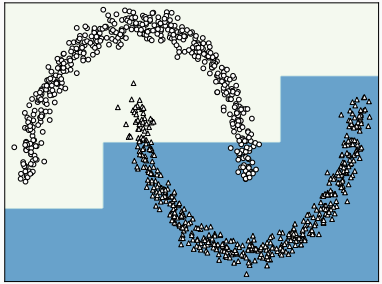
[See02] Matthias W. Seeger.
PAC-Bayesian generalisation error bounds for gaussian process classification.
JMLR, 2002.

References iii

[SH09] John Shawe-Taylor and David R. Hardoon.
Pac-bayes analysis of maximum entropy classification.
In *AISTATS*, 2009.

Naive Bayes classifier [BK15]

M=16



M=128