Случайные леса — пример одного из методов ансамблей (ensemble), основанных на агрегировании результатов ансамбля более простых оценивателей. Несколько неожиданный результат использования подобных методов ансамблей — то, что целое в данном случае оказывается больше суммы составных частей. Результат «голосования» среди достаточного количества оценивателей может оказаться лучше результата любого из отдельных участников «голосования»!

RF (random forest) — это множество решающих деревьев. В задаче регрессии их ответы усредняются, в задаче классификации принимается решение голосованием по большинству. Все деревья строятся независимо по следующей схеме:

- Выбирается подвыборка обучающей выборки размера samplesize (м.б. с возвращением) – по ней строится дерево (для каждого дерева — своя подвыборка).
- Для построения каждого расщепления в дереве просматриваем max_features случайных признаков (для каждого нового расщепления — свои случайные признаки).
- Выбираем наилучшие признак и расщепление по нему (по заранее заданному критерию). Дерево строится, как правило, до исчерпания выборки (пока в листьях не останутся представители только одного класса), но в современных реализациях есть параметры, которые ограничивают высоту дерева, число объектов в листьях и число объектов в подвыборке, при котором проводится расщепление.

```
class sklearn.ensemble.RandomForestClassifier(n_estimators=10,
criterion='gini', max_depth=None, min_samples_split=2,
min_samples_leaf=1, min_weight_fraction_leaf=0.0,
max_features='auto', max_leaf_nodes=None, min_impurity_split=1e-07,
bootstrap=True, oob_score=False, n_jobs=1,
random_state=None, verbose=0, warm_start=False,
class_weight=None)
```

Число деревьев — n_estimators (чем больше, тем лучше)

Число признаков для выбора расщепления — max_features

По умолчанию он равен sqrt(n) в задачах классификации

Минимальное число объектов, при котором выполняется расщепление — min_samples_split (по-умолчанию 2)

Ограничение на число объектов в листьях — min_samples_leaf (1, оптимально 5)

Максимальная глубина деревьев — max_depth

Критерий расщепления — criterion

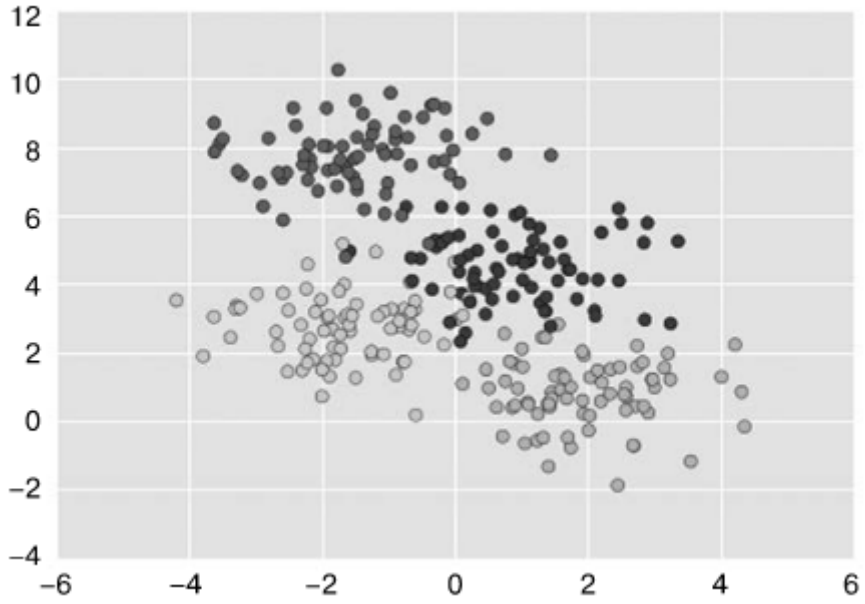
Для классификации реализованы критерии “gini” и “entropy”, которые соответствуют классическим критериям расщепления: Джини и энтропийному.

oob-ответы (ответы, которые выдавал бы алгоритм на обучающей выборке, если бы «обучался не на ней»)

О деревьях.

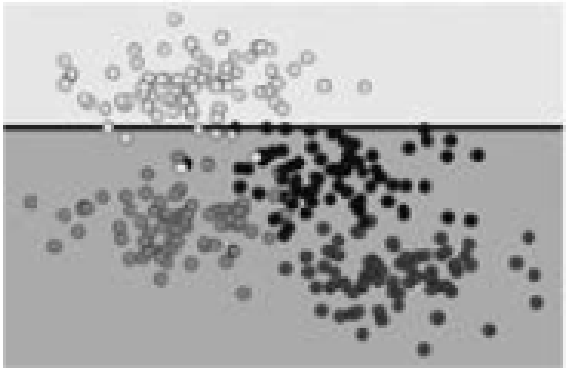
Деревья решений — исключительно интуитивно понятные способы классификации или маркирования объектов. По сути, все сводится к классификации путем задания серии уточняющих вопросов. В связанных с машинным обучением реализациях деревьев принятия решений вопросы обычно имеют вид выровненных по осям координат разбиений данных, то есть каждый узел дерева разбивает данные на две группы с помощью порогового значения одного из признаков. Рассмотрим пример:

Данные: двумерные, 4 класса

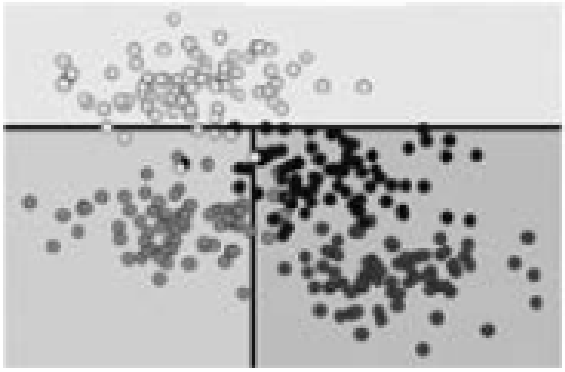


Простое дерево принятия решений для этих данных будет многократно разделять данные по одной или нескольким осям, в соответствии с определенным количественным критерием, и на каждом уровне маркировать новую область согласно большинству лежащих в ней точек. На рис. 5.69 приведена визуализация первых четырех уровней классификатора для этих данных, созданного на основе дерева принятия решений.

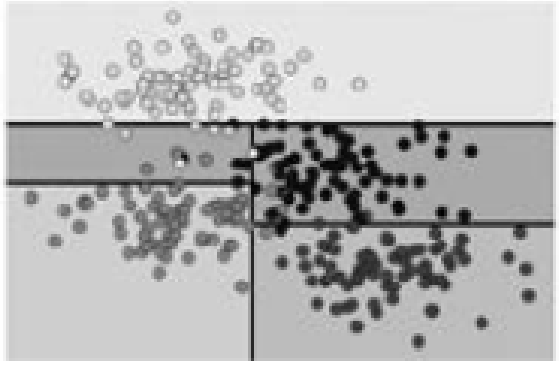
depth = 1



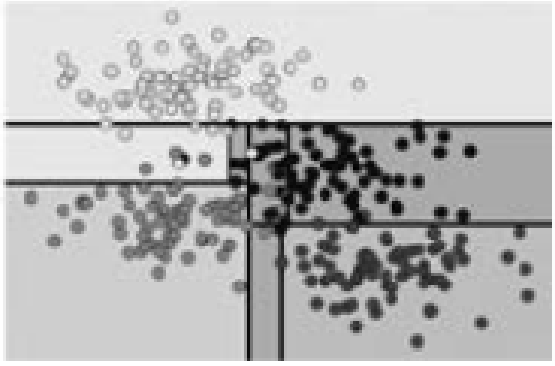
depth = 2

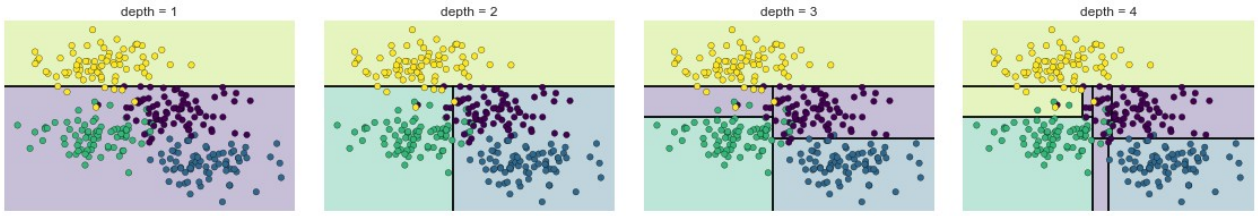


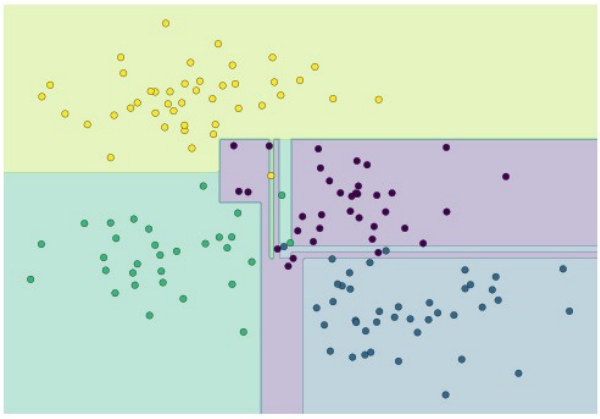
depth = 3

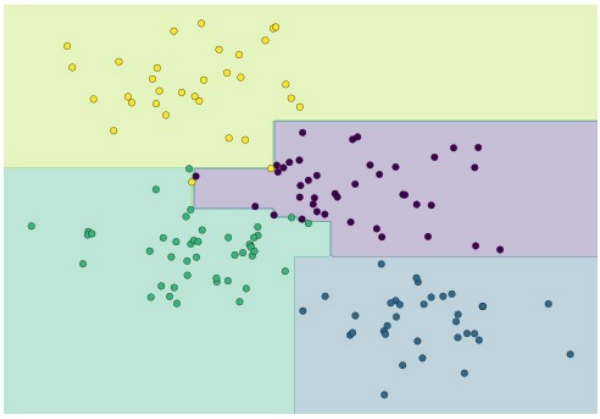


depth = 4









Обратите внимание, что после первого разбиения все точки в верхней ветке остаются неизменными, поэтому необходимости в дальнейшем ее разбиении нет. За исключением узлов, в которых присутствует только один цвет, на каждом из уровней все области снова разбиваются по одному из двух признаков.
Случайные леса — мощный метод, обладающий несколькими достоинствами.

Как обучение, так и предсказание выполняются очень быстро в силу простоты лежащих в основе модели деревьев принятия решений. Кроме того, обе задачи допускают эффективную параллелизацию, так как отдельные деревья представляют собой совершенно независимые сущности.

Вариант с несколькими деревьями дает возможность использования вероятностной классификации: решение путем «голосования» оценивателей дает оценку вероятности.

Непараметрическая модель исключительно гибка и может эффективно работать с задачами, на которых другие оцениватели оказываются недообученными.
Основной недостаток случайных лесов состоит в том, что результаты сложно интерпретировать.Чтобы сделать какие-либо выводы относительно смысла модели классификации, случайные леса — не лучший вариант.

Дерево решений как алгоритм машинного обучения - объединение логических правил вида "Значение признака x меньше a И Значение признака y меньше b… => Класс 1"

Математическое обоснование классификаторов:
Разбив исходный набор данных на две части по некому предикату, можно рассчитать энтропию каждого подмножества, после чего рассчитать среднее значение энтропии — если оно окажется меньшим чем энтропия исходного множества, значит предикат содержит некую обобщающую информацию о данных.
https://habrahabr.ru/post/171759/

```
s0 = вычисляем энтропию исходного множества

Если s0 == 0 значит:
   Все объекты исходного набора, принадлежат к одному классу
   Сохраняем этот класс в качестве листа дерева

Если s0 != 0 значит:
   Перебираем все элементы исходного множества:
      Для каждого элемента перебираем все его атрибуты:
         На основе каждого атрибута генерируем предикат, который разбивает исходное множеств
о на два подмножества
         Рассчитываем среднее значение энтропии
         Вычисляем ΔS
   Нас интересует предикат, с наибольшим значением ΔS
   Найденный предикат является частью дерева принятия решений, сохраняем его

   Разбиваем исходное множество на подмножества, согласно предикату
   Повторяем данную процедуру рекурсивно для каждого подмножества
```