**Что такое ансамблевые методы?**

***Ансамблевые методы*** — это парадигма машинного обучения, где несколько моделей (часто называемых «слабыми учениками») обучаются для решения одной и той же проблемы и объединяются для получения лучших результатов. Основная гипотеза состоит в том, что при правильном сочетании слабых моделей мы можем получить более точные и/или надежные модели.

Один слабый ученик

В машинном обучении, независимо от того, сталкиваемся ли мы с проблемой классификации или регрессии, выбор модели чрезвычайно важен, чтобы иметь какие-либо шансы получить хорошие результаты. Этот выбор может зависеть от многих переменных задачи: количества данных, размерности пространства, гипотезы распределения…

Слабое ***смещение (bias)*** и ***разброс (variance)*** модели, хотя они чаще всего изменяются в противоположных направлениях, являются двумя наиболее фундаментальными особенностями, ожидаемыми для модели. Действительно, чтобы иметь возможность «решить» проблему, мы хотим, чтобы в нашей модели было достаточно степеней свободы для разрешения базовой сложности данных, с которыми мы работаем, но мы также хотим, чтобы у нее было не слишком много степеней свободы, чтобы избежать ее высокого разброса и быть более устойчивой. Это хорошо известный **компромисс между смещением и разбросом**.



Иллюстрация компромисса между смещением и разбросом

В ансамблевой теории обучения мы вводим понятия ***слабых учеников*** (или ***базовых моделей***), которых можно использовать в качестве строительных блоков для проектирования более сложных моделей путем объединения нескольких из них. В большинстве случаев эти базовые модели работают сами по себе не так хорошо в связи с тем, что они имеют высокое смещение (например, модели с низкой степенью свободы), либо с тем, что имеют слишком большой разброс, чтобы быть устойчивыми (например, модели с высокой степенью свободы). Тогда идея ансамблевых методов состоит в том, чтобы попытаться уменьшить смещение и/или разброс таких слабых учеников, объединяя несколько из них вместе, чтобы создать ***сильного ученика*** (или ***модель ансамбля***), который достигает лучших результатов.

Объединение слабых учеников

Чтобы реализовать ансамблевый метод, нам сначала нужно отобрать наших слабых учеников для агрегирования. В основном (в том числе в хорошо известных методах бэггинга и бустинга) используется единственный базовый алгоритм обучения, так что у нас есть однородные слабые ученики, которые обучаются по-разному. Получаемая нами модель ансамбля называется «***однородной***». Тем не менее, существуют также некоторые методы, которые используют различные типы базовых алгоритмов обучения: некоторые ***разнородные слабые ученики*** затем объединяются в **«*разнородную ансамблевую модель*»**.

* **Бэггинг.**В этом случае часто рассматривают однородных слабых учеников, обучают их параллельно и независимо друг от друга, а затем объединяют их, следуя некоторому детерминированному процессу усреднения.

Слабых учеников можно объединить, чтобы получить модель с лучшими показателями. Способ объединения базовых моделей должен быть адаптирован к их типам. Модели с низким смещением и высоким разбросом следует объединять таким образом, чтобы сделать сильную модель более устойчивой, тогда как модели с низким разбросом и высоким смещением лучше объединять таким образом, чтобы сделать ансамблевую модель менее смещенной.

**Сфокусируем внимание на бэггинге**

В ***параллельных методах*** мы рассматриваем разных учеников независимо друг от друга друга и, таким образом, можно обучать их одновременно. Наиболее известным из таких подходом является «***бэггинг***» (от «*bootstrap aggregation*»), целью которого является создание ансамблевой модели, которая является **более надежной,** чем отдельные модели, ее составляющие.

Бутстрэп

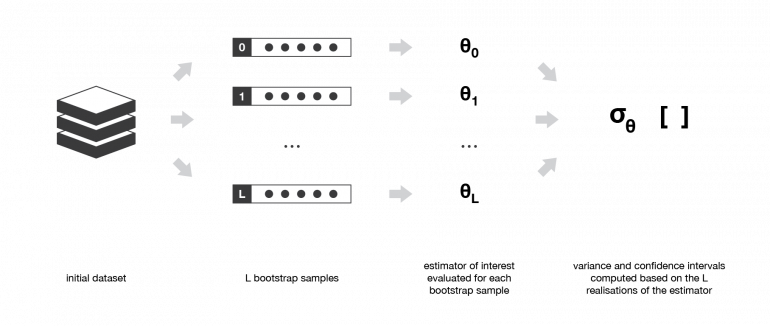
Давайте начнем с определения бутстрэпа. Этот статистический метод заключается в генерации выборок размера B (так называемых бутстрэп выборок) из исходного датасета размера N путем случайного выбора элементов с повторениями в каждом из наблюдений B.



Иллюстрация процесса бустрэпа

При некоторых допущениях эти выборки имеют довольно хорошие статистические свойства: в первом приближении их можно рассматривать как взятые непосредственно из истинного базового (и часто неизвестного) распределения данных, так и независимо друг от друга. Таким образом, их можно рассматривать как репрезентативные и независимые выборки истинного распределения данных (почти идентичные выборки). Гипотеза, которая должна быть проверена, чтобы сделать это приближение действительным, имеет две стороны. Во-первых, размер N исходного датасета должен быть достаточно большим, чтобы охватить большую часть сложности базового распределения, чтобы выборка из датасета была хорошим приближением к выборке из реального распределения **(репрезентативность)**. Во-вторых, размер датасета N должен быть достаточно большим по сравнению с размером бутстрэп выборок B, чтобы выборки не слишком сильно коррелировали **(независимость)**. Обратите внимание, что в дальнейшем мы иногда будем ссылаться на эти свойства (репрезентативность и независимость) бутстрэп выборок: читатель всегда должен помнить, что **это только приближение**.

Бутстрэп выборки часто используются, например, для оценки разброса или доверительных интервалов статистических оценок. По определению статистическая оценка является функцией некоторых наблюдений и, следовательно, случайной величины с разбросом, полученным из этих наблюдений. Чтобы оценить разброс такой оценки, нам нужно оценить его на нескольких независимых выборках, взятых из интересующего распределения. В большинстве случаев рассмотрение действительно независимых выборок потребовало бы слишком большого количества данных по сравнению с реально доступным количеством. Затем мы можем использовать бутстрэп, чтобы сгенерировать несколько бутстрэп выборок, которые можно рассматривать как «почти репрезентативные» и «почти независимые» (почти «независимые одинаково распределенные выборки»). Эти примеры бутстрэп выборок позволят нам аппроксимировать разброс оценки, оценивая его значение для каждой из них.



Бутстрэп часто используется для оценки разброса или доверительного интервала некоторых статистических оценок

**Бэггинг**

При обучении модели, независимо от того, имеем ли мы дело с  
проблемой классификации или регрессии, мы получаем функцию, которая принимает входные данные, возвращает выходные данные и определяется в отношении обучающего датасета. Из-за теоретического разброса обучающего датасета (мы напоминаем, что датасет является наблюдаемой выборкой, исходящей из истинно неизвестного базового распределения), подобранная модель также подвержена изменчивости: **если бы наблюдался другой датасет, мы получили бы другую модель**.

Идея бэггинга в таком случае проста: мы хотим подобрать несколько независимых моделей и «усреднить» их прогнозы, чтобы получить модель с меньшим разбросом. Однако на практике мы не можем подобрать полностью независимые модели, потому что для этого потребуется слишком много данных. Таким образом, мы полагаемся на хорошие «приблизительные свойства» бутстрэп выборок (репрезентативность и независимость) для подбора моделей, которые практически независимы.

Сначала мы генерируем несколько бутстрэп выборок так, чтобы каждая новая бутстрэп выборка выполняла роль (почти) еще одного независимого датасета, взятого из истинного распределения. Затем мы можем **обучить слабого ученика для каждой из этих выборок и, наконец, агрегировать их так, чтобы мы как бы «усреднили» их результаты** и, таким образом, получили модель ансамбля с разбросом меньшим, чем ее отдельные компоненты.

Итак, предположим, что у нас есть L бутстрап выборок (аппроксимации L независимых датасетов) размера B. Это обозначается:



Мы можем обучить L почти независимых слабых учеников (по одному на каждый датасет): 

А затем объединим их некоторым процессом усреднения, чтобы получить модель ансамбля с меньшим разбросом. Например, мы можем определить нашу сильную модель так, чтобы



Существует несколько возможных способов объединить несколько моделей, обученных параллельно. Для задачи регрессии выходные данные отдельных моделей могут быть буквально усреднены для получения выходных данных модели ансамбля. Для задачи классификации класс, предсказываемый каждой моделью, можно рассматривать как голос, а класс, который получает большинство голосов, является ответом модели ансамбля (это называется **мажоритарным голосованием**). Что касается задачи классификации, мы также можем рассмотреть вероятности каждого класса, предсказываемые всеми моделями, усреднить эти вероятности и сохранить класс с самой высокой средней вероятностью (это называется **мягким голосованием**). Средние значения или голоса могут быть простыми или взвешенными, если будут использоваться любые соответствующие им веса.

**Случайные леса**

***Деревья решений*** являются очень популярными базовыми моделями для ансамблевых методов. Сильные ученики, состоящие из нескольких деревьев решений, можно назвать «лесами». Деревья, составляющие лес, могут быть выбраны либо неглубокими (глубиной в несколько узлов), либо глубокими (глубиной в множество узлов, если не в полную глубину со всеми листьями). Неглубокие деревья имеют меньший разброс, но более высокое смещение, и тогда для них лучшим выбором станут **последовательные методы**, которые мы опишем позже. Глубокие деревья, с другой стороны, имеют низкое смещение, но высокий разброс и, таким образом, являются подходящим выбором для бэггинга, который в основном направлен на уменьшение разброса.

**Случайный лес**представляет собой метод бэггинга, где глубокие деревья, обученные на бутстрап выборках, объединяются для получения результата с более низким разбросом. Тем не менее, случайные леса также используют другой прием, чтобы несколько обученных деревьев были менее коррелированными друг с другом: при построении каждого дерева вместо выбора всех признаков из датасета для генерации бутстрэпа мы выбираем и сохраняем только случайное их подмножество для построения дерева (обычно одинаковое для всех бутстрэп выборок).

Выборка по признакам действительно приводит к тому, что все деревья не смотрят на одну и ту же информацию для принятия своих решений и, таким образом, уменьшают корреляцию между различными возвращаемыми выходными данными. Другое преимущество выборки по признакам заключается в том, что **она делает процесс принятия решений более устойчивым к отсутствующим данным**: значения наблюдения (из обучающего датасета или нет) с отсутствующими данными можно восстанавливать с помощью регрессии или классификации на основе деревьев, которые учитывают только те признаки, где данные не отсутствуют. Таким образом, алгоритм случайного леса сочетает в себе концепции бэггинга и выбора подпространства случайных объектов для создания более устойчивых моделей.