State of the Art
Ensemble Learning

References

[1] Q. Bai, H. Lam, and S. Sclaroff, "A bayesian framework for online classifier ensemble," in *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014, pp. 1584–1592.

[2] A. Lacoste, M. Marchand, F. Laviolette, and H. Larochelle, "Agnostic bayesian learning of ensembles," in *Proceedings of The 31st International Conference on Machine Learning (ICML)*, 2014, pp. 611–619.

[3] C. Cortes, M. Mohri, and U. Syed, "Deep boosting," in *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014, pp. 1179–1187.

[4] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter, "Efficient and robust automated machine learning," in *Advances in Neural Information Processing Systems (NIPS)*, 2015, pp. 2944–2952.

[5] C. Cortes, V. Kuznetsov, and M. Mohri, "Ensemble methods for structured prediction," in *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014, pp. 1134–1142.

[6] C. Cortes, V. Kuznetsov, M. Mohri, and M. Warmuth, "On-line learning algorithms for path experts with non-additive losses," in *Journal of Machine Learning Research Workshop and Conference Proceedings (JMLR)*, 2015, pp. 424–447.

[7] A. Lacoste, H. Larochelle, F. Laviolette, and M. Marchand, "Sequential model-based ensemble optimization," *arXiv preprint arXiv:1402.0796*, 2014.

[8] P. Shivaswamy and T. Jebara, "Variance penalizing adaboost," in *Advances in Neural Information Processing Systems*, 2011, pp. 1908–1916.

Deep Boosting (JMLR 2014)

Idea: Classifier set $H = \{H_1, H_2, \ldots, H_p\}$ where each set of classifiers H_i has increasing complexity.

Higher efficiency attainable if higher weight is placed on low complexity H_i and only use high complexity H_i when necessary. How can the mixture weights and learning guarantees be computed?

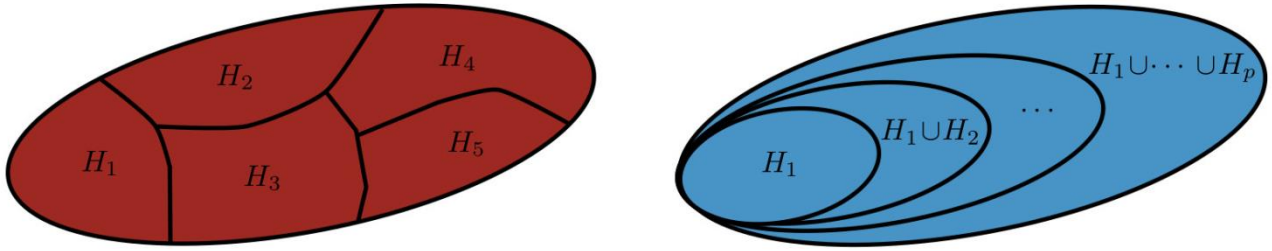


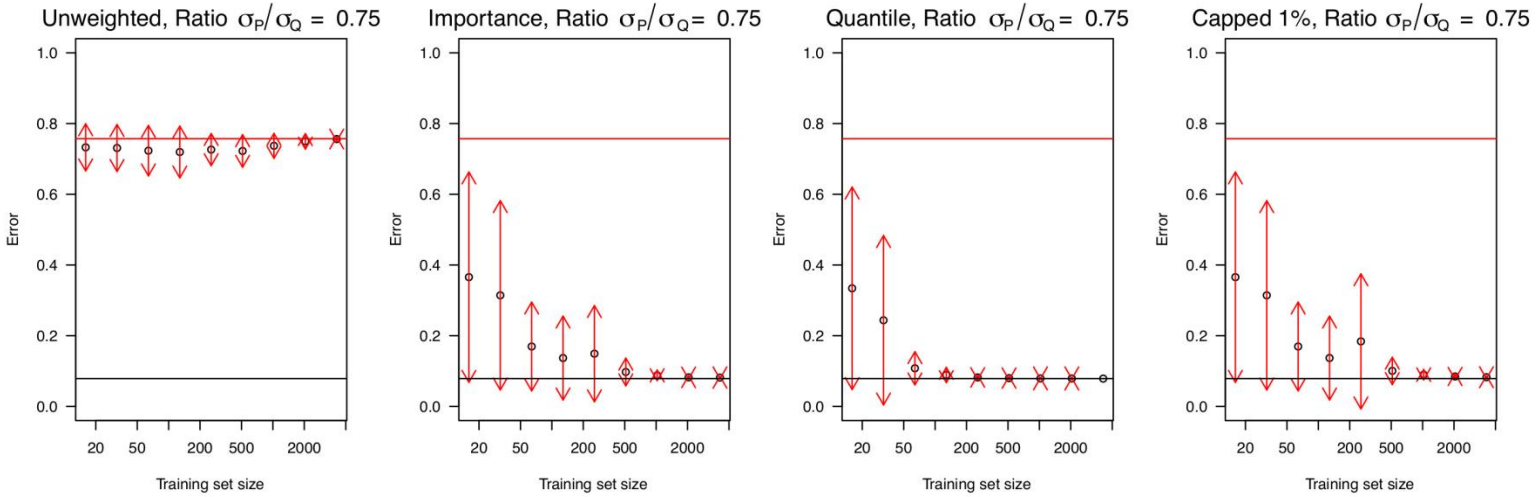
Figure 1. Base classifier set H decomposed in terms of sub-families $H_1, \ldots, H_p$ or their unions.

Solution:

1. Rademacher complexity for each H_i to construct a data-dependent learning bound for convex ensembles–based on margin-based learning (Bartlett and Mendelson, 2002 JMLR). Paper
2. Rademacher complexity is a generalization of the VC-dimension for finite H allowing for the complexity measure of H taking arbitrary real values.
3. Theorem: Paper

Variance Penalizing AdaBoost (NIPS 2011) Paper

Idea: AdaBoost essentially performs empirical risk minimization. An alternative to empirical risk minimization is variance penalization (balance the mean and variance).



Solution: Use empirical Bernstein bounds (Maurer and Pontil, 2009) and iteratively minimizes a cost function that balances the sample mean and the sample variance of the exponential loss.

1. Empirical Bernstein bounds is a concentration inequality like Hoeffding's inequality (mean and empirical mean) but also includes the variance.
2. Very useful to utilize the empirical Bernstein bounds as it can be used to provide learning bounds for importance weighting ensembles (Cortes et al., 2010 NIPS). Paper

Agnostic Bayesian Learning of Ensembles (JMLR 2014) Paper

Idea: Produce ensembles of classifiers based on holdout estimation of their generalization performances estimated using Bayesian inference.

1. Bayesian inference allows uncertainty about the performance and is used to weighted the predictors accordingly.
2. Finding the best (as opposed to the true) predictor among a class is known as agnostic PAC-learning. Additionally, the non-reliance on the assumption that the true underlying data generating function belongs to our model class is also at the center of agnostic PAC-learning.
3. $h \in H$ is a finite set of predictors obtained from one or many learning algorithms, with various hyperparameters.

Solution:

1. Assume risk of misclassification is a random variable r.
2. Assume prior on r, then we observe the losses L to compute the posterior $p(r|L)$.
3. Given $p(r|L)$, can compute the probability of $h \in H$ being the predictor with the lowest risk:

$$P(\forall g \in H : r_h \leq r_g | L) = E_{r \sim p(\cdot|L)}[\mathbf{1}\{r_h \leq r_g, \forall g \neq h\}]$$

Ensemble Methods for Structured Prediction (JMLR 2014/2015) Paper

Idea: Finite set of substructures $l > 1$, and a user defined loss function. Find the optimal substructure for prediction.

1. Used in speech analysis. Some predictors better then others at detecting perceptually distinct units of sound. Patch together results of predictors to correctly identify the word.
2. Boosting approach utilized.