Lecture 20: Boosting

CS3780/5780

Bagging Recap

- Given data set D of size n: Sample $D_1, D_2, \ldots, D_m$ by drawing n points from D with replacement
- Train models on each of $D_1, D_2, \ldots, D_m$, call them $h_1, h_2, \ldots, h_m$ (illustration)
- Bagged hypothesis $h_{Bagged} = \frac{1}{m} \sum_{i=1}^{m} h_i$
- Since we average m models we reduce variance (how much by depends on how uncorrelated the learn hypotheses are)
- As each prediction is average of many predictors we can obtain variance for each prediction and use it for uncertainty estimate

Ensemble Methods

- Train multiple models/hypotheses and combine them
- Ensemble methods are plug-and-play (can be used with any base algorithm for learning)
- Bagging: takes high variance-low bias methods and ensembles them to reduce variance : Fixes overfitting
- Boosting: Takes high bias-low variance methods and ensembles them to obtain low bias model : Fixes undercutting

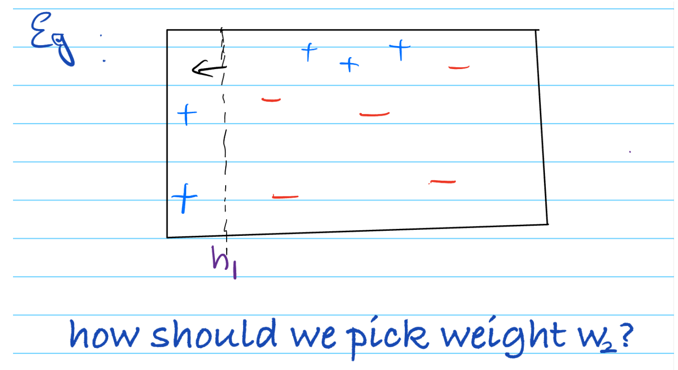
Random Forest Recap

- Use Bagging with max-depth decision trees
- However, while deciding each split (for each D_i), subsample k< d features and only use them for the split
- Random forest is one of the best out of the box classifier that works great!
- Only k and m are the parameters: $k = \sqrt{d}$ is often great choice, larger m helps.

Boosting

- Binary classification $(\mathscr{Y} = \{+1, -1\})$
- We are given dataset $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
- Assume access to a high bias (simple) classification algorithm that can barely do better than chance: weak learner
- Boosting answers the question: How can we ensemble weak learners to get an algorithm with 0 training error
  - In bagging we sampled with replacement uniformly from D
  - In Boosting we will sequentially sample points from D from carefully chosen weights over points in D

Adaboost Example



Boosting Algorithm

Skeleton Algorithm:

1. $\forall i \in [n],$ set $w_1[i] = \frac{1}{n}$ (initialize uniformly)
2. $H_0 = 0$ (initialize ensembles hypothesis to 0)
3. For t = 1 to T
   1. Create dataset D_t by drawing points from D according to weight w_t
   2. Use weak learning algorithm on D_t to obtain classifier h_t
   3. Add h_t to ensemble H_t
   4. Update weight w_t over points in D.
4. End
5. Return Ensembled classifier

Adaboost Algorithm

1. $\forall i \in [n],$ set $w_1[i] = \frac{1}{n}$ (initialize uniformly)
2. $H_0 = 0$ (initialize ensembles hypothesis to 0)
3. For t = 1 to T
   1. D_t ~ sample points in D using weights w_t
   2. h_t = Weak_Learning_Algo(D_t)
   3. $\epsilon_t = \sum_{i=1}^{n} w_t[i]\mathbf{1}\{h_t(x_i) \neq y_i\}$, $\alpha_t = \frac{1}{2}\log\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$
   4. $H_t = H_{t-1} + \alpha_t h_t$
   5. $\forall i \in [n], w_{t+1}[i] \propto w_t[i] \times \exp\left(-\alpha_t y_i h_t(x_i)\right)$
4. End
5. Return $h_{\text{Boost}}(x) = \text{sign}(H_T(x))$

Weak Learning Hypothesis

- WLH: For any weights over points in D, the weak learning algorithm can produce hypothesis whose weighted classification error is less than $\frac{1}{2} - \gamma$ for some $\gamma > 0$

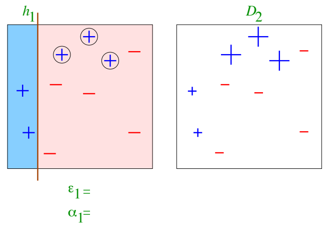
Generalization of Boosting

- Each weak learner is high bias but low variance classifier
- We only combine order log(n) of these weak learners
- So Boosted classifier will not have too high of a variance

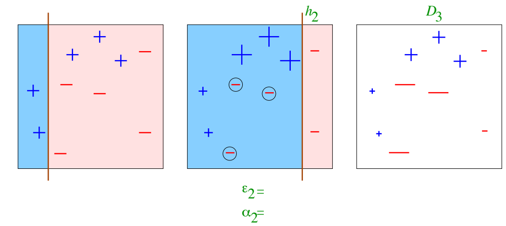
Boosting Theorem

- If weak learning hypothesis holds with some margin $\gamma > 0$, then Adaboost classifier will have 0 training error on D in at most $T = \frac{\log(n)}{\gamma^2}$ iterations
- Proof: See notes

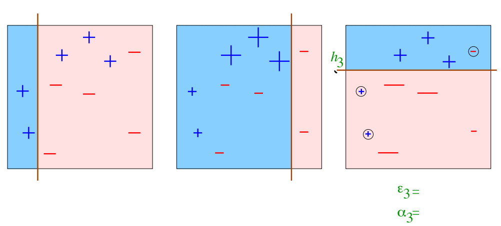
Round 1



Round 2



Round 3



Final Classifier