CSCI 5622 Machine Learning

ML Ensembles; Bagging and Boosting

DATE	READ	DUE
Today, Oct 12	Bagging & Boosting	Exper. 1 plan
Wed, Oct 14	Stacking & Rand Forests	Literature Review
Mon, Oct 19	Clustering	Peer Fdbk Lit Rev

www.RodneyNielsen.com/teaching/CSCI5622-F09/

Instructor: Rodney Nielsen

Assistant Professor Adjunct, CU Dept. of Computer Science

Research Assistant Professor, DU, Dept. of Electrical & Computer Engr.

Research Scientist, Boulder Language Technologies

PAC Learnability

- The concept class F is said to be PAC learnable by L, which uses H, if:
  - For all f in F, $\mathcal{D}$ over X, $0 < \varepsilon < 0.5$, $0 < \delta < 0.5$,
  - There is a $(1-\delta)$ chance that L will output an h such that $error_{\mathcal{D}}(h) \leq \varepsilon$
  - In time that is polynomial in $1/\varepsilon$, $1/\delta$, $|X|$, and *encoding size*(f)

PAC Learnability

- **Bound |training set| for *consistent* learners**
  - $N \geq (1/\varepsilon)(\ln|H| + \ln(1/\delta))$
- ***any* learner with a bias**
  - $N \geq (1/2\varepsilon^2)(\ln|H| + \ln(1/\delta))$
- **Unbiased Learners**
  - $N \geq (1/\varepsilon)(2^n \ln 2 + \ln(1/\delta))$
- **Given *N* training instances, a learner will *probably*, with $p = (1-\delta)$, successfully learn to *approximate* the *correct* function *f*, with error less than ε**

Ensemble Learning

- **Train multiple classifiers**
- **Use them all in some combination to make the final predictions**

ML Bootstrap AGGregatING (Bagging)

- Build several classifiers using different random subsets of the training data
- Average outputs for regression
- Most frequent vote for classification
- Typically improves performance a fair amount
  - Requires that changing the training data leads to meaningful change in the model learned

Bagging

- Improve performance by aggregating decisions from classifiers trained on different sets of data
- Generally have a single training set
- Train on several random bootstrap samples

Bootstrap Sampling

- Original training dataset is size *N*
- Draw *N* instances from the original training data at random with replacement
  - About 63% of original data will be in bootstrap
- If this produces large changes in the classifiers, Bagging will improve classification
  - Requires that learned classifiers be unstable
  - E.g., ANNs, Decision Trees are unstable, while *k*-NN is not

Bagging Results

- **Breiman achieved results ranging from a 20% to 47% reduction in the error rate**

Bagging Probabilities

- Some classifiers output class probability estimates
- These can be averaged over the bootstrapped classifiers to give final predictions
- Predict class with highest probability
- Rarely works much better than voting
- But probability estimates are important in some applications

Boosting

- Bagging uses non-focused random bootstrap resampling of training data
  - Essentially smooths the decision boundaries and adds smooth curves
- Is there a way to focus on the regions in space where the classifier is performing poorly?
  - Boosting

ML AdaBoost (Adaptive Boosting)

- Boosts a weak base learner to achieve much better performance
- This occurs over a series of *T* learning trials
- The training data is given a weighting distribution over the examples D_t
- Weights start out equally distributed across the training data
- After each iteration, D_t is updated to focus the learner on the difficult instances

AdaBoost

- Base learner can be any algorithm that can use fractional/weighted instances
- Or training data can just be resampled according to the weight distributions

AdaBoost Algorithm

- $D_1(i) \leftarrow 1/N$
- For $t = 1..T$
  - $h_t \leftarrow WeakLearn(D_t)$
  - $\varepsilon_t \leftarrow \Sigma_{i:h_t(x_i) \neq y_i} D_t(i)$
  - $\alpha_t \leftarrow 0.5 \ln((1-\varepsilon_t)/\varepsilon_t)$
  - $D_{t+1}(i) \leftarrow D_t(i) \exp(-\alpha_t y_i h_t(x_i)) \ / \ Z_t$
- $h(x) \leftarrow \text{sign}(\Sigma_{t=1..T} \alpha_t h_t(x_i))$

AdaBoost Error Bounds

- **Error on the training data drops exponentially**
  - $\varepsilon_h(X_{Tr}) \leq \exp(-2\ \Sigma_t (0.5 - \varepsilon_t^2))$
- **Generalization error**

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\frac{d(1+\log 2N/d) - \log \eta/4}{N}}$$

$$\varepsilon_h(\mathbf{X}_P) \leq \varepsilon_h(\mathbf{X}_{Tr}) + O\left(\sqrt{\frac{dT}{N}}\right)$$

  - **Can overfit as $T \rightarrow \infty$**

ML Project Discussion

- **Questions?**

Project Discussion

ML

- **Literature Review**
  - **Your words**
  - **Citations**
  - **Quotes**
  - **Block quotes**
- **Do not copy from other papers without noting it and giving them credit!!!**

Evaluation Metrics

ML

- Accuracy
- Precision, Recall, and F-measure
- Purity, …

Training for Special Data

- Data partitioning when not i.i.d.
- When should related data be distributed across data sets and when shouldn't it be?

Presentations Next Week

ML

- **Volunteers?**

Project Reviews

ML

- **Extended office hours**