STAT406- Methods of Statistical Learning Lecture 17

Matias Salibian-Barrera

UBC - Sep / Dec 2016

Random forests

(1) `for(b in 1:B)`

   (a) Draw a bootstrap sample from the training data

   (b) Grow a "random forest tree" as follows: for each terminal node:

      (i) Randomly select `m` features

      (ii) Pick the best split among these

      (iii) Split the node into two children

(2) Return the ensemble of trees $(T_b)_{1 \leq b \leq B}$

Out-of-bag error estimates

- Each bagged tree is trained on a bootstrap sample
- Predict the observations not in the bootstrap sample with that tree
- One will have “about” $B/3$ predictions for each point in the training set
- These can be used to estimate the prediction error (classification error rate) without having to use CV

Example

`Example`

Random forests

- Out of bag error estimate
- For each training observation $(y_i, \mathbf{x}_i)$, obtain a prediction using only those trees in which $(y_i, \mathbf{x}_i)$ was **NOT** used
- In other words, let $\mathcal{I}_i$ the set of trees (bootstrap samples) where $(y_i, \mathbf{x}_i)$ does not appear, then

$$\hat{y}_i \; = \; \frac{1}{|\mathcal{I}_i|} \sum_{j \in \mathcal{I}} T_j(\mathbf{x}_i)$$

Random forests

- This error estimate can be computed at the same time as the trees are being built
- When this error estimate is stabilized we can stop adding trees to the ensemble

Example

```
Example
```

