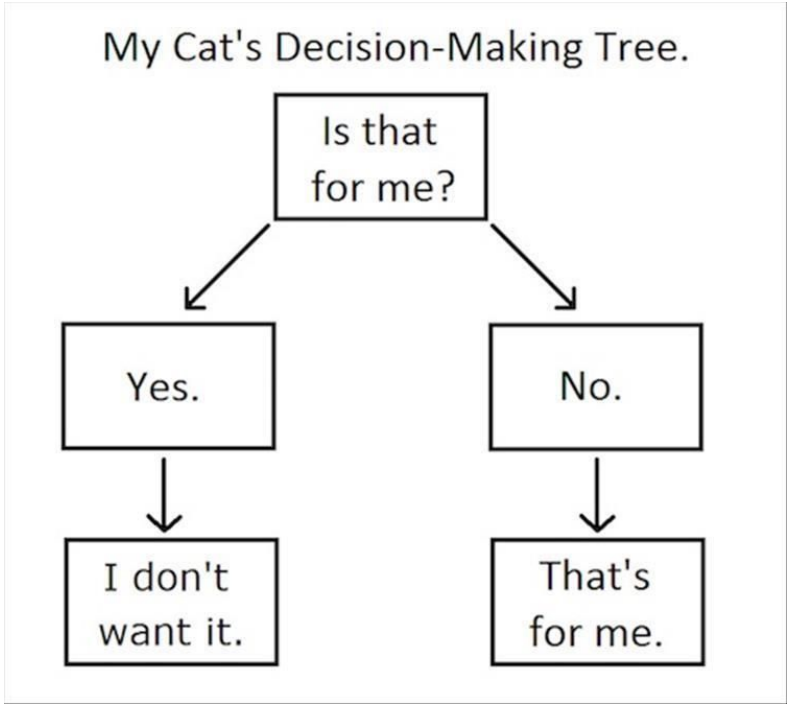
Обучение с учителем: Деревья решений (Decision Trees). Случайный лес (Random Forest).

Екатерина Кондратьева


My Cat's Decision-Making Tree.
Is that for me?
Yes.
No.
I don't want it.
That's for me.

Деревья Решений

Дерево принятия решений (также может называться деревом классификации или регрессионным деревом) — средство поддержки принятия решений, использующееся в машинном обучении, анализе данных и статистике. Структура дерева представляет собой «листья» и «ветки». На рёбрах («ветках») дерева решения записаны атрибуты, от которых зависит целевая функция, в «листьях» записаны значения целевой функции, а в остальных узлах — атрибуты, по которым различаются случаи. Чтобы классифицировать новый случай, надо спуститься по дереву до листа и выдать соответствующее значение.

Деревья принятия решений характеризуются:

- Критерием Информативности (в sklearn `gini`,`entropy`)
- Критерий Останова

https://www.hse.ru/mirror/pubs/share/215285956

Чем характеризуется дерево?

1. Видом предикатов в вершинах (одномерные, многомерные, метрические);

2. Функционалом качества Q(X, j, t);

3. Критерием останова;

4. Методом обработки пропущенных значений;

5. Методом стрижки.

https://basegroup.ru/sites/default/files/treegraph.gif

Классификация на деревьях решений

Основные параметры класса `sklearn.tree.DecisionTreeClassifier`:

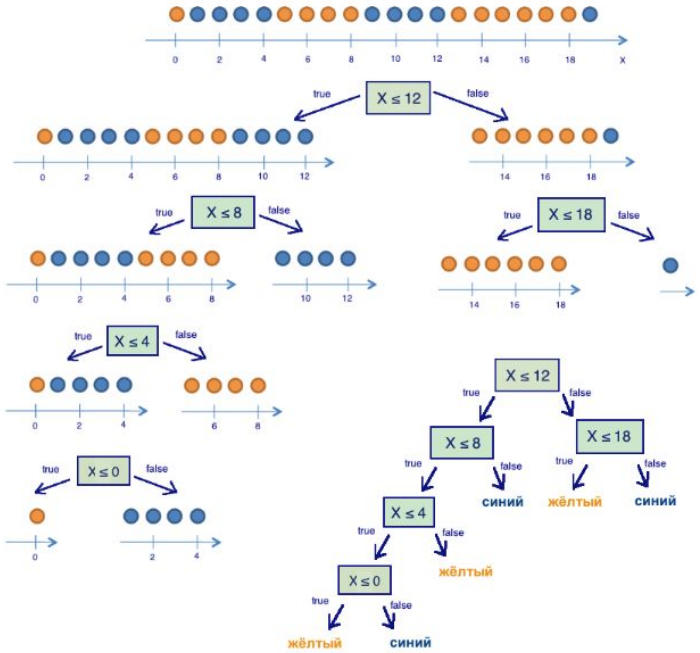
- `max_depth` – максимальная глубина дерева

- `max_features` — максимальное число признаков, по которым ищется лучшее разбиение в дереве (это нужно потому, что при большом количестве признаков будет "дорого" искать лучшее (по критерию типа прироста информации) разбиение среди *всех* признаков)

- `min_samples_leaf` – минимальное число объектов в листе. У этого параметра есть понятная интерпретация: скажем, если он равен 5, то дерево будет порождать только те классифицирующие правила, которые верны как минимум для 5 объектов

https://habr.com/ru/company/ods/blog/322534/

Понятие энтропии. Принцип построения решающего правила в деревьях



Что означает глубина дерева (max depth)?

https://habr.com/ru/company/ods/blog/322534/

Критерии качества разбиения в задаче:

Неопределенность Джини (Gini impurity): $G = 1 - \sum_k (p_k)^2$

Максимизацию этого критерия можно интерпретировать как максимизацию числа пар объектов одного класса, оказавшихся в одном поддереве.

Ошибка классификации (misclassification error): $E = 1 - \max_k p_k$

В случае задачи бинарной классификации (p+– вероятность объекта иметь метку +) энтропия и неопределенность

Критерии реализованы в [алгоритмах]

Регрессия на деревьях решений

При прогнозировании количественного признака идея построения дерева остается та же, но меняется критерий качества:

- Дисперсия вокруг среднего: $D = \frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - \frac{1}{\ell}\sum_{i=1}^{\ell} y_i)^2$

где ℓ – число объектов в листе, yi – значения целевого признака. Попросту говоря, минимизируя дисперсию вокруг среднего, мы ищем признаки, разбивающие выборку таким образом, что значения целевого признака в каждом листе примерно равны.


LIFE IS JUST
A RANDOM FOREST

Случайные Леса Решений

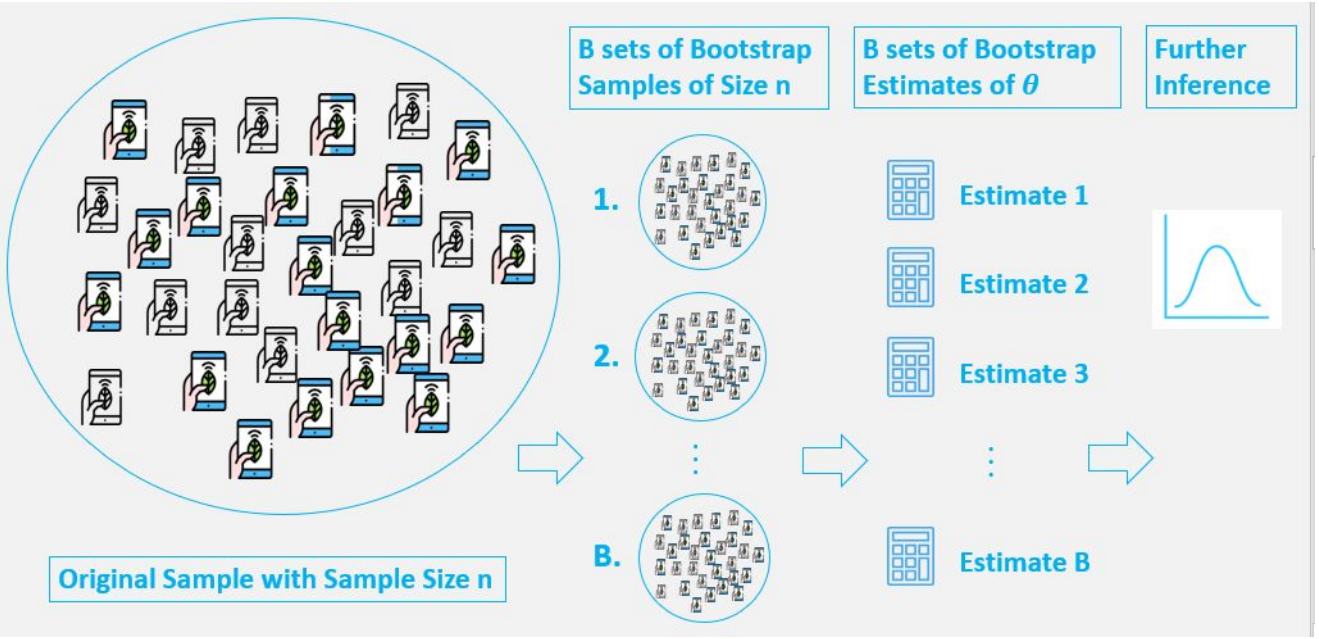
Random forest (с англ. — «случайный лес») — алгоритм машинного обучения заключающийся в использовании комитета (ансамбля) решающих деревьев.

- Сгенерируем случайную подвыборку с повторениями размером N из обучающей выборки.
- Построим решающее дерево, классифицирующее образцы данной подвыборки, причём в ходе создания очередного узла дерева будем выбирать набор признаков, на основе которых производится разбиение (не из всех M признаков, а лишь из m случайно выбранных).
- Дерево строится до полного исчерпания подвыборки и не подвергается процедуре прунинга (англ. pruning — отсечение ветвей) (в отличие от решающих деревьев, построенных по таким алгоритмам, как CART или C4.5).

Алгоритм применяется для задач классификации, регрессии и кластеризации.

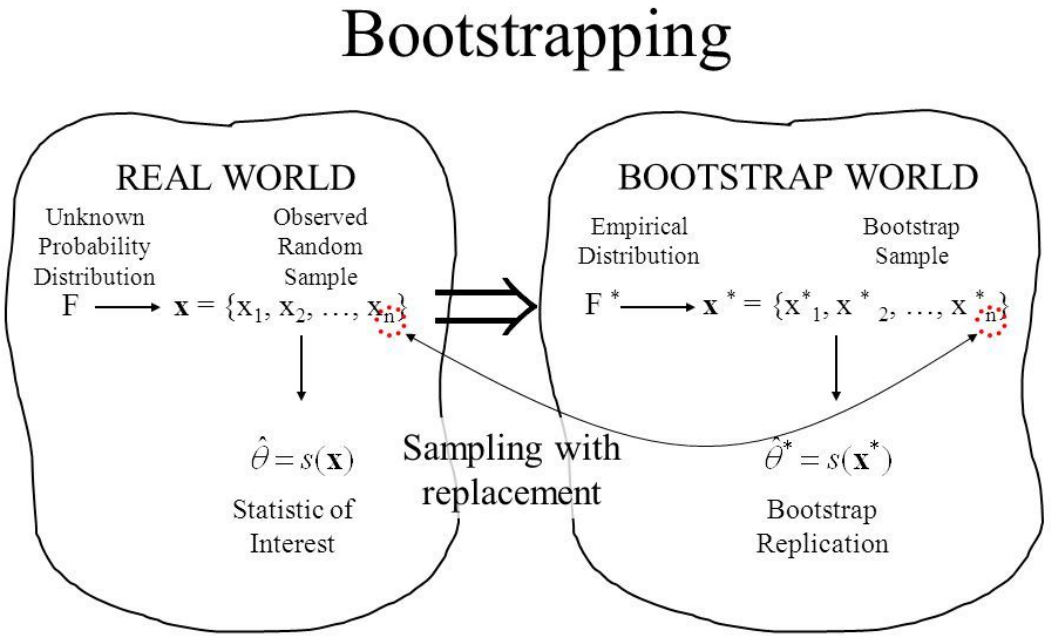
https://ru.wikipedia.org/wiki/Random_forest

Bootstrap



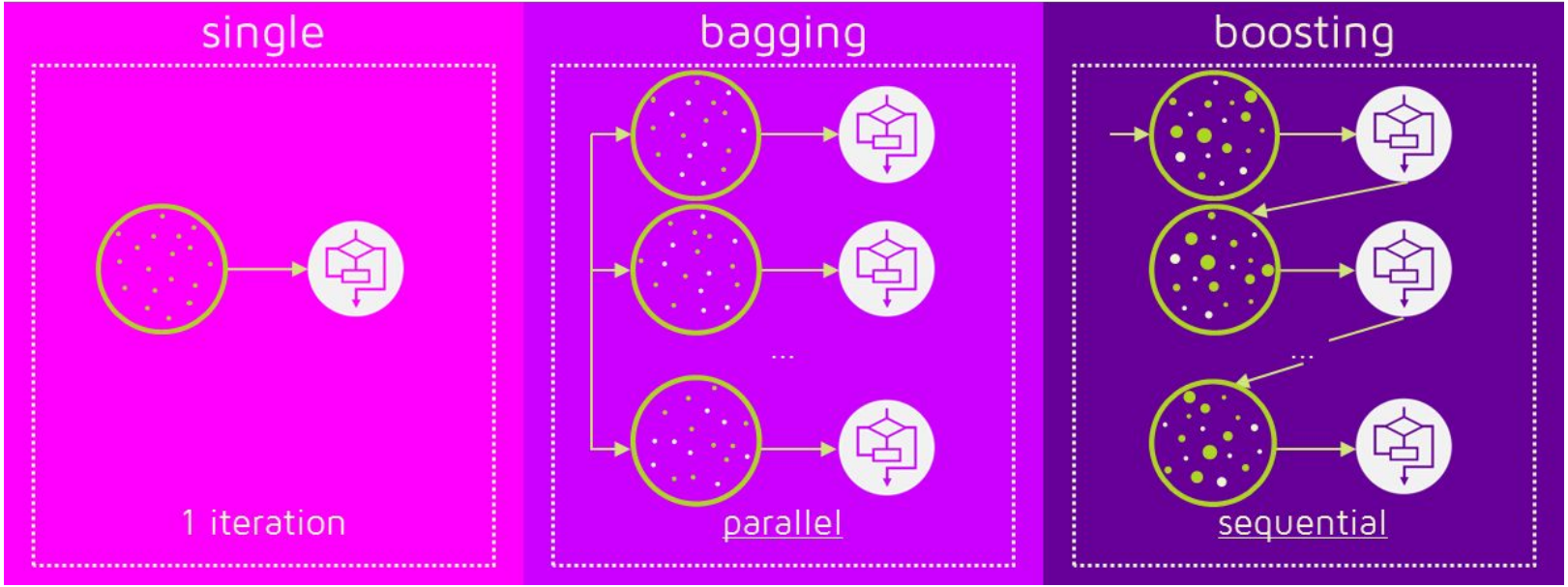
https://miro.medium.com/max/1400/1*iH5w0MBdiOlxDOCX6nmqqw.png

Bootstrap



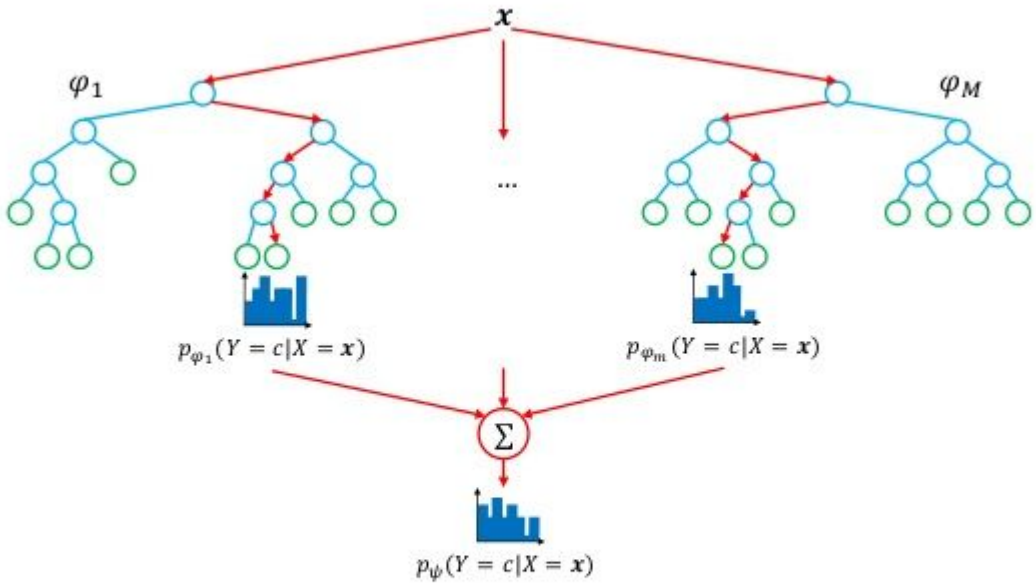
After Efron and Tibshirani, An Introduction to the Bootstrap, 1993

Bagging.



https://quantdare.com/wp-content/uploads/2016/04/bb3.png

Random forests



Randomization

- Bootstrap samples
- Random selection of $K \leqslant p$ split variables } Random Forests
- Random selection of the threshold } Extra-Trees

https://www.kdnuggets.com/wp-content/uploads/rand-forest-2.jpg

Вопросы для самопроверки:

1. В чем отличие Decision Trees от Random Forest?
2. На что влияют критерии построения решающего правила в деревьях?
3. Как интерпретировать результат модели RFC?
4. Почему важно варьировать `max_depth` дерева?

Источники:

1. Лекция https://ru.coursera.org/lecture/supervised-learning/rieshaiushchiie-dieriev-ia-HZxD1

2. https://chrisalbon.com/machine_learning/trees_and_forests/visualize_a_decision_tree/

3. https://habr.com/ru/post/171759/

4. https://www.hse.ru/mirror/pubs/share/215285956