Decision Tree Ensembles

Random Forest & Gradient Boosting

CSE 416 Quiz Section

4/26/2018

Kaggle Titanic Data

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S

Kaggle Titanic Data - Training Variable Selection

Drop Label Drop Drop Drop

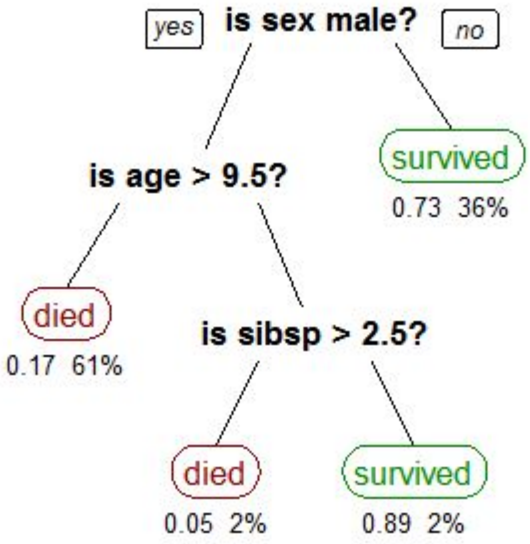
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S

Kaggle Titanic Data - Training Set

Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	3	male	22	1	0	7.25	S
1	1	female	38	1	0	71.2833	C
1	3	female	26	0	0	7.925	S
1	1	female	35	1	0	53.1	S
0	3	male	35	0	0	8.05	S
0	3	male		0	0	8.4583	Q
0	1	male	54	0	0	51.8625	S

Decision Tree

Titanic Survival Classification Tree



Decision Tree

Like Mr. Bean's car, a decision tree is

- **Super Simple** - They are often easier to interpret than even linear models.
- **Very Efficient** - The computation cost is minimal.
- **Weak** - It has low predictive power on its own. It's in a class of models called the "weak learners".

Random Forest

1. Randomly sample the rows (w/replacement) and columns (w/o replacement) at each node and build a deep tree.

Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	3	male	22	1	0	7.25	S
1	1	female	38	1	0	71.2833	C
1	3	female	26	0	0	7.925	S
1	1	female	35	1	0	53.1	S
0	3	male	35	0	0	8.05	S
0	3	male		0	0	8.4583	Q
0	1	male	54	0	0	51.8625	S
0	3	male	2	3	1	21.075	S
1	3	female	27	0	2	11.1333	S
1	2	female	14	1	0	30.0708	C

Random Forest

1. Randomly sample the rows (w/replacement) and columns (w/o replacement) at each node and build a deep tree.
2. Repeat many times (1,000+)

Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	3	male	22	1	0	7.25	S
1	1	female	38	1	0	71.2833	C
1	3	female	26	0	0	7.925	S
1	1	female	35	1	0	53.1	S
0	3	male	35	0	0	8.05	S
0	3	male		0	0	8.4583	Q
0	1	male	54	0	0	51.8625	S
0	3	male	2	3	1	21.075	S
1	3	female	27	0	2	11.1333	S
1	2	female	14	1	0	30.0708	C

Random Forest

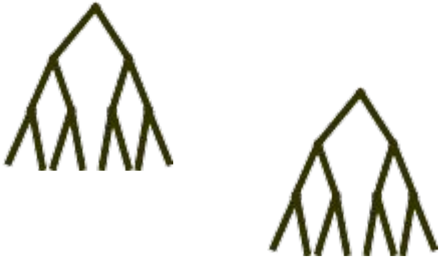
1. Randomly sample the rows (w/replacement) and columns (w/o replacement) at each node and build a deep tree.
2. Repeat many times (1,000+)
3. Ensemble trees by majority vote (ie. if 300 out of 1,000 trees predicts a given individual dies then probability of death is 30%).

Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	3	male	22	1	0	7.25	S
1	1	female	38	1	0	71.2833	C
1	3	female	26	0	0	7.925	S
1	1	female	35	1	0	53.1	S
0	3	male	35	0	0	8.05	S
0	3	male		0	0	8.4583	Q
0	1	male	54	0	0	51.8625	S
0	3	male	2	3	1	21.075	S
1	3	female	27	0	2	11.1333	S
1	2	female	14	1	0	30.0708	C

Random Forest - Tree 1

Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	3	male	22	1	0	7.25	S
1	1	female	38	1	0	71.2833	C
1	3	female	26	0	0	7.925	S
1	1	female	35	1	0	53.1	S
0	3	male	35	0	0	8.05	S
0	3	male		0	0	8.4583	Q
0	1	male	54	0	0	51.8625	S
0	3	male	2	3	1	21.075	S
1	3	female	27	0	2	11.1333	S
1	2	female	14	1	0	30.0708	C

Random Forest - Tree 2



Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	3	male	22	1	0	7.25	S
1	1	female	38	1	0	71.2833	C
1	3	female	26	0	0	7.925	S
1	1	female	35	1	0	53.1	S
0	3	male	35	0	0	8.05	S
0	3	male		0	0	8.4583	Q
0	1	male	54	0	0	51.8625	S
0	3	male	2	3	1	21.075	S
1	3	female	27	0	2	11.1333	S
1	2	female	14	1	0	30.0708	C

Random Forest - Several Trees

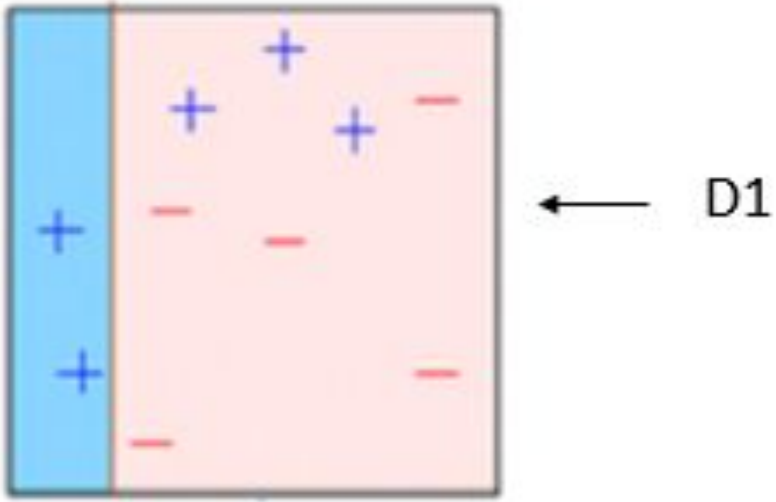
Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	3	male	22	1	0	7.25	S
1	1	female	38	1	0	71.2833	C
1	3	female	26	0	0	7.925	S
1	1	female	35	1	0	53.1	S
0	3	male	35	0	0	8.05	S
0	3	male		0	0	8.4583	Q
0	1	male	54	0	0	51.8625	S
0	3	male	2	3	1	21.075	S
1	3	female	27	0	2	11.1333	S
1	2	female	14	1	0	30.0708	C

Random Forest

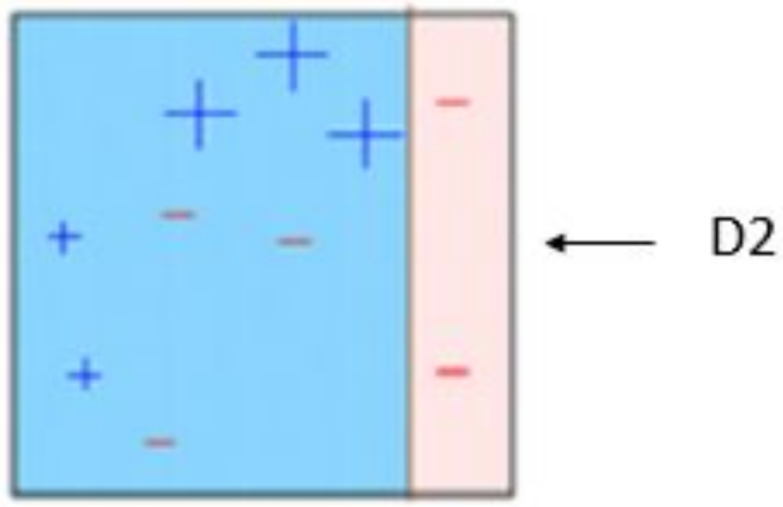
Like a Honda CR-V, Random Forest is

- **Versatile** - It can do classification, regression, missing value imputation, clustering, feature importance, and works well on most data sets right out of the box.
- **Efficient** - Trees can be grown in parallel.
- **Low Maintenance** - Parameter tuning is often not needed. You can tune number of columns to subsample, but it usually doesn’t change much.

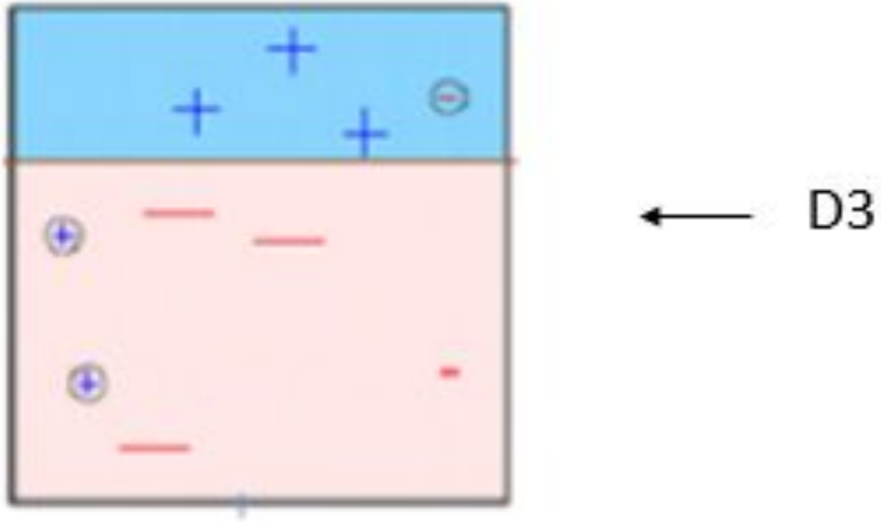
Adaboost Example - Tree Stump 1



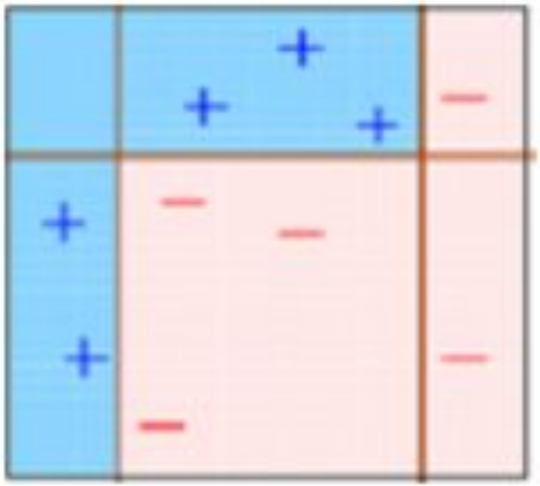
Adaboost Example - Tree Stump 2



Adaboost Example - Tree Stump 3



Adaboost Example - Ensemble



D4

$$\hat{y} = sign\left(\sum_{t=1}^{T} \hat{\mathbf{w}}_t f_t(\mathbf{x})\right)$$

Gradient Boosting

Given this process, how quickly do you think this leads to overfitting?

Gradient Boosting

Given this process, how quickly do you think this leads to overfitting?

The surprising answer is not very fast.

Gradient Boosting

Like the original hummer, Gradient Boosting is

- **Powerful** - On most real world data sets, it is hard to beat in predictive power. It can handle missing values natively. It is fairly robust to unbalanced data.
- **High Maintenance** - There are many parameters to tune. Extra precautions must be taken to prevent overfitting.
- **Expensive** - Boosting is inherently sequential and computationally expensive. However, it is a lot faster now with new tools like XGBoost (UW) and Lightgbm (Microsoft).