RANDOM FORESTS – METHOD SUMMARY

REGRESSION | CLASSIFICATION | NONPARAMETRIC | BLACK-BOX | FEATURE SELECTION

General idea

- **Bagging ensemble** of M tree **base learners** fitted on **bootstrap** data samples

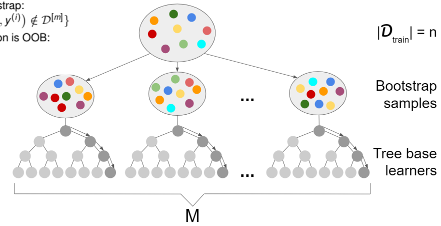
  $\Rightarrow$ Reduce **variance** by ensembling while slightly increasing **bias** by bootstrapping

  - Use unstable, **high-variance** base learners by letting trees grow to full size
  - Promoting **decorrelation** by random subset of candidate features for each split
- **Predict** via averaging (regression) or majority vote (classification) of base learners

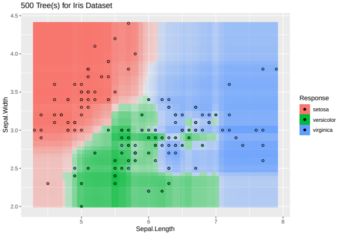
Hypothesis space

$$\mathcal{H} = \left\{ f(\mathbf{x}) : f(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} \sum_{t=1}^{T^{[m]}} c_t^{[m]} \mathbb{I}(\mathbf{x} \in Q_t^{[m]}) \right\}$$

- IB observations for m-th bootstrap:
  $\mathrm{IB}^{[m]} = \{i \in \{1, \ldots, n\} \mid (\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}^{[m]}\}$
- OOB observations for m-th bootstrap:
  $\mathrm{OOB}^{[m]} = \{i \in \{1, \ldots, n\} \mid (\mathbf{x}^{(i)}, y^{(i)}) \notin \mathcal{D}^{[m]}\}$
- Nr. of trees where i-th observation is OOB:
  $S_{\mathrm{OOB}}^{(i)} = \sum_{m=1}^{M} \mathbb{I}(i \in \mathrm{OOB}^{[m]})$.



Schematic depiction of bagging process



Prediction surface for `iris` data with 500-tree ensemble

RANDOM FORESTS – METHOD SUMMARY

Empirical risk & Optimization Just like tree base learners

Out-of-bag (OOB) error

- Ensemble prediction for obs. outside individual trees' bootstrap training sample $\Rightarrow$ unseen test sample
- Use resulting loss as unbiased estimate of **generalization error**
- Mainly useful for tuning and less for model comparison as we usually compare all models uniformly by CV

Feature importance

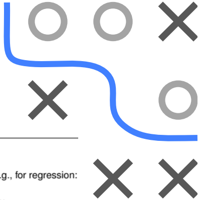
- Based on **improvement in split criterion:** aggregate improvements by all splits using j-th feature
- Based on **permutation:** permute j-th feature in OOB observations and compute impact on OOB error

Hyperparameters

- **Ensemble size**, i.e., number of trees
- **Complexity** of base learners, e.g., tree depth, min-split, min-leaf-size
- **Number of split candidates**, i.e., number of features to be considered at each split
  $\Rightarrow$ frequently used heuristics with total of p features: $\lfloor \sqrt{p} \rfloor$ for classification, $\lfloor p/3 \rfloor$ for regression

USING THE OUT-OF-BAG ERROR ESTIMATE

- Gives us a (proper) estimator of GE, computable during training
- Can even compute this for all smaller ensemble sizes (after we fitted M models)



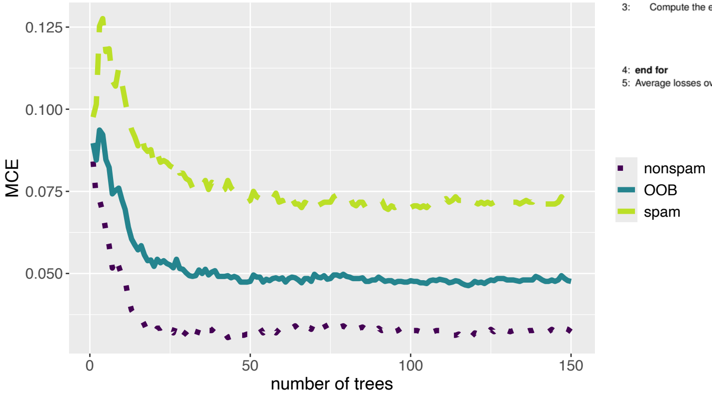
Out-Of-Bag error estimation

1: **Input:** $\mathrm{OOB}^{[m]}, \hat{b}^{[m]} \forall m \in \{1, \dots, M\}$
2: **for** $i = 1 \to n$ **do**
3: Compute the ensemble OOB prediction for observation i, e.g., for regression:

$$\hat{f}^{(i)}_{\mathrm{OOB}} = \frac{1}{S_{\mathrm{OOB}^{(i)}}} \sum_{m=1}^{M} \mathbb{I}(i \in \mathrm{OOB}^{[m]}) \cdot \hat{f}^{[m]}(\mathbf{x}^{(i)})$$

4: **end for**
5: Average losses over all observations:

$$\widehat{\mathrm{GE}}_{\mathrm{OOB}} = \frac{1}{n} \sum_{i=1}^{n} L(y^{(i)}, \hat{f}^{(i)}_{\mathrm{OOB}})$$



RANDOM FORESTS – PROS & CONS

Advantages

- + Retains most of **trees'** advantages (e.g., feature selection, feature interactions)
- + Fairly **good predictor**: mitigating base learners' variance through bagging
- + Quite **robust** w.r.t. small changes in data
- + Good with **high-dimensional** data, even in presence of noisy features
- + Easy to **parallelize**
- + Robust to its hyperparameter configuration
- + Intuitive measures of **feature importance**

Disadvantages

- – Loss of individual trees' **interpretability**
- – Can be suboptimal for **regression** when extrapolation is needed
- – **Bias** toward selecting features with many categories (same as CART)
- – Rather large model size and slow inference time for large ensembles
- – Typically inferior in **performance** to tuned gradient tree boosting.