State of the Art
Ensemble Learning
References


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?

![Base classifier set $H$ decomposed in terms of sub-families $H_1, \ldots, H_p$ or their unions.](image)

**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)  

**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:

\[
P(\forall g \in H : r_h \leq r_g | L) = E_{r \sim p(.)|L} [1\{r_h \leq r_g, \forall g \neq h\}]
\]
Ensemble Methods for Structured Prediction (JMLR 2014/2015)  

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.