Reading Notes for Week 5

Random Forest

Unlike logistic regression and LS-SVM where the model falls victim to overfitting when model complexity is increasing, the random forest model converges to the optimal model by adding more tree models. A very interesting aspect of random forest is that its generalization error(irreducible part of the error) is upper-bounded.

The paper basically go through how we should train a random forest model, how to evaluate it later on. To train a random forest model, we have two frameworks. The first is by bagging, which is bootstrap and aggregation. The idea is that we train different classifiers on the bootstrapped dataset, each spitting out a weak learner that has low variance and high bias. Since bagging is generally a variance reduction ensemble method, a very important observation is that due to the correlations between different trees trained on bootstrapped datasets, the variance may not be reduced so neatly since we can derive the variance of our predictors as follows:

$$\begin{aligned}\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{n}\sum_i X_i\right) \\ &= \frac{1}{n^2}\sum_{i,j}\mathrm{Cov}(X_i, X_j) \\ &= \frac{n\sigma^2}{n^2} + \frac{n(n-1)\rho\sigma^2}{n^2} \\ &= \rho\sigma^2 + \frac{1-\rho}{n}\sigma^2\end{aligned}$$

where n is the number of classifiers we trained. Notice that even when we increase the number of classifiers, the variance of the ensemble model is still irreducible to 0 due to the first term. So we have random feature selection techniques while training different tree instances. But the paper uses empirical results to show that such techniques improve classification tasks but not regression tasks.

Gradient Boosting

Unlike bagging, boosting is an ensemble method focusing on reducing the bias of the model.Adaboost is an example of this collection of methods. But adaptive boosting assumes we are using an exponential loss function while gradient boosting is a more general technique that allows an arbitrary differentiable loss function. Basically we can use gradient boosting(GB in short) to solve regression problems and classification problems. For regression, we typically choose MAE or huber loss instead of squared loss so that our training process will not be robust to the outliers. Moreover, in the squared loss case, the gradient of the loss function is the same as the residuals. But in huber loss, if we fit an additive model on residual, we will be less robust to outliers than if we do so on negative gradient.