STAT406- Methods of Statistical Learning Lecture 18

Matias Salibian-Barrera

UBC - Sep / Dec 2016

Random forests

- Feature ranking - relative importance of each variable
- Given a single tree T, at each node t split we can compute the sum of reductions in sum of squares (or gini or deviance measures) m_t^2
- We assign this squared measure m_t^2 to the variable (feature) used in the split

Random forests

- To each feature, we assign the sum of "squared gains" attributed to it
- For the i-th variable X_i we have

$$\mathcal{J}_i^2(T) = \begin{cases} m_t^2 & \text{if split involved } X_i \\ 0 & \text{otherwise} \end{cases}$$

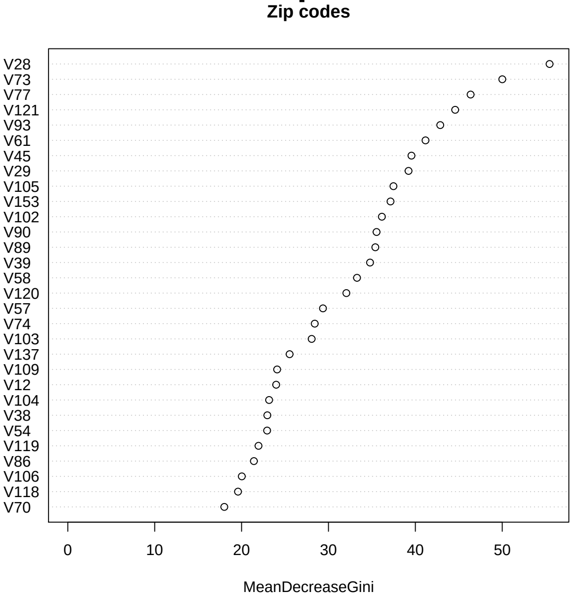
Random forests

- For a random forest we use

$$\mathcal{J}_i^2 = \sum_T \mathcal{J}_i^2(T)$$

- In other words, we sum (or average) the importance of the variable across the trees in the forest

Random Forest - plot



Computer

- Whichever software you use, learn it
- We can **help** with `R`
- We **won't teach all of** `R`
- **You are responsible** for learning it
- There are tons of on-line resoures
- Example: `http://swirlstats.com/`

Philosophy of the class

- We're here to **help you learn** (vs. teaching you)
- We'll encourage **engagement**, **curiosity** and **generosity**
- We'll have **zero tolerance for plagiarism**
- We favour **steady work** through the Term (vs. sleeping until finals)

Ensembles

- Ensembles of classifiers
- Combine classifiers trained on the same (or similar [e.g. bootstrapped]) data
- Consensus is reached by (equally weighted) voting or averaged estimated probabilities.
- Bagging and Random Forests are examples of ensembles.

Boosting

- Originally proposed for classification
- Main idea: sequentially re-train a simple classifier assigning more importance to points that were previously misclassified

Boosting

- The end result is a weighted average of all the classifiers
- Interesting ideas:
  - Not all components of the ensemble are treated equally
  - Members of the ensemble use information about other members
  - The underlying loss function has a “margin” (unlike 0-1 losses)

Boosting - AdaBoost.M1

Algorithm. Data $(y_{\mathbf{i}}, \mathbf{x}_{\mathbf{i}})$, with $y_{\mathbf{i}} \in \{-1, 1\}$

- Set initial weights $w_{\mathbf{i}} = 1/n$, $1 \leq \mathbf{i} \leq n$
- For $\mathbf{j} = 1, \ldots, K$
- Build a classifier $T_{\mathbf{j}}(\mathbf{x})$ to the data using weights $w_{\mathbf{i}}$, $1 \leq \mathbf{i} \leq n$

Boosting - AdaBoost.M1

- Let

$$e_{\mathbf{j}} = \sum_{\mathbf{i}=1}^{n} w_{\mathbf{i}} \, I\left(y_{\mathbf{i}} \neq T_{\mathbf{j}}(\mathbf{x}_{\mathbf{i}})\right) \Big/ \sum_{\ell=1}^{n} w_{\ell}$$

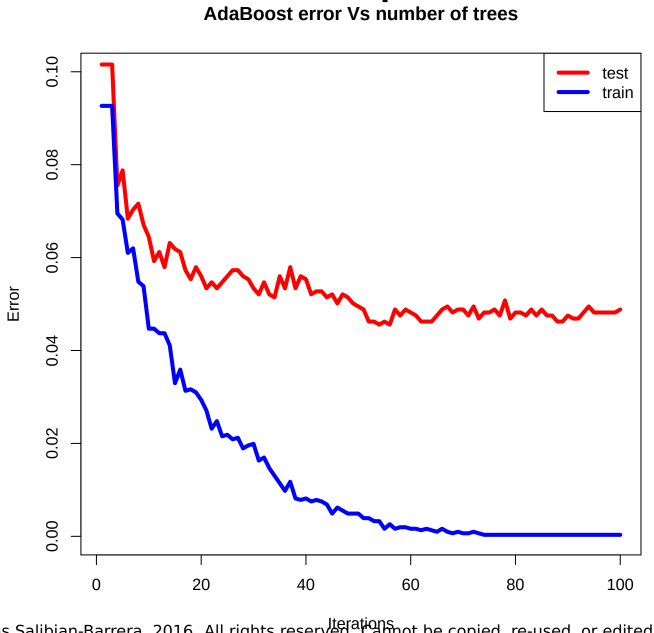
- Let $\alpha_{\mathbf{j}} = \log\left((1 - e_{\mathbf{j}})/e_{\mathbf{j}}\right)$ and

$$w_{\mathbf{i}} = w_{\mathbf{i}} \exp\left(\alpha_{\mathbf{j}} \, I\left(y_{\mathbf{i}} \neq T_{\mathbf{j}}(\mathbf{x}_{\mathbf{i}})\right)\right), \; \mathbf{i} = 1, \ldots, n$$

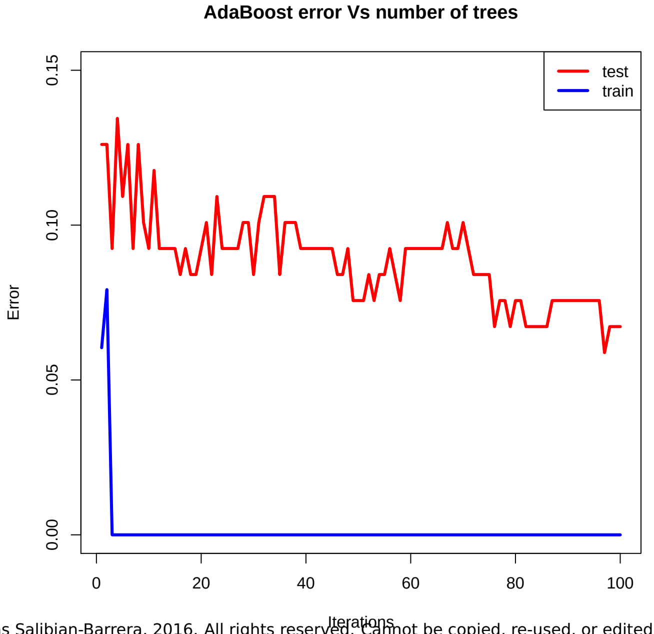
- Final classifier:

$$T(\mathbf{x}) = \operatorname{sign}\left(\sum_{\mathbf{j}=1}^{K} \alpha_{\mathbf{j}} \, T_{\mathbf{j}}(\mathbf{x})\right)$$

Error evolution - Spam data



Error evolution - Isolet



Boosting

- Boosting is fitting an **additive model**
- … using a **forward search algorithm**
- … and a **specific loss function**

Boosting

- Think of classifiers of the form

$$G(x) = \sum_{j=1}^{K} \beta_j f(\mathbf{x}, \gamma_j)$$

  where $f(\mathbf{x}, \gamma_j)$ are simple base classifiers (e.g. trees)

Boosting

- Given a data set $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$

$$\min_G \sum_{i=1}^{n} L\left(y_i, G(\mathbf{x}_i)\right) =$$

$$= \min_{\boldsymbol{\beta}, \gamma} \sum_{i=1}^{n} L\Big(y_i, \sum_{j=1}^{K} \beta_j \, f(\mathbf{x}_i, \gamma_j)\Big)$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)'$

Boosting

- Find approximate solutions sequentially
- Start with $f_0(\mathbf{x}) = 0$
- `for(j in 1:K)`
- Find

$$(\beta_j, \gamma_j) \;=\; \arg\min_{\beta,\gamma} \sum_{i=1}^{n} L\big(y_i, f_{j-1}(\mathbf{x}_i) + \beta\, f(\mathbf{x}_i, \gamma)\big)$$

- Let $f_j(\mathbf{x}) = f_{j-1}(\mathbf{x}) + \beta_j\, f(\mathbf{x}, \gamma_j)$