Bagging, Random Forests

Dripta Mj
Department of Mathematics
Ramakrishna Mission Vivekananda Educational and Research Institute
Belur Math, India

Ensemble methods

- Idea is to use multiple learners and combine their predictions.
  - E.g. in ensemble of classifiers, predictions from a set of classifiers are combined

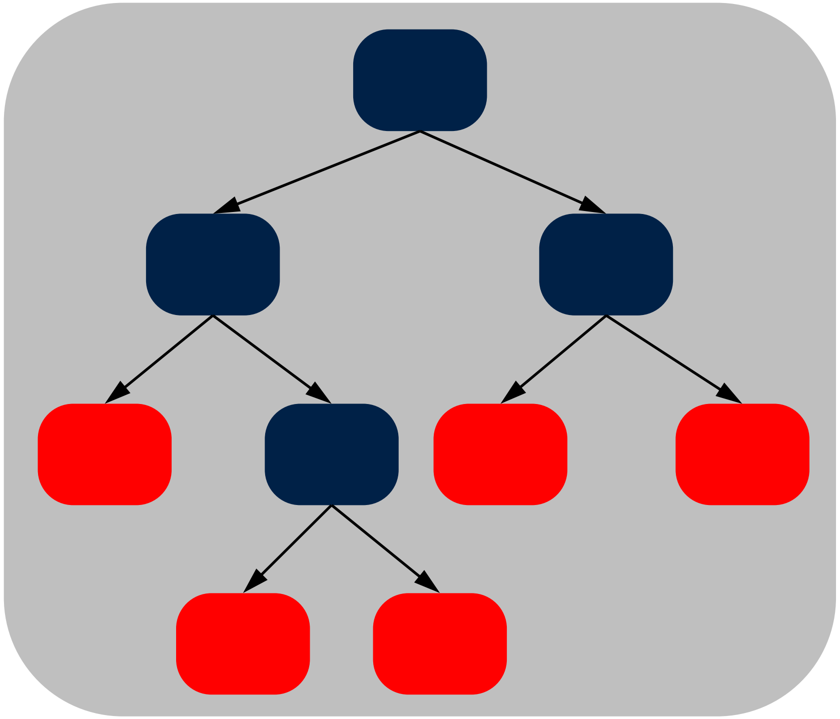
- Consider a committee of M models with uncorrelated errors, then by simply averaging the outputs of the M models the average error can be reduced by a factor of M.
  - Although in practice the errors are typically correlated and so the reduction is smaller.

- Ensemble methods can transform a "weak" learner into a strong model by taking combinations of the former.

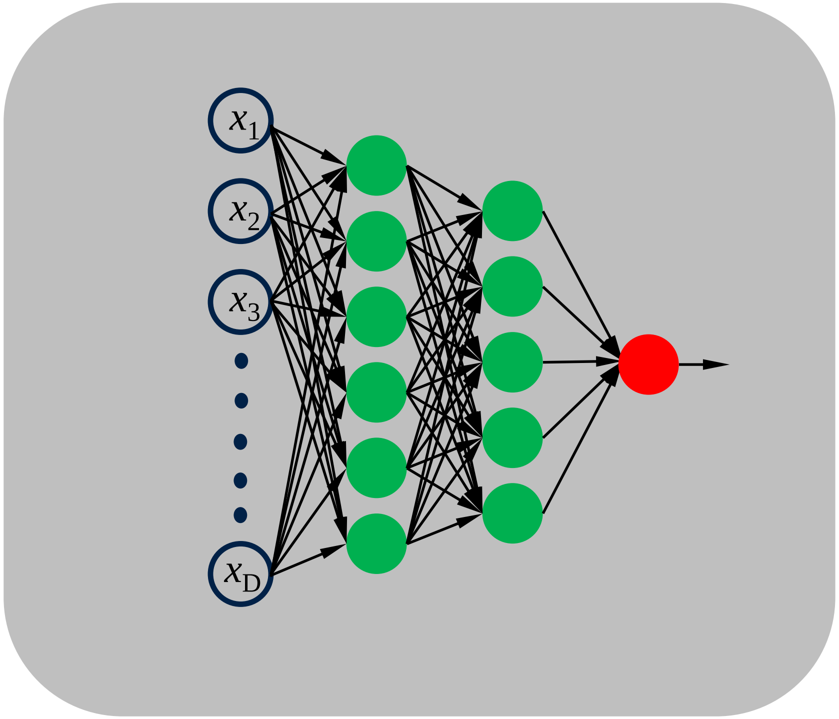
- Ensemble methods combine models such that the ensemble achieves better performance than an individual model on average.

Base Models – examples

Decision Tree



Neural Network



Can we reduce variance?

Original decomposition:

$$\mathbb{E}_{\mathbf{x},y,\mathcal{D}}\left[(g_{\mathcal{D}}(\mathbf{x}) - y)^2\right] = \underbrace{\mathbb{E}_{\mathbf{x},\mathcal{D}}\left[\left(g_{\mathcal{D}}(\mathbf{x}) - \overline{g}(\mathbf{x})\right)^2\right]}_{\text{Variance}} + \underbrace{\mathbb{E}_{\mathbf{x}}\left[\left(\overline{g}(\mathbf{x}) - \overline{y}(\mathbf{x})\right)^2\right]}_{\text{Bias}^2} + \underbrace{\mathbb{E}_{\mathbf{x},y}\left[\left(\overline{y}(\mathbf{x}) - y\right)^2\right]}_{\text{Noise}}$$

- Suppose we have M different training datasets: $\mathcal{D}_1$, $\mathcal{D}_2$, …., $\mathcal{D}_M$
- Can train a separate model on each of them: $g_{\mathcal{D}_1}, g_{\mathcal{D}_2}, \ldots., g_{\mathcal{D}_M}$
- Predictions can be obtained as the average of the trained models

$$\widehat{g}(\mathbf{x}) = \frac{1}{M}\sum_{m=1}^{M} g_{\mathcal{D}_m}(\mathbf{x}) \to \overline{g}(\mathbf{x}) \qquad \text{as} \quad M \to \infty$$

- As $\widehat{g}(\mathbf{x}) \to \overline{g}(\mathbf{x})$, the variance term $\mathbb{E}\left[(\widehat{g}(\mathbf{x}) - \overline{g}(\mathbf{x}))^2\right] \to 0$
- Issue: Don't have M different training datasets.

Bootstrap Aggregating

- Bagging: Bootstrap Aggregating
- Bootstrap: Replicate given dataset by sampling with replacement.
- Example:

$$\text{Original data}: \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \mathbf{x}^{(4)}, \mathbf{x}^{(5)}\}$$

$$\text{Bootstrap 1}: \{\mathbf{x}^{(4)}, \mathbf{x}^{(1)}, \mathbf{x}^{(3)}, \mathbf{x}^{(4)}, \mathbf{x}^{(2)}\}$$

$$\text{Bootstrap 2}: \{\mathbf{x}^{(5)}, \mathbf{x}^{(5)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \mathbf{x}^{(2)}\}$$

$$\text{Bootstrap 3}: \{\mathbf{x}^{(2)}, \mathbf{x}^{(1)}, \mathbf{x}^{(1)}, \mathbf{x}^{(3)}, \mathbf{x}^{(1)}\}$$

- Bootstrap samples are independent realizations of the original data.

Bagging Algorithm

for $m = 1$ to M **do**

- Draw a bootstrap sample dataset $\mathcal{D}_m$ from the training dataset $\mathcal{D}$.
  - The size of $\mathcal{D}_m$ should be same as $\mathcal{D}$.
- Train a base model T_m on the dataset $\mathcal{D}_m$.

end for

- Output ensemble models: $\{T_1, T_2, ..., T_M\}$
- Prediction for a new example $\mathbf{x}^*$:
  - Regression:

$$\overline{y}_M(\mathbf{x}^*) = \frac{1}{M}\sum_{m=1}^{M} T_m(\mathbf{x}^*)$$

  - Classification:

$$\overline{y}_M(\mathbf{x}^*) = \text{majority vote}\{C_1(\mathbf{x}^*), C_2(\mathbf{x}^*), ..., C_M(\mathbf{x}^*)\}$$

where $C_m(\mathbf{x}^*)$ is the class prediction of the mth model.

Bagging – Random Forests

- Bagging gives the average of predictions of a model fit to many Bootstrap samples.
- Bagging reduces the variance as it averages the fits from many independent datasets (bootstrap samples).
- Issue with Bagging:
  - Similar decision trees can be formed by different Bootstrap samples.
- Random Forests address the issue.
- In Random Forests, each Bootstrap sample produces a different decision tree.
- The final output is the average of the predictions from all the trees.

Bootstrap samples

Original Dataset

Instance	Weather	Health	Teaching	Topic Importance	Going to class?
1	Hot	Good	Interesting	Medium	Yes
2	Cold	Average	Boring	High	Yes
3	Cold	Sick	Mediocre	Medium	No
4	Mild	Average	Interesting	High	Yes
5	Rainy	Sick	Mediocre	Low	No
6	Hot	Good	Boring	High	Yes
7	Rainy	Good	Mediocre	Medium	No
8	Mild	Good	Mediocre	Medium	Yes

BOOTSTRAP-1

Instance	Weather	Health	Teaching	Topic Importance	Going to class?
7	Rainy	Good	Mediocre	Medium	No
2	Cold	Average	Boring	High	Yes
6	Hot	Good	Boring	High	Yes
2	Cold	Average	Boring	High	Yes
5	Rainy	Sick	Mediocre	Low	No
6	Hot	Good	Boring	High	Yes
8	Mild	Good	Mediocre	Medium	Yes
7	Rainy	Good	Mediocre	Medium	No

BOOTSTRAP-2

Instance	Weather	Health	Teaching	Topic Importance	Going to class?
6	Hot	Good	Boring	High	Yes
6	Hot	Good	Boring	High	Yes
1	Hot	Good	Interesting	Medium	Yes
4	Mild	Average	Interesting	High	Yes
5	Rainy	Sick	Mediocre	Low	No
1	Hot	Good	Interesting	Medium	Yes
7	Rainy	Good	Mediocre	Medium	No
5	Rainy	Sick	Mediocre	Low	No

BOOTSTRAP-3

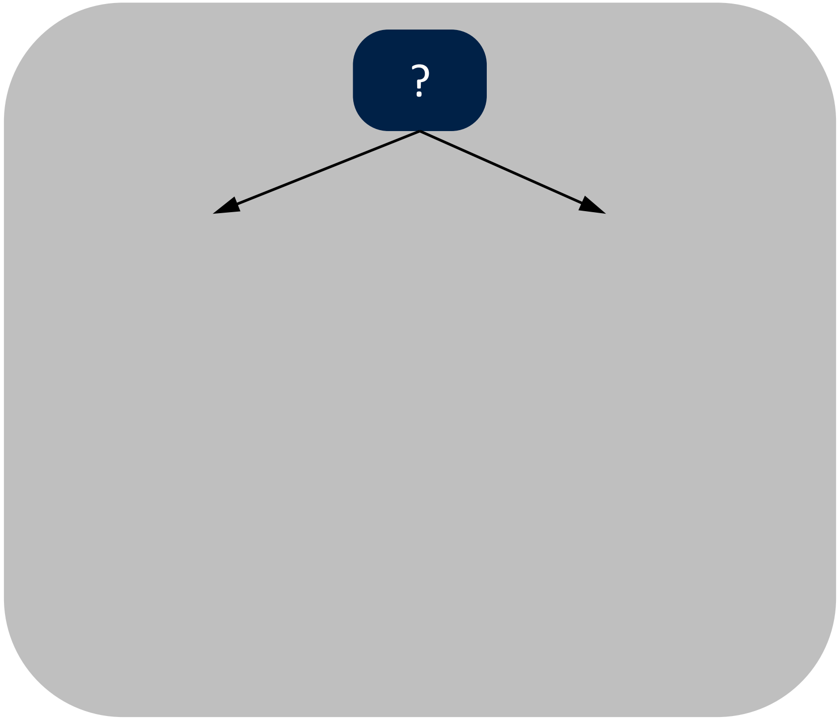
Instance	Weather	Health	Teaching	Topic Importance	Going to class?
6	Hot	Good	Boring	High	Yes
2	Cold	Average	Boring	High	Yes
4	Mild	Average	Interesting	High	Yes
4	Mild	Average	Interesting	High	Yes
1	Hot	Good	Interesting	Medium	Yes
5	Rainy	Sick	Mediocre	Low	No
2	Cold	Average	Boring	High	Yes
4	Mild	Average	Interesting	High	Yes

Random Forests – example

- k variables are selected at random, where $k < D$. Default: $k = \sqrt{D}$
  - Here $k = 2$.
- Of the k selected features, the best feature (according to some criteria) is used for splitting.

BOOTSTRAP-1

Instance	Weather	Health	Teaching	Topic Importance	Going to class?
7	Rainy	Good	Mediocre	Medium	No
2	Cold	Average	Boring	High	Yes
6	Hot	Good	Boring	High	Yes
2	Cold	Average	Boring	High	Yes
5	Rainy	Sick	Mediocre	Low	No
6	Hot	Good	Boring	High	Yes
8	Mild	Good	Mediocre	Medium	Yes
7	Rainy	Good	Mediocre	Medium	No

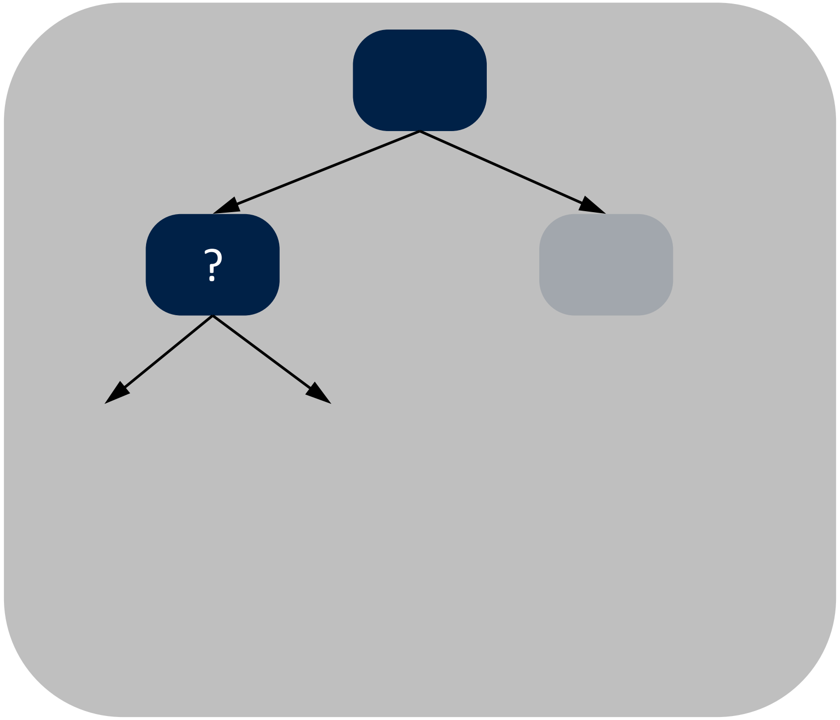


Random Forests – example

- k variables are selected at random, where $k < D$. Default: $k = \sqrt{D}$
  - Here $k = 2$.
- Of the k selected features, the best feature (according to some criteria) is used for splitting.
- At the next node, k features are again selected at random and splitting is done using the best feature.

BOOTSTRAP-1

Instance	Weather	Health	Teaching	Topic Importance	Going to class?
7	Rainy	Good	Mediocre	Medium	No
2	Cold	Average	Boring	High	Yes
6	Hot	Good	Boring	High	Yes
2	Cold	Average	Boring	High	Yes
5	Rainy	Sick	Mediocre	Low	No
6	Hot	Good	Boring	High	Yes
8	Mild	Good	Mediocre	Medium	Yes
7	Rainy	Good	Mediocre	Medium	No



Random Forests – example

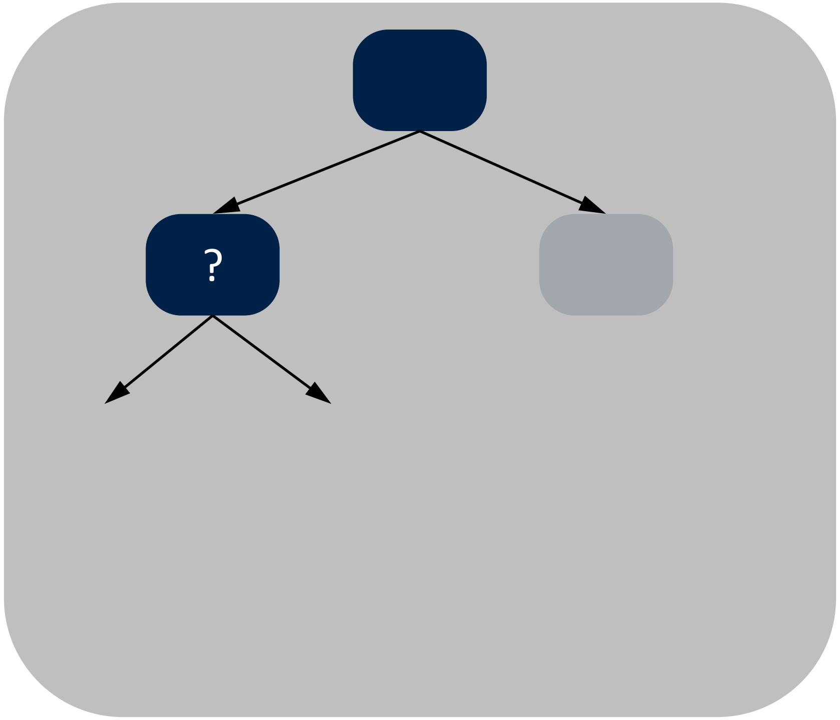
- k variables are selected at random, where $k < D$. Default: $k = \sqrt{D}$
  - Here $k = 2$.

- Of the k selected features, the best feature (according to some criteria) is used for splitting.

- At the next node, k features are again selected at random and splitting is done using the best feature.

BOOTSTRAP-1

Instance	Weather	Health	Teaching	Topic Importance	Going to class?
7	Rainy	Good	Mediocre	Medium	No
2	Cold	Average	Boring	High	Yes
6	Hot	Good	Boring	High	Yes
2	Cold	Average	Boring	High	Yes
5	Rainy	Sick	Mediocre	Low	No
6	Hot	Good	Boring	High	Yes
8	Mild	Good	Mediocre	Medium	Yes
7	Rainy	Good	Mediocre	Medium	No

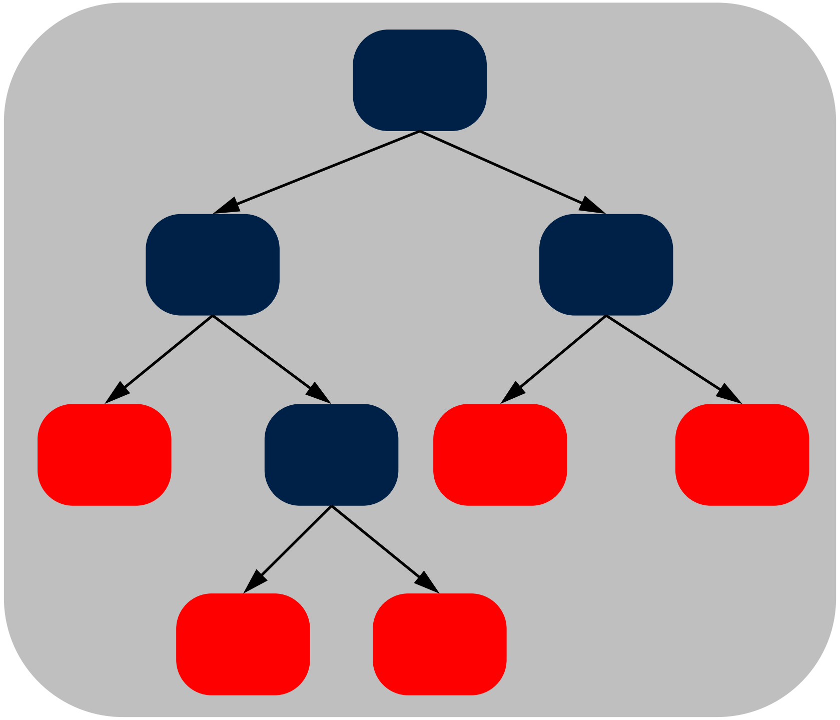


Random Forests – example

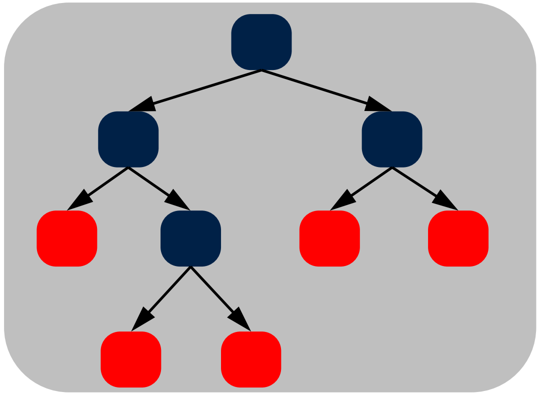
- k variables are selected at random, where $k < D$. Default: $k = \sqrt{D}$
  - Here $k = 2$.
- Of the k selected features, the best feature (according to some criteria) is used for splitting.
- At the next node, k features are again selected at random and splitting is done using the best feature.
- The process is repeated till the end.

BOOTSTRAP-1

Instance	Weather	Health	Teaching	Topic Importance	Going to class?
7	Rainy	Good	Mediocre	Medium	No
2	Cold	Average	Boring	High	Yes
6	Hot	Good	Boring	High	Yes
2	Cold	Average	Boring	High	Yes
5	Rainy	Sick	Mediocre	Low	No
6	Hot	Good	Boring	High	Yes
8	Mild	Good	Mediocre	Medium	Yes
7	Rainy	Good	Mediocre	Medium	No

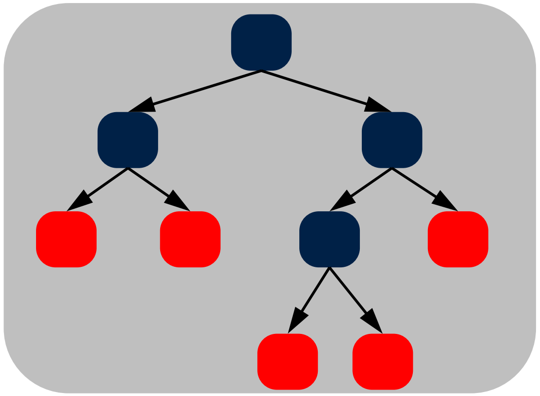


Tree ensembles



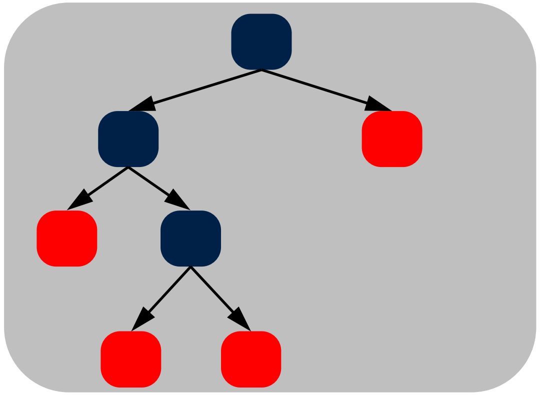
BOOTSTRAP-1

Instance	Weather	Health	Teaching	Topic Importance	Going to class?
7	Rainy	Good	Mediocre	Medium	No
2	Cold	Average	Boring	High	Yes
6	Hot	Good	Boring	High	Yes
2	Cold	Average	Boring	High	Yes
5	Rainy	Sick	Mediocre	Low	No
6	Hot	Good	Boring	High	Yes
8	Mild	Good	Mediocre	Medium	Yes
7	Rainy	Good	Mediocre	Medium	No



BOOTSTRAP-2

Instance	Weather	Health	Teaching	Topic Importance	Going to class?
6	Hot	Good	Boring	High	Yes
6	Hot	Good	Boring	High	Yes
1	Hot	Good	Interesting	Medium	Yes
4	Mild	Average	Interesting	High	Yes
5	Rainy	Sick	Mediocre	Low	No
1	Hot	Good	Interesting	Medium	Yes
7	Rainy	Good	Mediocre	Medium	No
5	Rainy	Sick	Mediocre	Low	No



BOOTSTRAP-3

Instance	Weather	Health	Teaching	Topic Importance	Going to class?
6	Hot	Good	Boring	High	Yes
2	Cold	Average	Boring	High	Yes
4	Mild	Average	Interesting	High	Yes
4	Mild	Average	Interesting	High	Yes
1	Hot	Good	Interesting	Medium	Yes
5	Rainy	Sick	Mediocre	Low	No
2	Cold	Average	Boring	High	Yes
4	Mild	Average	Interesting	High	Yes

Algorithm – regression

for m = 1 to M **do**

- Draw a bootstrap sample dataset $\mathcal{D}_m$ from the training dataset $\mathcal{D}$. The size of $\mathcal{D}_m$ should be same as $\mathcal{D}$.
- Construct a decision tree T_m using the bootstraped dataset $\mathcal{D}_m$ based on the following rules:
  - Select k features randomly from the D features.
  - From the k features, select the best feature (based on some criteria) for splitting
  - Split the node using the best feature.
  - Repeat the process till the stopping criteria is attained.

end for

- Output tree ensembles: $\{T_1, T_2, ..., T_M\}$
- Prediction at a new point $\mathbf{x}^*$:
  - Regression:

$$\overline{y}_M(\mathbf{x}^*) = \frac{1}{M}\sum_{m=1}^{M} T_m(\mathbf{x}^*)$$

Algorithm – classification

for m = 1 to M **do**

- Draw a bootstrap sample dataset $\mathcal{D}_m$ from the training dataset $\mathcal{D}$. The size of $\mathcal{D}_m$ should be same as $\mathcal{D}$.
- Construct a decision tree T_m using the bootstraped dataset $\mathcal{D}_m$ based on the following rules:
  - Select k features randomly from the D features.
  - From the k features, select the best feature (based on some criteria) for splitting
  - Split the node using the best feature.
  - Repeat the process till the stopping criteria is attained.

end for

- Output tree ensembles: $\{T_1, T_2, ..., T_M\}$
- Prediction at a new point $\mathbf{x}^*$:
  - Classification:

$$\overline{y}_M(\mathbf{x}^*) = \text{majority vote}\{C_1(\mathbf{x}^*), C_2(\mathbf{x}^*), ..., C_M(\mathbf{x}^*)\}$$

where $C_m(\mathbf{x}^*)$ is the class prediction of the mth random forest.

Out-of-bag error

- Test error can be assessed without cross-validation or validation set
- On an average, each bagged tree uses around two-third of the original training dataset.
  - The left-out examples are known as "out-of-bag" (OOB) examples.
- Example:

$$\text{Original data}: \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \mathbf{x}^{(4)}, \mathbf{x}^{(5)}\}$$

Bootstraps		OOB examples
Bootstrap 1 :	$\{\mathbf{x}^{(4)}, \mathbf{x}^{(1)}, \mathbf{x}^{(3)}, \mathbf{x}^{(4)}, \mathbf{x}^{(2)}\}$	$\{\mathbf{x}^{(5)}\}$
Bootstrap 2 :	$\{\mathbf{x}^{(5)}, \mathbf{x}^{(5)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \mathbf{x}^{(2)}\}$	$\{\mathbf{x}^{(1)}, \mathbf{x}^{(4)}\}$
Bootstrap 3 :	$\{\mathbf{x}^{(2)}, \mathbf{x}^{(1)}, \mathbf{x}^{(1)}, \mathbf{x}^{(3)}, \mathbf{x}^{(1)}\}$	$\{\mathbf{x}^{(4)}, \mathbf{x}^{(5)}\}$

- The prediction for the nth example $\mathbf{x}^{(n)}$ can be made using the bagging trees where $\mathbf{x}^{(n)}$ was an OOB example.
- So roughly there will be around $M/3$ predictions for each example.

Out-of-bag error

- Final OOB prediction:
  - Regression: Average of the predicted outputs.
  - Classification: Majority vote.

- In this way OOB predictions can obtained for all the N examples in the training dataset.

- OOB error: Error can be computed from the OOB predictions of the N examples.
  - The resulting error provides an estimate of the test error for the bagged model.

Why RFs work?

- The main difference with bagging is that the features are chosen from random subsets.
- The decision trees in bagging can end up being correlated as the same features are tend to be used repeatedly for splitting different bootstrap samples.
  - In Random Forests the splitting features are selected from random subsets and so the correlation between trees decreases.
- Also by restricting the number of features the computations are reduced; the trees are learnt faster.