Бэггинг_2

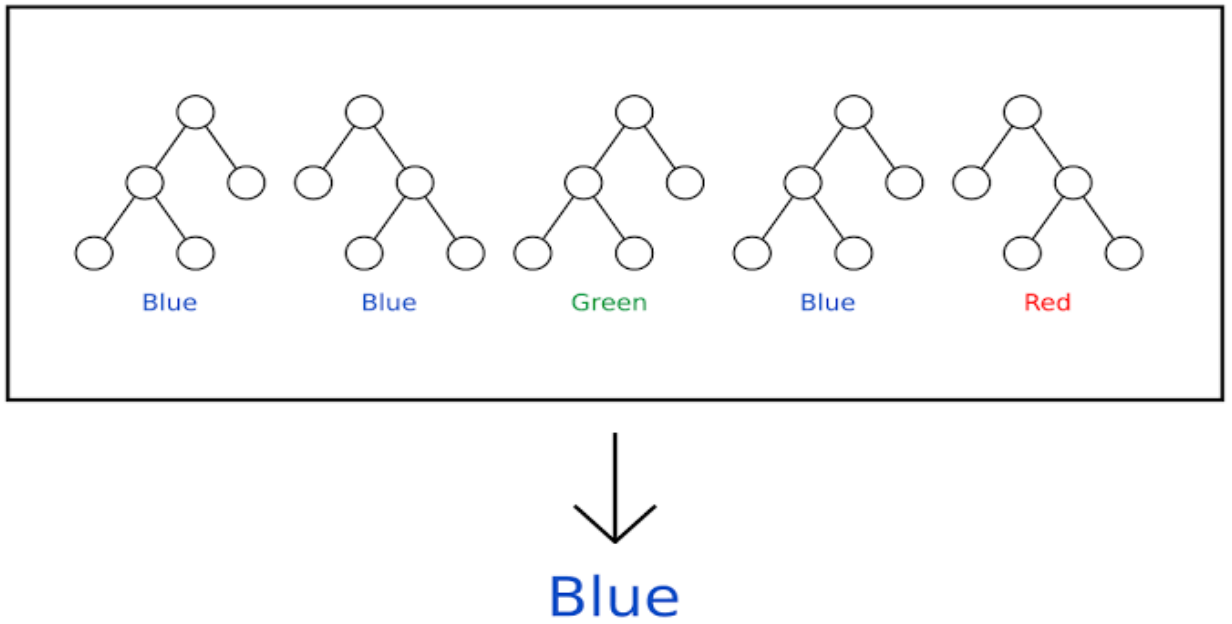
ФИНАНСОВЫЙ УНИВЕРСИТЕТ

ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Ансамбли

Под обучением ансамбля моделей понимается процедура обучения конечного набора базовых классификаторов, результаты прогнозирования которых, затем объединяются и формируют прогноз агрегированного классификатора.
Целью создания ансамбля моделей является повышение точности прогноза агрегированного классификатора по сравнению с точностью прогнозирования каждого индивидуального базового классификатора

Bagging



Интуитивно понятно, что комбинирование базовых классификаторов даст более точный результат, чем каждый индивидуальный классификатор, если базовые классификаторы с одной стороны достаточно точны, а с другой – если они дают разные результаты, т.е. ошибаются на разных множествах. Проиллюстрируем это следующим искусственным примером:
Пусть у нас есть 25 базовых классификатора, при этом

- вероятность ошибки каждого классификатора 35%
- результаты прогноза классификаторов независимы

Ансамбли

Вероятность ошибки объединенного классификатора, прогнозирующего то значение класса, за которое «проголосовало» большинство - более 12 базовых классификаторов, согласно биномиальному закону распределения будет равно 0,06.
Т.е. ошибка ансамбля будет всего 6% по сравнению с 35% ошибки каждого базового классификатора. Рассмотренный пример является нереалистичным, так как на практике результаты прогнозирования базовых классификаторов сильно коррелированы, потому что наибольшая вероятность ошибки любого классификатора происходит на границе классов, а наименьшая – вдали от границ.

$$P(\text{больше 12 классификаторов ошиблись}) = \sum_{i=13}^{25} \binom{25}{i} 0,35^{i} \cdot 0,65^{25-i} = 0,06$$

Биномиальный закон распределения.

Пусть заданы числа $n \in N$ и p $(0 \le p \le 1)$. Тогда каждому целому числу из промежутка $[0; n]$ можно поставить в соответствие вероятность, рассчитанную по формуле Бернулли. Получим закон распределения случайной величины (назовём её β)

β	0	…	k	…	n
P	…	…	$C_n^k p^k (1-p)^{n-k}$	…	…

Будем говорить, что случайная величина β распределена по закону Бернулли. Такой случайной величиной является частота появления события A в n повторных независимых испытаниях, если в каждом испытании событие A происходит с вероятностью p.

Ансамбли

Хорошим примером ансамблей считается теорема Кондорсе «о жюри присяжных» (1784). Если каждый член жюри присяжных имеет независимое мнение, и если вероятность правильного решения члена жюри больше 0.5, то тогда вероятность правильного решения присяжных в целом возрастает с увеличением количества членов жюри и стремится к единице. Если же вероятность быть правым у каждого из членов жюри меньше 0.5, то вероятность принятия правильного решения присяжными в целом монотонно уменьшается и стремится к нулю с увеличением количества присяжных.

N — количество присяжных

p — вероятность правильного решения присяжного

μ — вероятность правильного решения всего жюри

m — минимальное большинство членов жюри, $m = \mathrm{floor}(N/2) + 1$

C_N^i — число сочетаний из N по i

$$\mu = \sum_{i=m}^{N} C_N^i p^i (1-p)^{N-i}$$

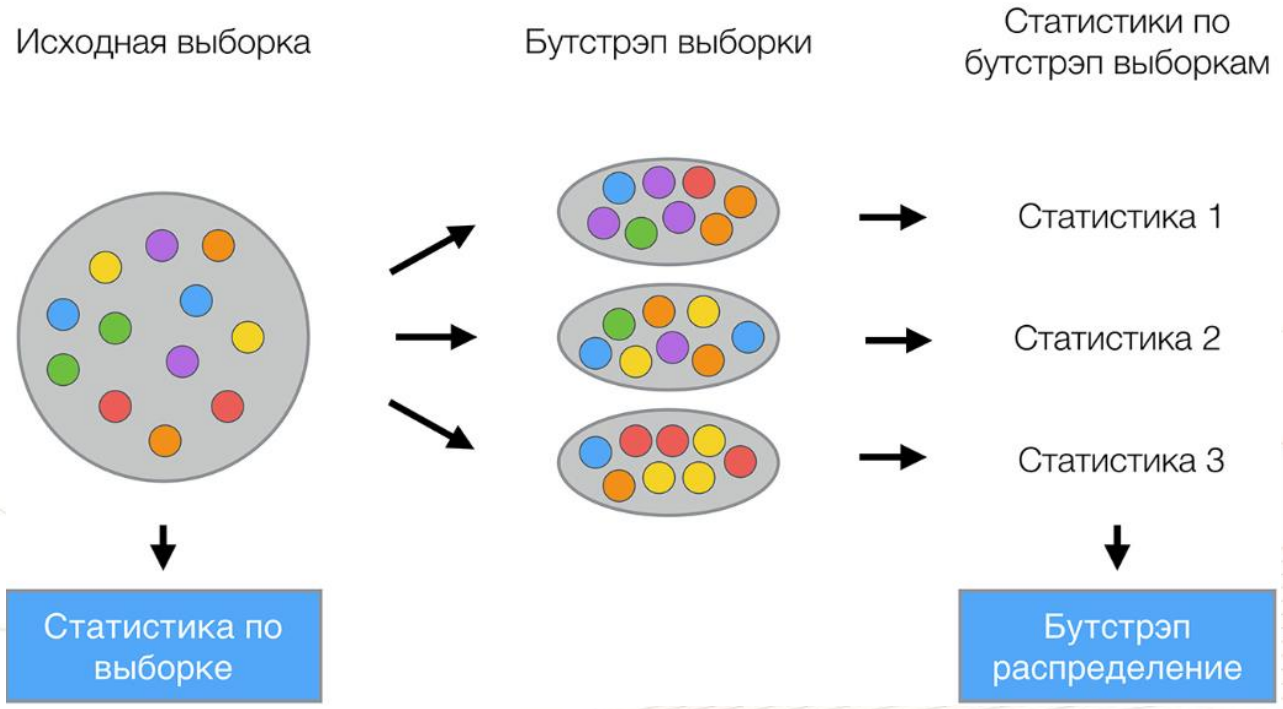
Если $p > 0.5$, то $\mu > p$

Если $N \to \infty$, то $\mu \to 1$

Бэггинг

	Date	Title	Budget	Domestic/TotalGross	Director	Rating	Runtime
0	2013-11-22	The Hunger Games: Catching Fire	130000000	424668047	Francis Lawrence	PG-13	146
1	2013-05-03	Iron Man 3	200000000	408813994	Shane Black	PG-13	129
2	2013-11-22	Frozen	150000000	400738009	Chris Buck,Jennifer Lee	PG	108
3	2013-07-03	Despicable Me 2	76000000	368061265	Pierre Coffin,Chris Renaud	PG	98
4	2013-06-14	Man of Steel	225000000	291045518	Zack Snyder	PG-13	143
5	2013-10-04	Gravity	100000000	274092705	Alfonso Cuaron	PG-13	91
6	2013-06-21	Monsters University	NaN	268492764	Dan Scanlon	G	107
7	2013-12-13	The Hobbit: The Desolation of Smaug	NaN	258366855	Peter Jackson	PG-13	161
8	2013-05-24	Fast & Furious 6	160000000	238679850	Justin Lin	PG-13	130
9	2013-03-08	Oz The Great and Powerful	215000000	234911825	Sam Raimi	PG	127
10	2013-05-16	Star Trek Into Darkness	190000000	228778661	J.J. Abrams	PG-13	133
11	2013-11-08	Thor: The Dark World	170000000	206362140	Alan Taylor	PG-13	120
12	2013-06-21	World War Z	190000000	202359711	Marc Forster	PG-13	116
13	2013-03-22	The Croods	135000000	187168425	Kirk De Micco,Chris Sanders	PG	98
14	2013-06-28	The Heat	43000000	159582186	Paul Feig	R	117
15	2013-08-07	We're the Millers	37000000	150394119	Rawson Marshall Thurber	R	110
16	2013-12-13	American Hustle	40000000	150117807	David O. Russell	R	138
17	2013-05-10	The Great Gatsby	105000000	144840419	Baz Luhrmann	PG-13	143

Bagging (от Bootstrap aggregation) — это один из первых и самых простых видов ансамблей. Бэггинг основан на статистическом методе бутстрэпа, который позволяет оценивать многие статистики сложных распределений.

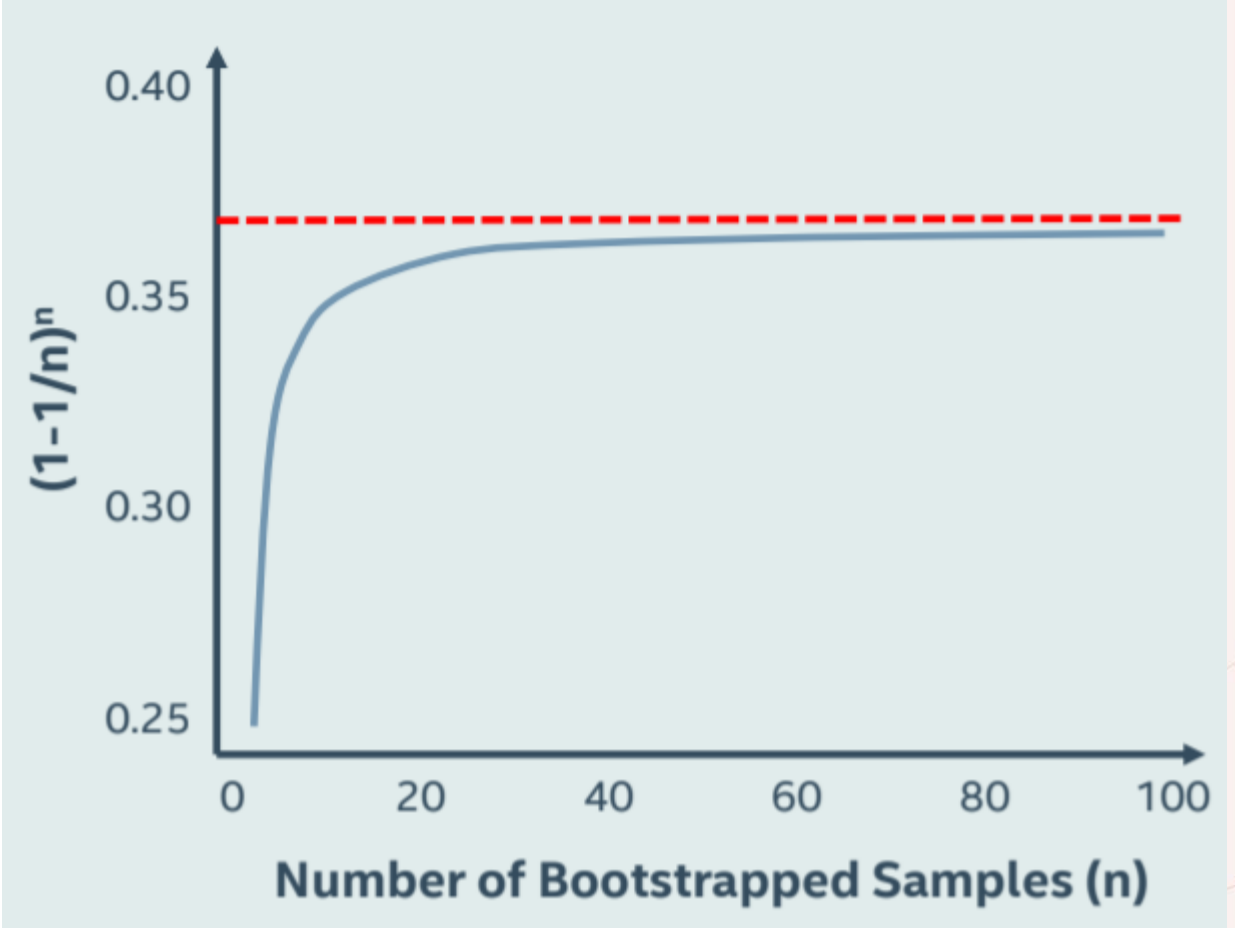


Улучшение: используйте много деревьев

- По заданному набору данных создайте *n* бутстрэп выборки
- Для данной записи *x*

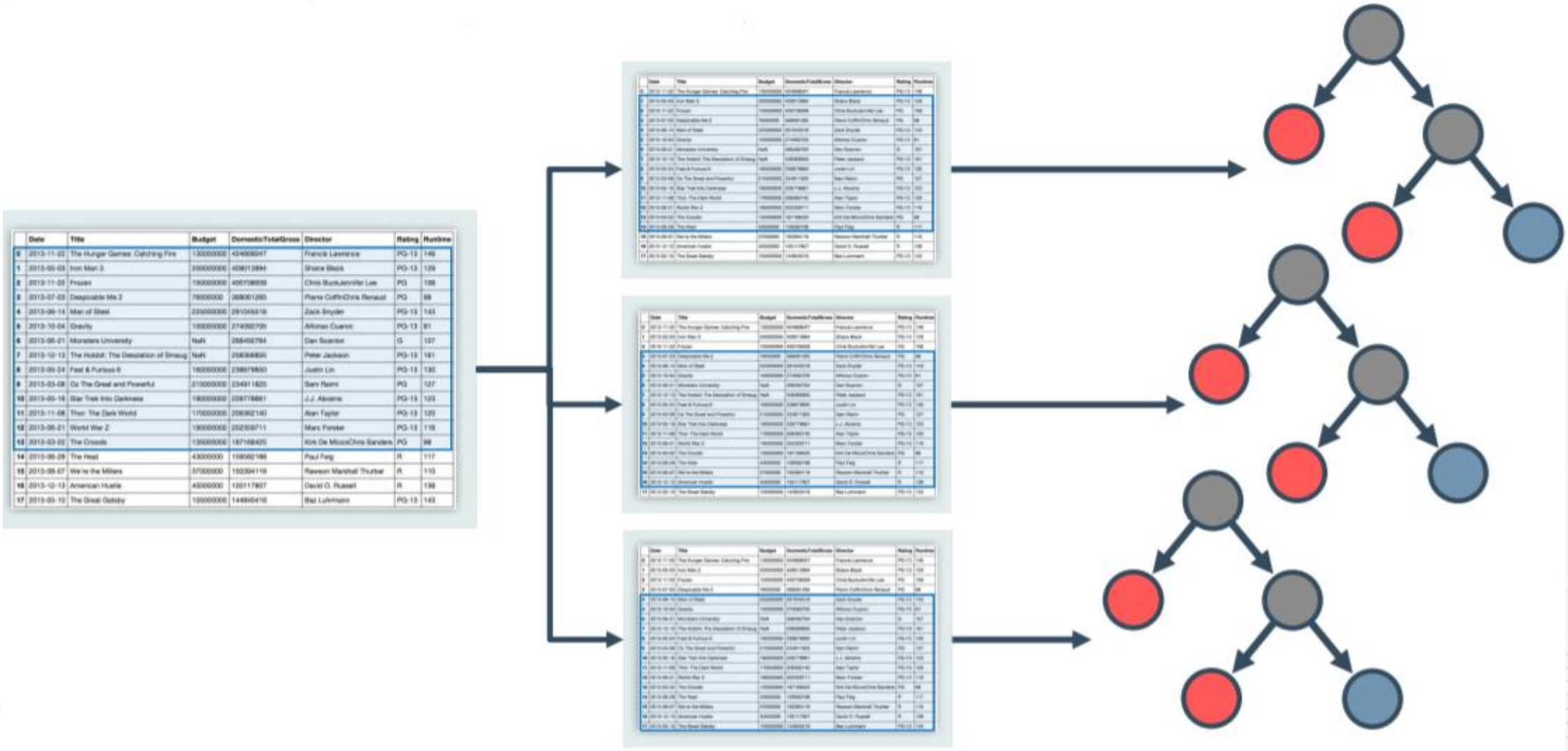
$$P(rec\ x\ not\ selected) = (1 - {}^{1}/_{n})^{n}$$

- Каждый образец начальной загрузки содержит примерно 2/3 записей



Улучшение: используйте много деревьев

Вырастить дерево решений для каждой бутстрэп выборки

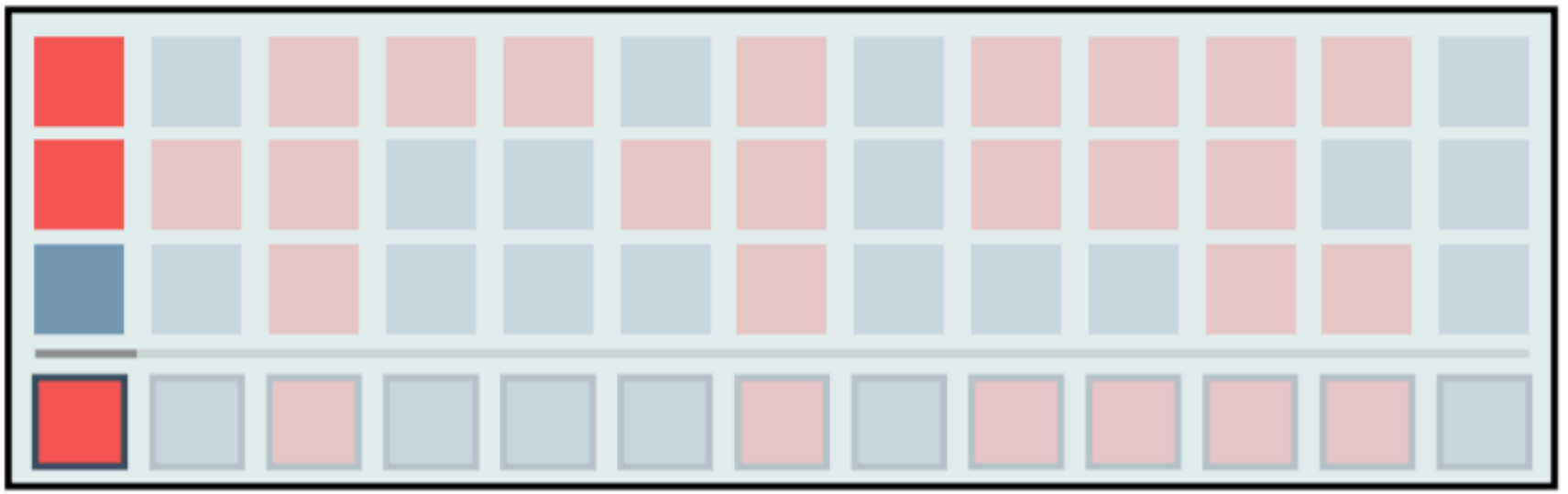


Агрегация результатов

Деревья голосуют и усредняют результат для каждой примера

Наблюдение (пример обучающей выборки)

Голосование для формирования единого классификатора

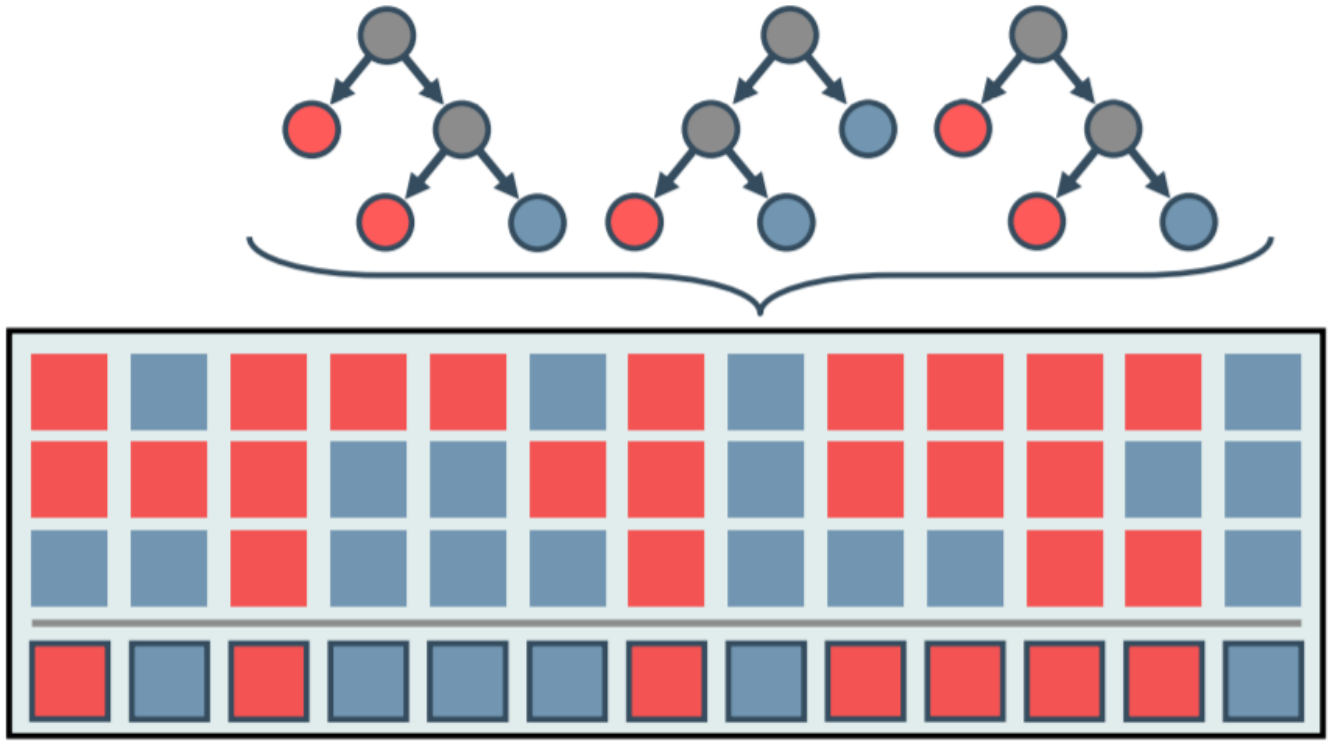


Результат

Агрегация результатов

Деревья голосуют и усредняют результат для каждой примера

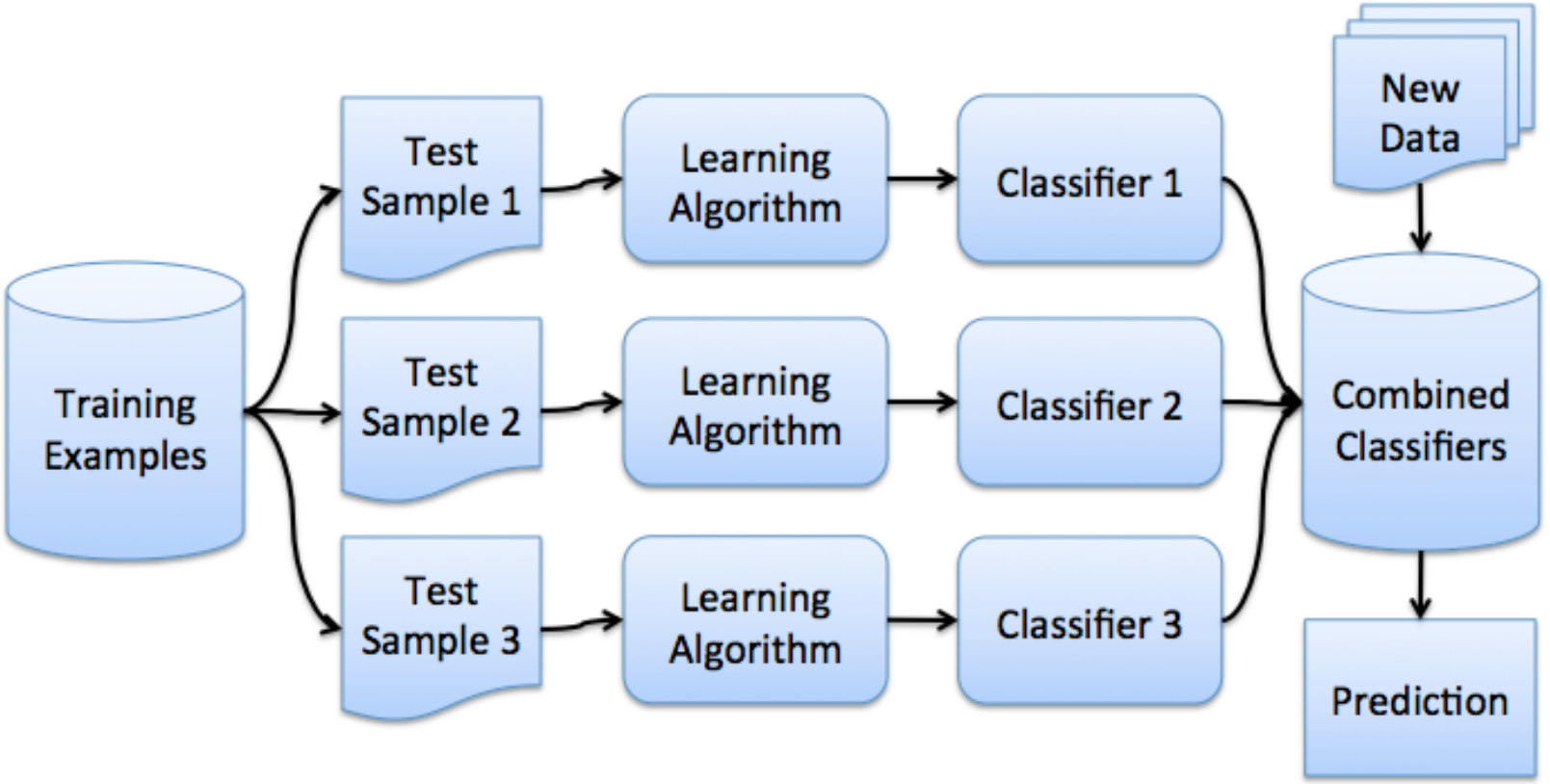
Голосование для формирования единого классификатора



- Улучшение точности классификации (регрессии) бэггинга увеличивается с увеличением количества деревьев
- Максимальное улучшение обычно достигает ~ 50 деревьев

Бэггинг

Пусть имеется обучающая выборка. С помощью бутстрэпа сгенерируем из неё выборки. Теперь на каждой выборке обучим свой классификатор. Итоговый классификатор будет усреднять ответы всех этих алгоритмов (в случае классификации это соответствует голосованию)



Бэггинг

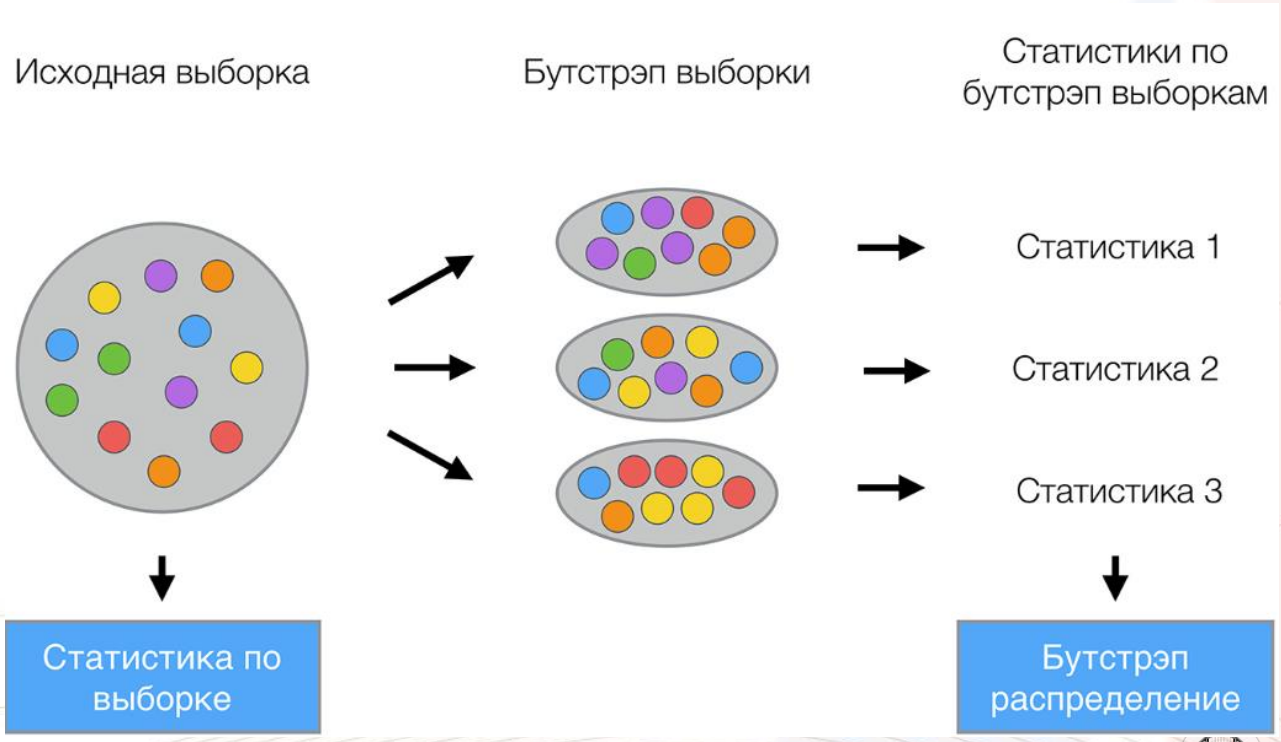
Если задан стандартный тренировочный набор ***D*** размера ***n***, бэггинг образует ***m*** новых тренировочных наборов D_i , каждый размером ***n'***, путём выборки из ***D*** *равномерно и с возвратом*.

При сэмплинге с возвратом некоторые наблюдения могут быть повторены в каждой D_i. Если $n'=n$, то для больших n ожидается, что множество D_i имеет $(1 - 1/e) \approx 63{,}2\%$ долю уникальных экземпляров из ***D***, остальные будут повторениями.

Этот вид сэмплинга известен как бутстрэп-сэмплинг.

Эти ***m*** моделей сглаживается с помощью вышеупомянутых ***m*** бутстрэп-выборок и комбинируются путём усреднения (для регрессии) или голосования (для классификации).

Bagging (от Bootstrap aggregation) — это один из первых и самых простых видов ансамблей. Бэггинг основан на статистическом методе бутстрэпа, который позволяет оценивать многие статистики сложных распределений.



Преимущества перед отдельными деревьями

- Бэггинг позволяет снизить дисперсию (variance) обучаемого классификатора, уменьшая величину, на сколько ошибка будет отличаться, если обучать модель на разных наборах данных, или другими словами, предотвращает переобучение. Эффективность бэггинга достигается благодаря тому, что базовые алгоритмы, обученные по различным подвыборкам, получаются достаточно различными, и их ошибки взаимно компенсируются при голосовании, а также за счёт того, что объекты-выбросы могут не попадать в некоторые обучающие подвыборки.
- Меньшая изменчивость, чем деревья решений
- Можно выращивать деревья параллельно

BaggingClassifier: синтаксис

Импортируем класс, содержащий метод классификации

```
from sklearn.ensemble import BaggingClassifier
```

Создадим экземпляр класса

```
BC = BaggingClassifier(n_estimators=50)
```

Обучим модель на обучающем выборке, а затем прогнозируем ожидаемое значение на тестовой выборке

```
BC = BC.fit(X_train, y_train)
y_predict = BC.predict(X_test)
```

Используйте BaggingRegressor для регрессии.
Настройте параметры перекрестной проверки