

Случайный лес. Ансамблирование

Алексей Ярошенко



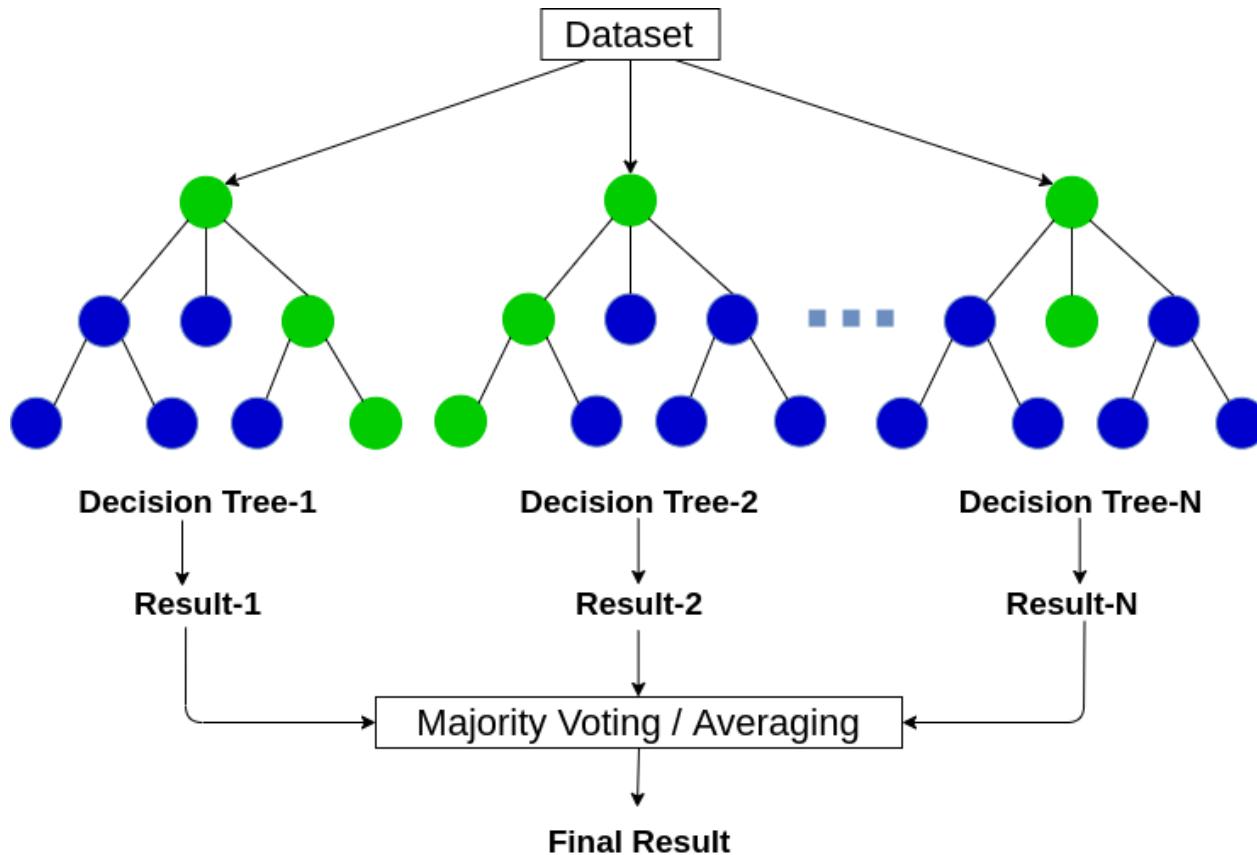
Проверить,
включена ли
запись лекции



А теперь соберем
случайный лес

Случайный лес

Обучили N деревьев. Предсказания усреднили. Расходимся



Одно дерево

0.572

```
tree = DecisionTreeClassifier(random_state=18)
tree.fit(X_train, y_train)
y_proba = tree.predict_proba(X_test)[:, 1]
print(f'Accuracy: {accuracy_score(y_test, y_proba > 0.5)}')
```

Accuracy: 0.572736520854527

Много деревьев

0.572

```
n_estimators = 100
forest = []

for i in range(n_estimators):
    tree = DecisionTreeClassifier(random_state=18)
    tree.fit(X_train, y_train)
    forest.append(tree)

preds = []
for tree in forest:
    y_pred = tree.predict_proba(X_test)[:, 1]
    preds.append(y_pred)

# берем среднее между деревьями
y_proba_forest = np.mean(preds, axis=0)

print(f'Accuracy: {accuracy_score(y_test, y_proba_forest > 0.5)}')
```

Accuracy: 0.572736520854527

Что не так?

Bagging

Добавим случайности

0.572 → 0.668. Что мы здесь сделали?

```
n_estimators = 100
forest = []
np.random.seed(18)

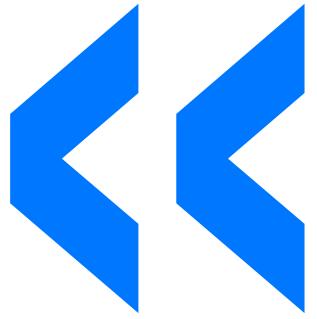
for i in range(n_estimators):
    tree = DecisionTreeClassifier(random_state=18)
    # =====
    take_idx = np.random.choice(a=len(y_train), size=len(y_train), replace=True)
    tree.fit(X_train[take_idx], y_train[take_idx])
    # =====
    forest.append(tree)

preds = []
for tree in forest:
    y_pred = tree.predict_proba(X_test)[:, 1]
    preds.append(y_pred)

# берем среднее между деревьями
y_proba_forest = np.mean(preds, axis=0)

print(f'Accuracy: {accuracy_score(y_test, y_proba_forest > 0.5)}')
```

Accuracy: 0.6683621566632757



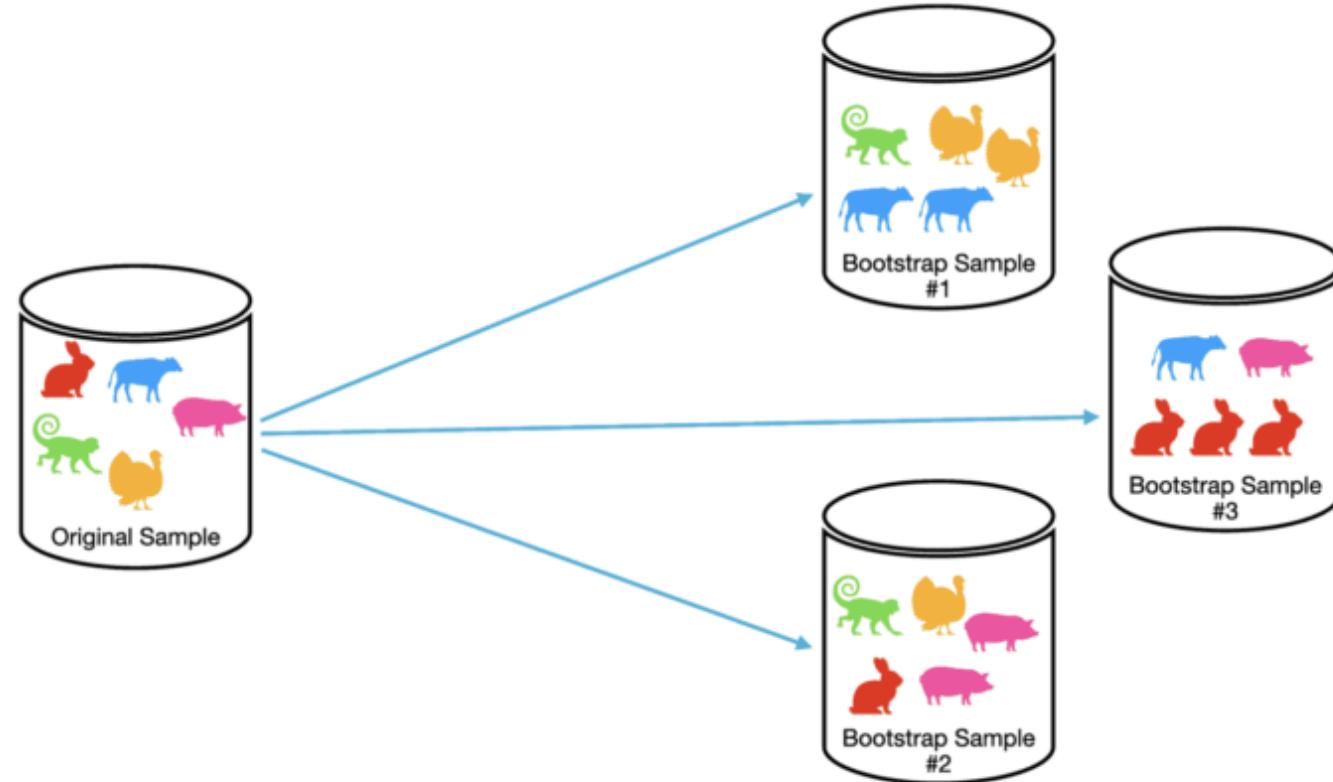
Вопрос на собеседовании

Что вообще такое Bootstrap и зачем он нужен?

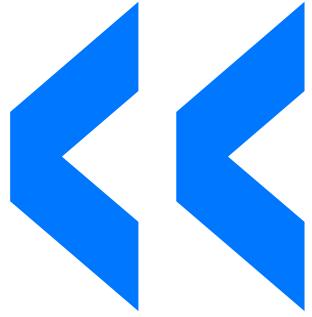
Bootstrap sampling

Выборки:

- Того же размера
- С возвращениями

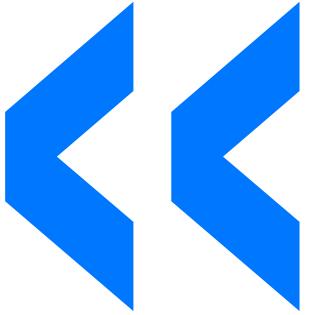


```
# =====
take_idx = np.random.choice(a=len(y_train), size=len(y_train), replace=True)
tree.fit(X_train[take_idx], y_train[take_idx])
# =====
```



Вопрос на собеседовании

А зачем выборка того же размера, если можно взять просто сэмпл?



Вопрос на собеседовании

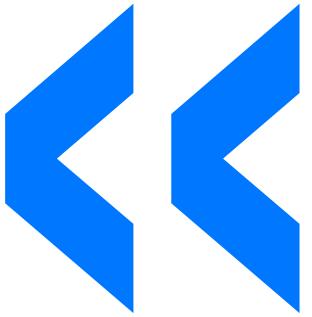
Когда использовать Bootstrap? Какие преимущества над другими методами? И какими другими?

Что мы сделали

- Bootstrap подвыборки
- Усреднили результаты по деревьям (агрегировали)

Bootstrap + Aggregation = Bagging

Out Of Bag score (OOB)



Вопрос на собеседовании

Посчитайте долю объектов, не попавших в обучение при бэггинге

1-й шаг:

$$P_1 = \frac{1}{N} - \text{вероятность взять объект при выборе 1 объекта}$$

Вероятность НЕ взять объект в обучение

$P_1 = \frac{1}{N}$ - вероятность взять объект при выборе 1 объекта

$P_0 = 1 - \frac{1}{N}$ - вероятность НЕ взять объект при выборе 1 объекта

$P_0 = \left(1 - \frac{1}{N}\right)^N$ - вероятность не взять объект в подвыборку

$\left(1 - \frac{1}{N}\right)^N \rightarrow \frac{1}{e} \approx 0.368 \approx 37\%$

Out Of Bag выборка

Не пропадать же датасету :)
Давайте на нем
валидироваться!

- инферим ОOB на тех деревьях, на которых он не учился
 - предсказания усредняем
 - считаем разные метрики

Попало в обучение дерева

Не попало в обучение дерева

Формально

$$\text{OOB} = \sum_{i=1}^l L \left(y_i, \frac{1}{\sum_{n=1}^N [x_i \notin X_n^l]} \sum_{n=1}^N [x_i \notin X_n^l] b_n(x_i) \right)$$

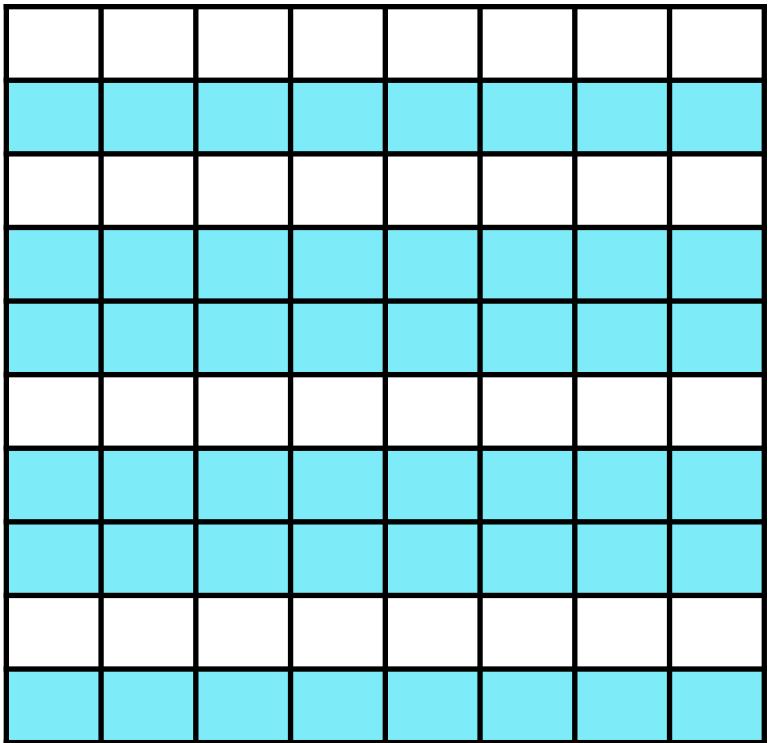
X_n^l - выборка для модели b_n

l - размер исходной выборки

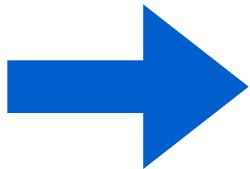
N - число моделей в ансамбле

Метод случайных подпространств (Random Subspace Method)

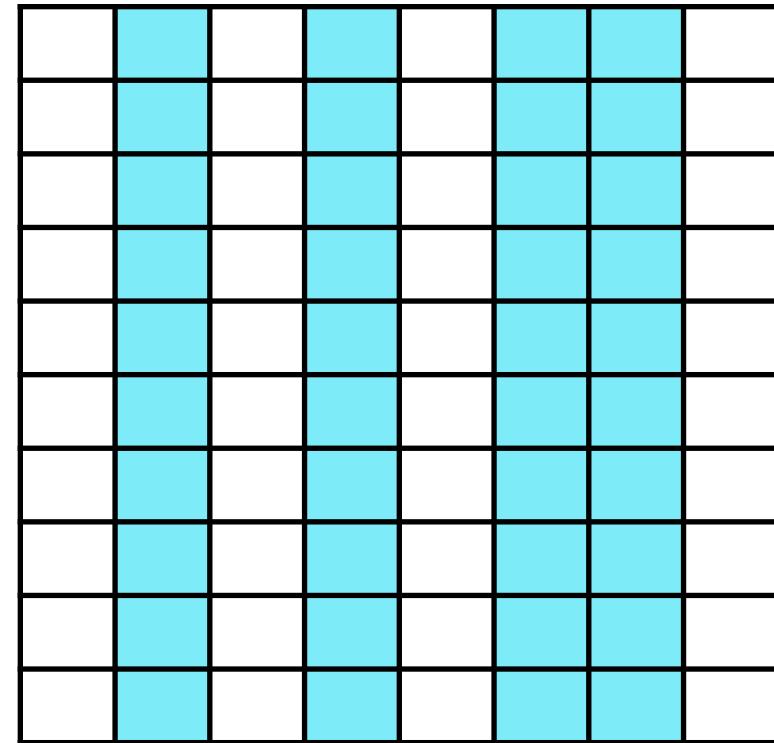
Bagging



Семплируем выборку того же размера с повторениями



Метод случайных подпространств



Берем случайную подвыборку признаков в сплит

А сколько сэмплировать признаков

Рекомендуют:

- Для регрессии: $\frac{m}{3}$
- Для классификации: \sqrt{m}

Попробуем сэмплировать признаки

0.572 → 0.585

```
n_estimators = 100
forest = []
np.random.seed(18)

for i in range(n_estimators):
    tree = DecisionTreeClassifier(random_state=18, max_features='sqrt')
    tree.fit(X_train, y_train)
    forest.append(tree)

preds = []
for tree in forest:
    y_pred = tree.predict_proba(X_test)[:, 1]
    preds.append(y_pred)

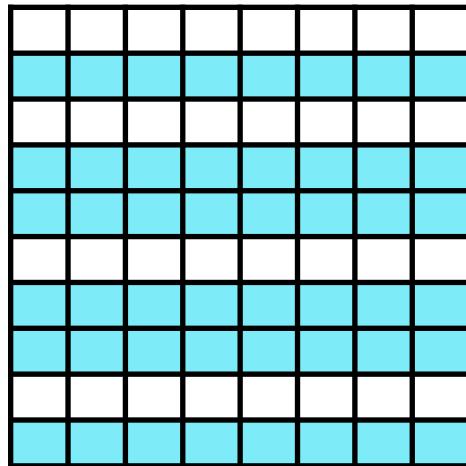
# берем среднее между деревьями
y_proba_forest = np.mean(preds, axis=0)

print(f'Accuracy: {accuracy_score(y_test, y_proba_forest > 0.5)}')
```

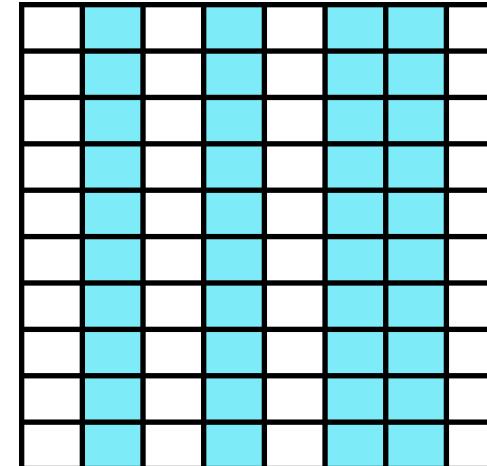
Accuracy: 0.5849440488301119

Random Forest

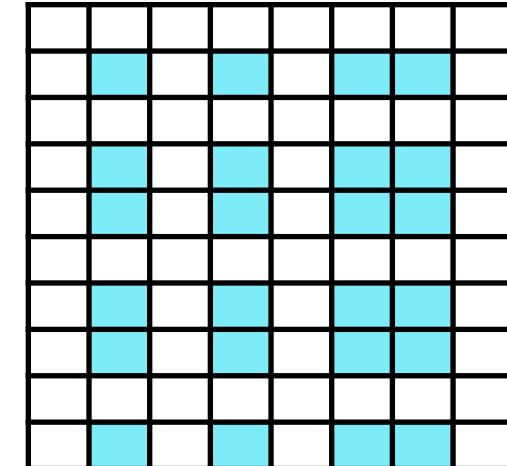
Bagging



Random Subspace



Random Forest



Для каждого дерева сэмплируем выборку того же размера с повторениями + случайный набор признаков

Соберем все в месте

0.572 -> [0.585 , 0.668] -> 0.671

```
n_estimators = 100
forest = []
np.random.seed(18)

for i in range(n_estimators):
    tree = DecisionTreeClassifier(random_state=18, max_features='sqrt')
    # =====
    take_idx = np.random.choice(a=len(y_train), size=len(y_train), replace=True)
    tree.fit(X_train[take_idx], y_train[take_idx])
    # =====
    forest.append(tree)

preds = []
for tree in forest:
    y_pred = tree.predict_proba(X_test)[:, 1]
    preds.append(y_pred)

# берем среднее между деревьями
y_proba_forest = np.mean(preds, axis=0)

print(f'Accuracy: {accuracy_score(y_test, y_proba_forest > 0.5)}')
```

Accuracy: 0.671414038657172

Почему это работает

На примере MSE, из чего состоит ошибка

$$\begin{aligned}\mathbb{E}[(y - a)^2] &= \mathbb{E}[y^2 - 2ay + a^2] = \mathbb{E}[y^2] - 2\mathbb{E}[ay] + \mathbb{E}[a^2] = \\ \mathbb{E}[y^2] - 2f\mathbb{E}[a] + \mathbb{E}[a^2] &= \mathbb{E}[y^2] - 2f\mathbb{E}[a] + \mathbb{E}[a^2] + \\ ((\mathbb{E}[y])^2 - (\mathbb{E}[y])^2) + ((\mathbb{E}[a])^2 - (\mathbb{E}[a])^2) &= \\ (\mathbb{E}[y^2] - (\mathbb{E}[y])^2) + (\mathbb{E}[a^2] - (\mathbb{E}[a])^2) + (\mathbb{E}[y])^2 - 2f\mathbb{E}[a] + (\mathbb{E}[a])^2 = \\ \mathbb{D}[y] + \mathbb{D}[a] + (\mathbb{E}[f - a])^2\end{aligned}$$

Дисперсия данных,
"неустранимая ошибка"

Смещение модели (**bias**)
Дисперсия модели (**variance**)

Bias & Variance Tradeoff

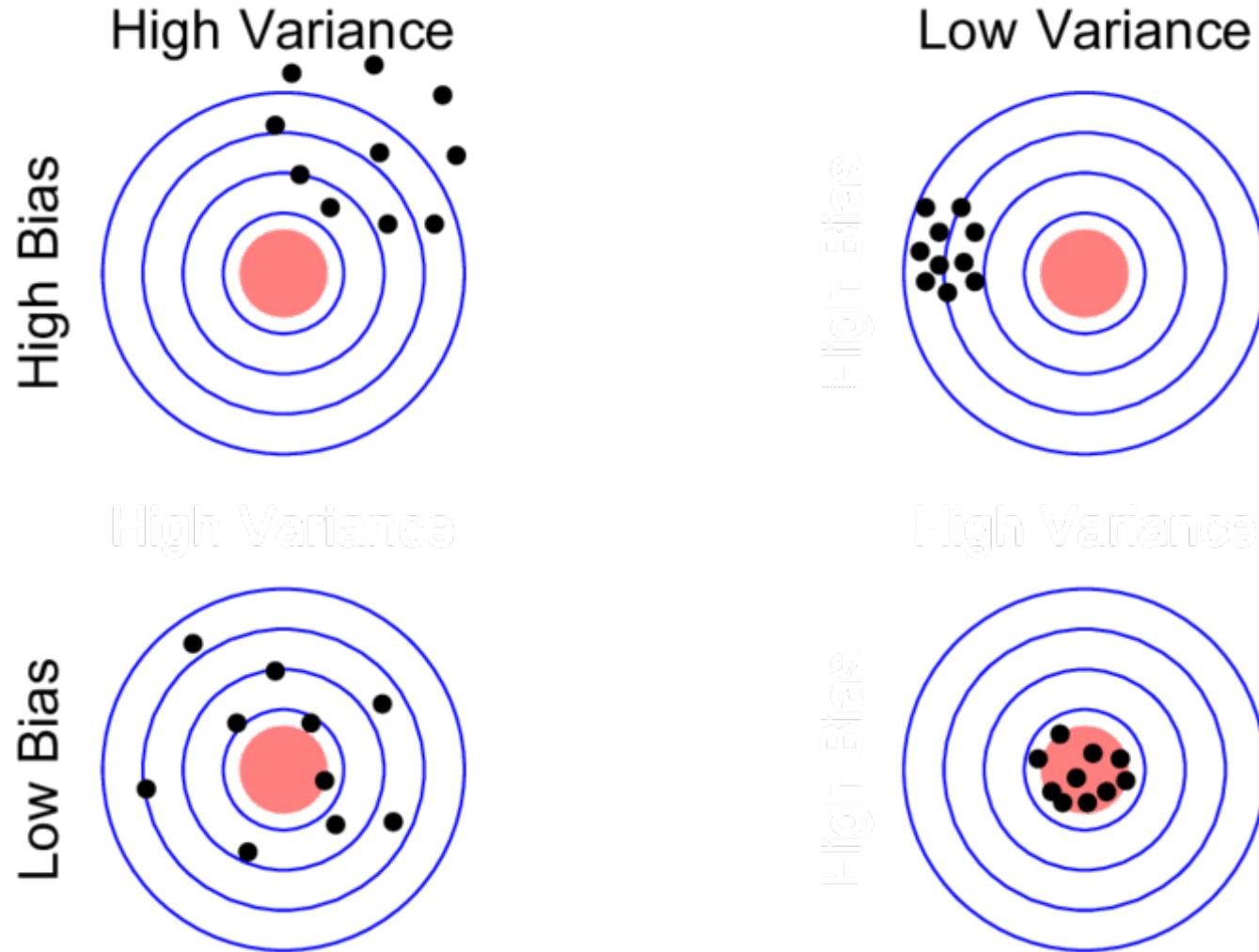
Ошибка предсказания модели состоит из 3-х частей:

1. смещение / bias
2. разборс / дисперсия / variance
3. неустранимый шум в данных σ^2

$$MSE = Bias[\hat{f}] + Var[\hat{f}] + \sigma^2$$

Если зафиксировать ошибку и шум данных, то увеличивая bias, снижаем variance и наоборот

Bias & Variance Tradeoff по простому



Глубокие деревья. Интуиция, почему низкий bias

Когда деревья глубокие, мы можем выучить обучающую выборку

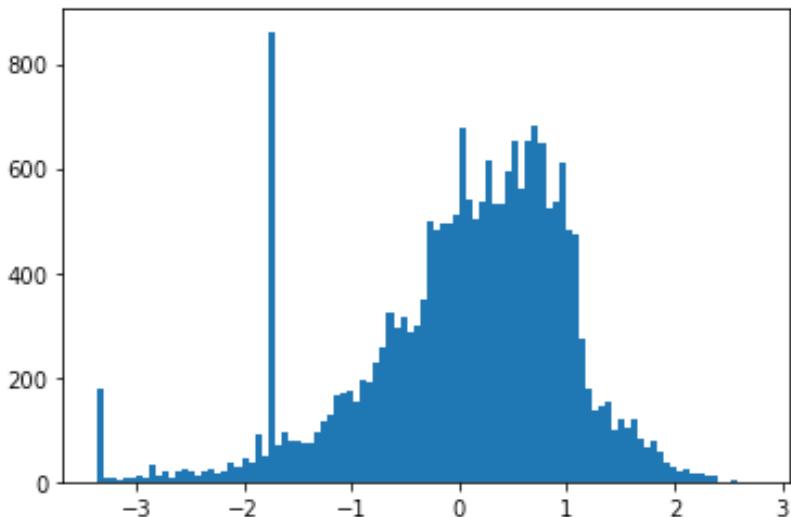
Это значит сделать $\text{bias} = 0$ на обучающей выборке

Нам нужны именно такие переобучение глубокие деревья с **низким bias и высоким variance**.

Усреднение предсказаний снижает variance.

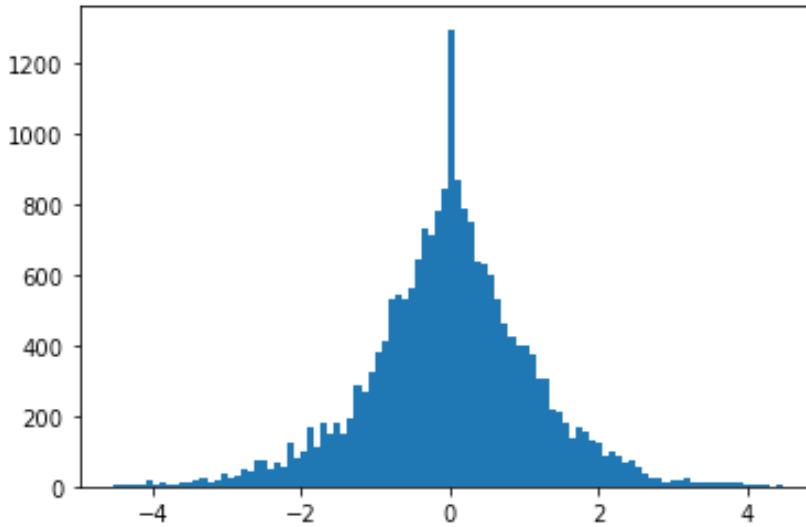
Получаем низкий bias и низкий variance

Глубокие деревья. Посмотрим, как выглядят



Дерево глубины 2:

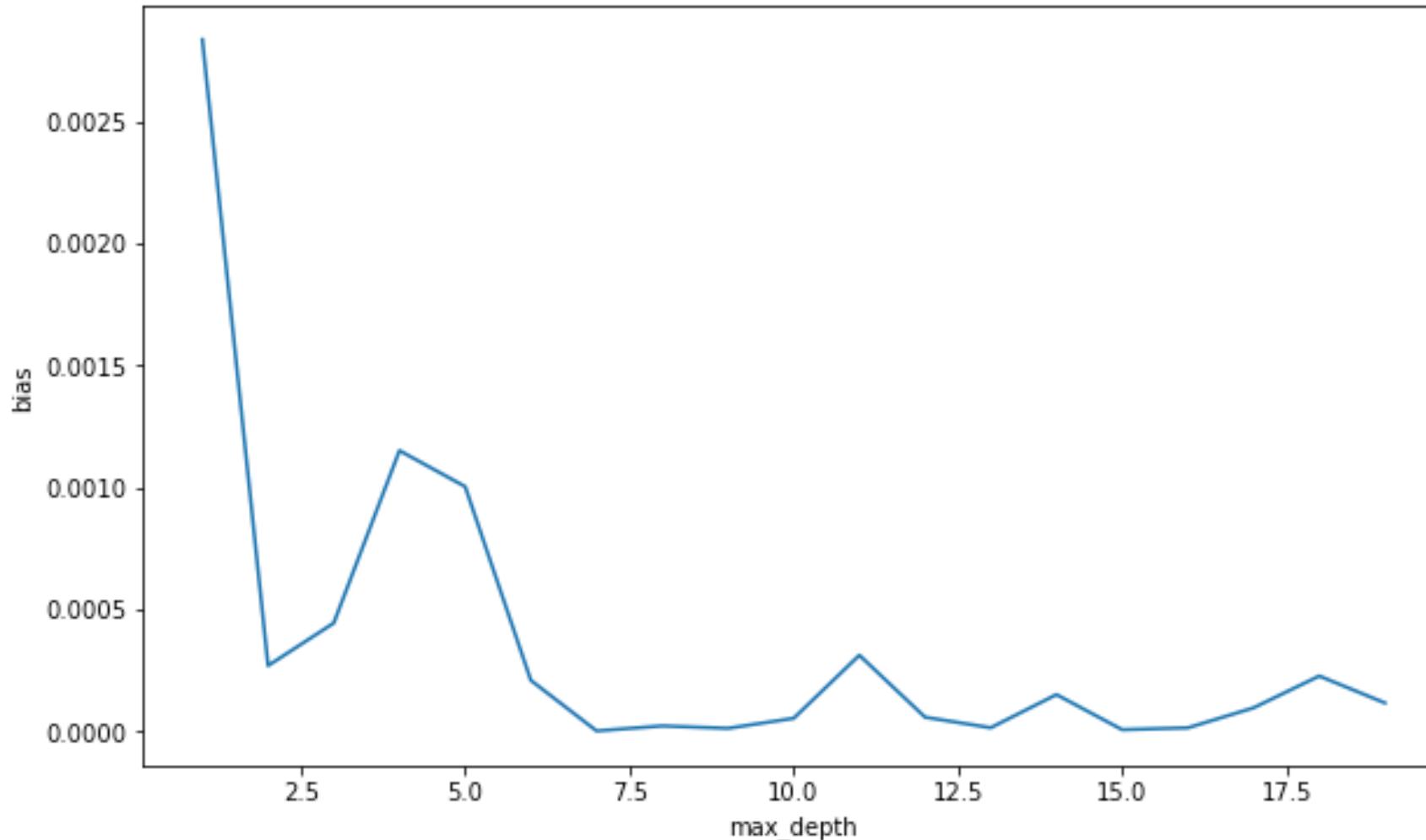
Ошибка предсказаний
смещена относительно 0.
Т.е. высокий bias



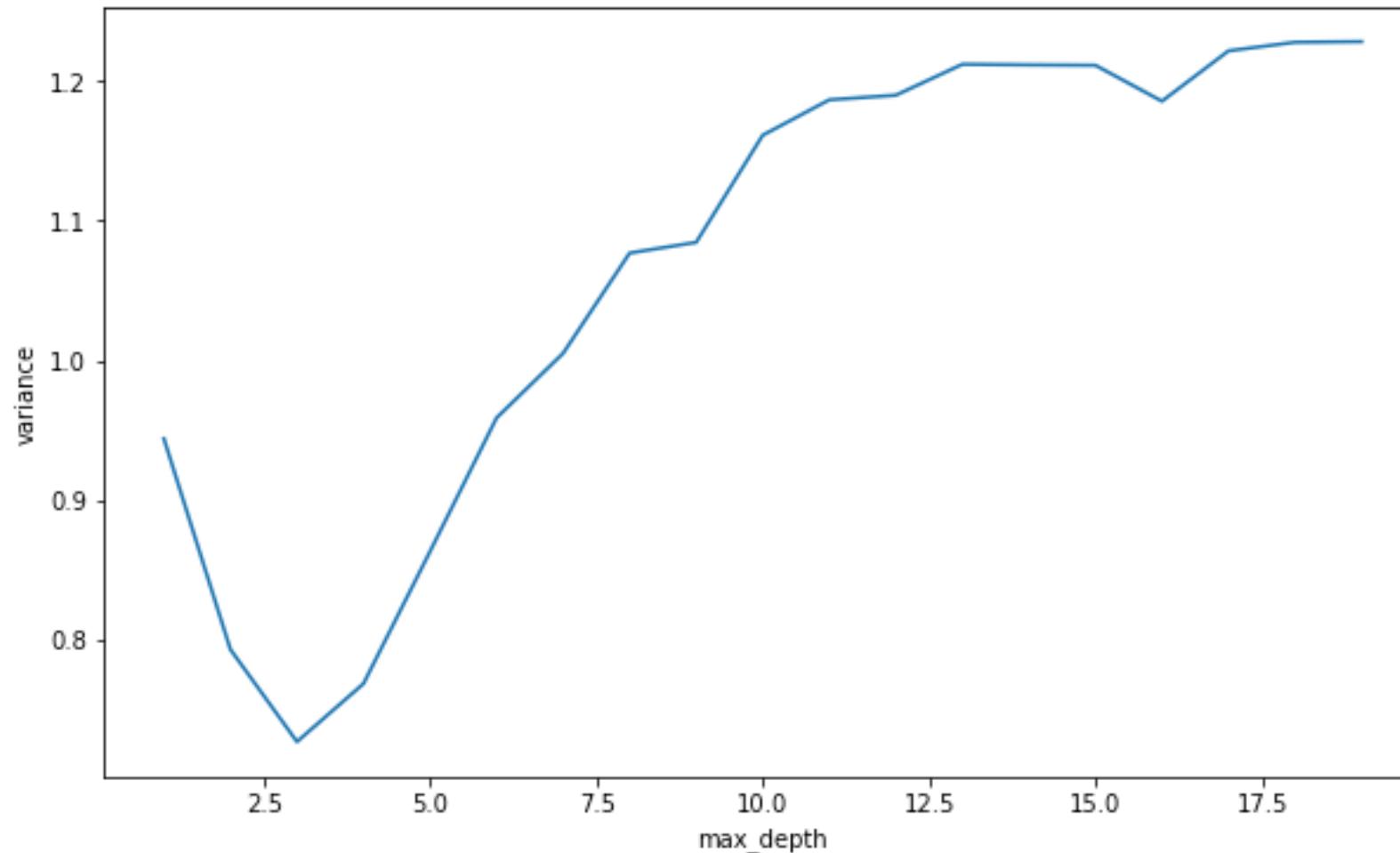
Дерево глубины 16:

Ошибка предсказаний
равномерно распределена
относительно 0
Т.е. низкий bias, но больше
variance

Как зависит от глубины дерева bias



Как зависит от глубины дерева variance



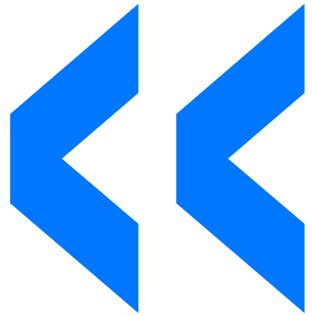
Снижаем variance

$D[tree] = \sigma^2$ - дисперсия 1 дерева

$D\left[\frac{1}{N} \sum tree\right] = (1 - r)\frac{\sigma^2}{N} + r\sigma^2$, где r - корреляция между деревьями

Т.е. чем больше нескоррелированных деревьев, тем меньше variance при сохранении bias

Extreme Random Forest



Вопрос не совсем на собеседовании

А как можно еще снизить корреляцию между
деревьями?

Шаг 1: Выборка

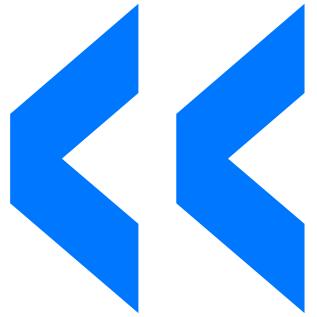
Выкинем Bootstrap, будем брать просто выборки без возвращений.

Шаг 2: Сплиты деревьев

1. Для каждой отобранный фичи выберем не лучший сплит среди возможных значений, а возьмем случайное значение фичи. Просто сэмплируем случайное значение из равномерного распределения между $\min(F)$ и $\max(F)$
2. Лучшую фичу выбираем как обычно: по Gini, энтропии, дисперсии и т.д. среди случайно насэмплированного множества сплитов

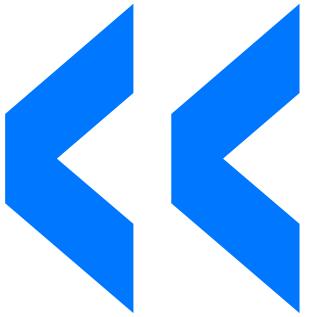
Random Forest - возвращения +
случайные сплиты фичей =
Extreme Random Forest

Мудрость толпы



Реальный эксперимент 1

Группе нужно было назвать, сколько конфет в банке.
Их было **850**, а **средняя оценка группы** оказалась **871**
— намного точнее, чем личные догадки каждого.



Реальный эксперимент 2

Поиск субмарины «Скорпион», затонувшей в 1968 году где-то в Атлантическом океане.

Чтобы найти её, решили собрать группу людей из разных отраслей. Независимо друг от друга они представили свои предположения. Результаты собрали воедино и вычислили примерное местоположение «Скорпиона».

Спустя пять месяцев субмарину нашли в 200 метрах от указанного места

Чтобы коллективное решение было точным, нужно

- **независимость** мнений
- **независимость** участников
- децентрализация (возможность основываться на **независимых** данных)
- **агрегирование** (объединение личных мнений в коллективное решение)

Суд присяжных

p - вероятность правильного приговора

$K = \lfloor \frac{N}{2} \rfloor + 1$ - минимальное большинство для правильного приговора

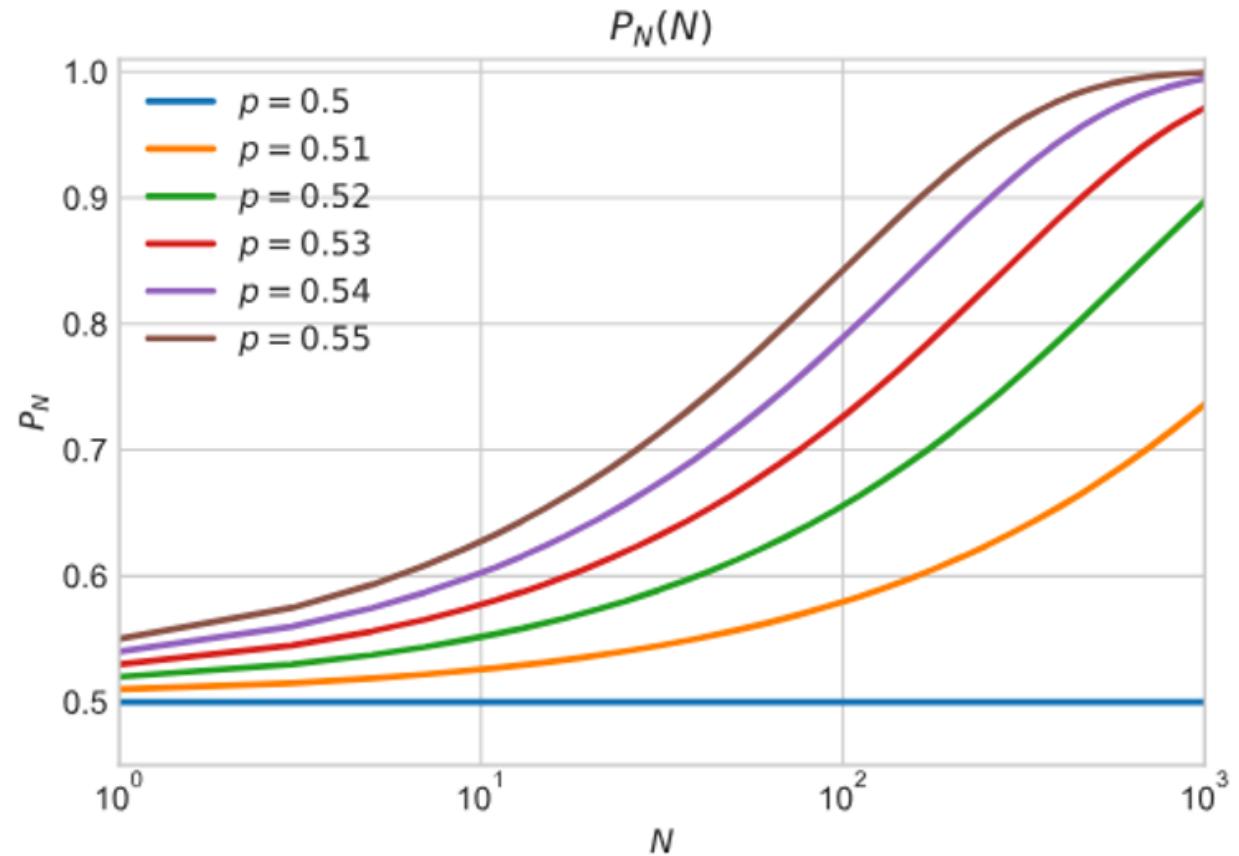
$P = \sum_{i=K}^N C_N^i p^i (1-p)^{N-i}$ - вероятность вынести правильное решение

Суд присяжных

$$P = \sum_{i=K}^N C_N^i p^i (1-p)^{N-i}$$

Если $p > 0.5$, то чем больше присяжных, тем вероятнее правильное решение

Почему тогда не учат 1000000 деревьев?



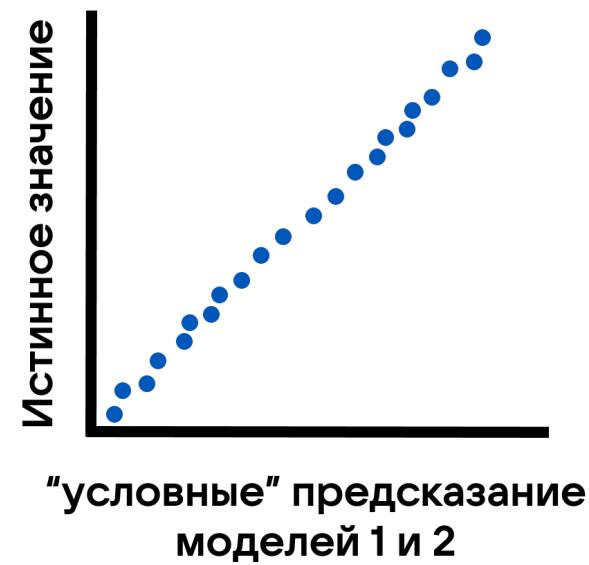
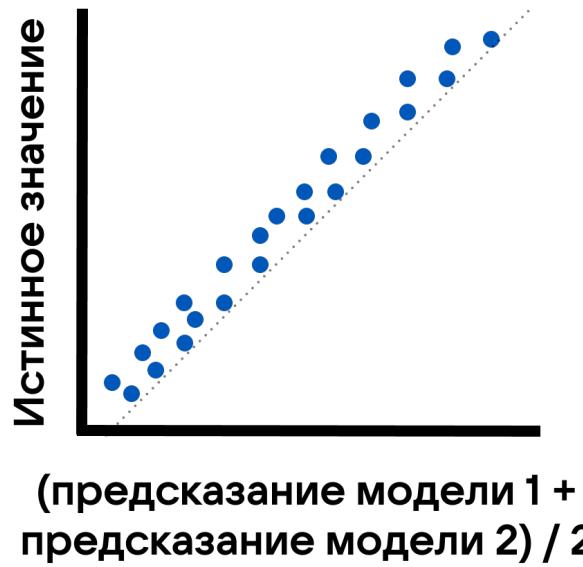
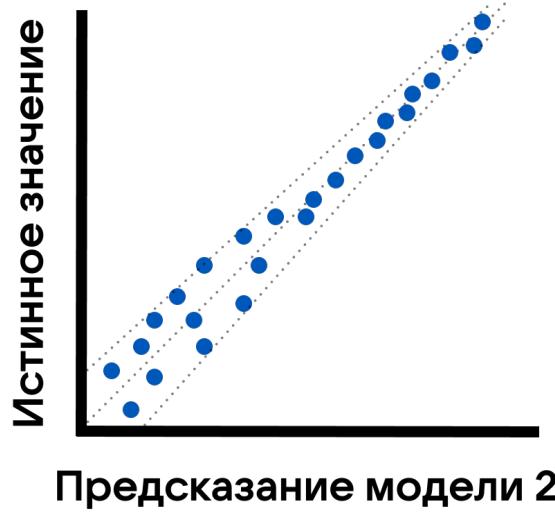
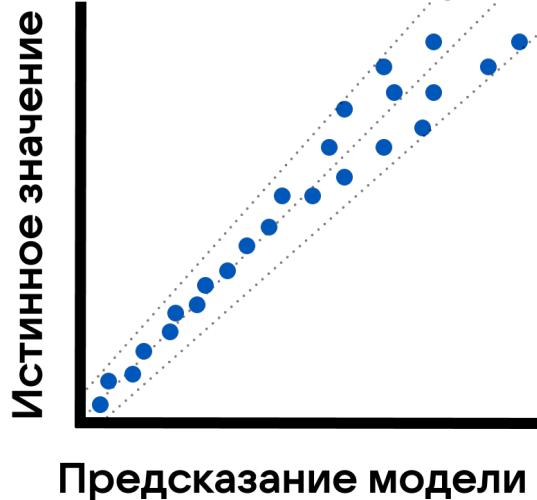
Ансамблирование

А как толпа может голосовать

- Простое голосование (Simple Voting): $f(x) = \frac{1}{m} \sum_{i=1}^m b_i(x)$
- Взвешенное голосование (Weighted Voting): $f(x) = \frac{1}{m} \sum_{i=1}^m w_i b_i(x)$
- Смесь экспертов (Mixture of Experts): $f(x) = \frac{1}{m} \sum_{i=1}^m w_i(x) b_i(x)$

Интуиция

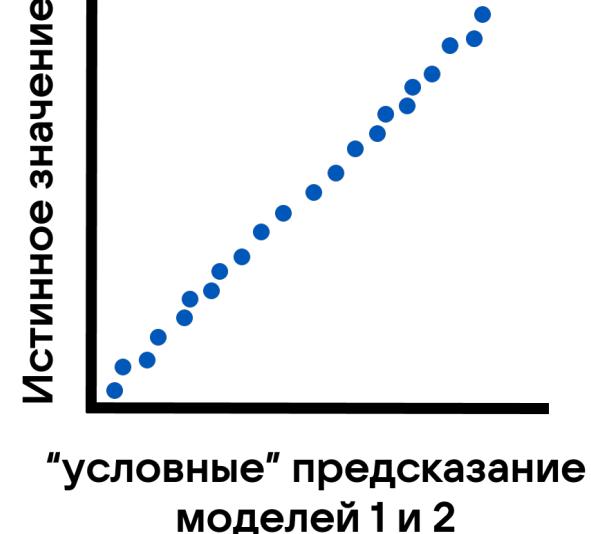
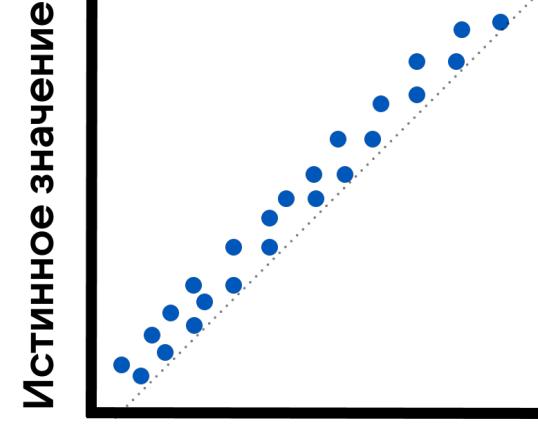
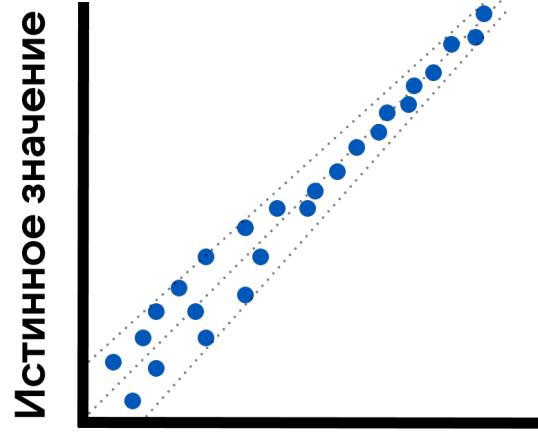
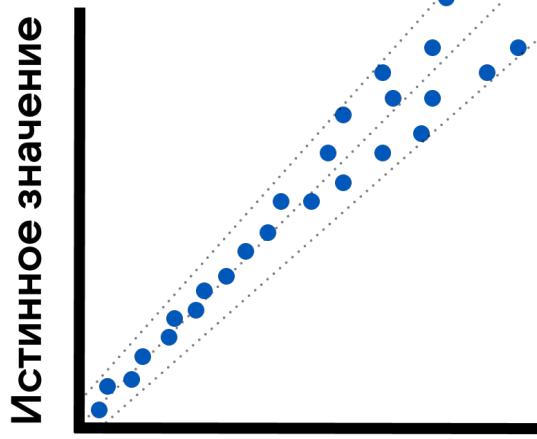
А что если использовать "условные", лучшие предсказания



Мета-алгоритм

Решает, а насколько модели можно доверять в именно этом конкретном объекте. Смесь экспертов (Mixture of Experts)

$$f(x) = \frac{1}{m} \sum_{i=1}^m w_i(x)b_i(x)$$



Blending

Блендинг (Blending)

A

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1

B

x_1	x_2	x_3	y
2.24	1.24	0.53	0
8.52	7.01	0.9	0
0.53	0.24	1.3	1
0.54	3.33	3.41	1

C

x_1	x_2	x_3	y
1.12	5.5	3.33	?
0.53	7.53	3.33	?
3.32	8.52	1.38	?
9.1	3.33	0.53	?

x_{b1}	x_{b2}	x_{b3}	...	x_{bM}	y
					0
					0
					1
					1

B_{meta}

x_{b1}	x_{b2}	x_{b3}	...	x_{bM}	y
					?
					?
					?
					?

C_{meta}

Блендинг (Blending)

A

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1

B

x_1	x_2	x_3	y
2.24	1.24	0.53	0
8.52	7.01	0.9	0
0.53	0.24	1.3	1
0.54	3.33	3.41	1

C

x_1	x_2	x_3	y
1.12	5.5	3.33	?
0.53	7.53	3.33	?
3.32	8.52	1.38	?
9.1	3.33	0.53	?

Учим на A



X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
					0
					0
					1
					1

B_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
					?
					?
					?
					?

C_{meta}

Блендинг (Blending)

A

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1

B

x_1	x_2	x_3	y
2.24	1.24	0.53	0
8.52	7.01	0.9	0
0.53	0.24	1.3	1
0.54	3.33	3.41	1

C

x_1	x_2	x_3	y
1.12	5.5	3.33	?
0.53	7.53	3.33	?
3.32	8.52	1.38	?
9.1	3.33	0.53	?

Предсказываем
на B

Модель b_1

Получаем 1-й
столбец
 B_{meta}

Xb_1	Xb_2	Xb_3	...	Xb_M	y
0.1					0
0.25					0
0.56					1
0.89					1

B_{meta}

Xb_1	Xb_2	Xb_3	...	Xb_M	y
					?
					?
					?
					?

C_{meta}

Блендинг (Blending)

A

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1

B

x_1	x_2	x_3	y
2.24	1.24	0.53	0
8.52	7.01	0.9	0
0.53	0.24	1.3	1
0.54	3.33	3.41	1

C

x_1	x_2	x_3	y
1.12	5.5	3.33	?
0.53	7.53	3.33	?
3.32	8.52	1.38	?
9.1	3.33	0.53	?

Предсказываем
на C



X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.1					0
0.25					0
0.56					1
0.89					1

B_{meta}

Получаем 1-й
столбец
 C_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.33					?
0.28					?
0.57					?
0.99					?

C_{meta}

Блендинг (Blending)

A

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1

B

x_1	x_2	x_3	y
2.24	1.24	0.53	0
8.52	7.01	0.9	0
0.53	0.24	1.3	1
0.54	3.33	3.41	1

C

x_1	x_2	x_3	y
1.12	5.5	3.33	?
0.53	7.53	3.33	?
3.32	8.52	1.38	?
9.1	3.33	0.53	?

Учим на A



X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.1					0
0.25					0
0.56					1
0.89					1

B_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.33					?
0.28					?
0.57					?
0.99					?

C_{meta}

Блендинг (Blending)

A

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1

B

x_1	x_2	x_3	y
2.24	1.24	0.53	0
8.52	7.01	0.9	0
0.53	0.24	1.3	1
0.54	3.33	3.41	1

C

x_1	x_2	x_3	y
1.12	5.5	3.33	?
0.53	7.53	3.33	?
3.32	8.52	1.38	?
9.1	3.33	0.53	?

Предсказываем
на B

Модель b₂

Получаем 2-й
столбец
 B_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.1	0.02				0
0.25	0.34				0
0.56	0.45				1
0.89	0.68				1

B_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.33					?
0.28					?
0.57					?
0.99					?

C_{meta}

Блендинг (Blending)

A

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1

B

x_1	x_2	x_3	y
2.24	1.24	0.53	0
8.52	7.01	0.9	0
0.53	0.24	1.3	1
0.54	3.33	3.41	1

C

x_1	x_2	x_3	y
1.12	5.5	3.33	?
0.53	7.53	3.33	?
3.32	8.52	1.38	?
9.1	3.33	0.53	?

Предсказываем
на C



X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.1	0.02				0
0.25	0.34				0
0.56	0.45				1
0.89	0.68				1

B_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.33	0.29				?
0.28	0.78				?
0.57	0.4				?
0.99	0.66				?

C_{meta}

Получаем 2-й
столбец
 C_{meta}

Блендинг (Blending)

A

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1

B

x_1	x_2	x_3	y
2.24	1.24	0.53	0
8.52	7.01	0.9	0
0.53	0.24	1.3	1
0.54	3.33	3.41	1

C

x_1	x_2	x_3	y
1.12	5.5	3.33	?
0.53	7.53	3.33	?
3.32	8.52	1.38	?
9.1	3.33	0.53	?

Учим на A



X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.1	0.02				0
0.25	0.34				0
0.56	0.45				1
0.89	0.68				1

B_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.33	0.29				?
0.28	0.78				?
0.57	0.4				?
0.99	0.66				?

C_{meta}

Блендинг (Blending)

A

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1

B

x_1	x_2	x_3	y
2.24	1.24	0.53	0
8.52	7.01	0.9	0
0.53	0.24	1.3	1
0.54	3.33	3.41	1

C

x_1	x_2	x_3	y
1.12	5.5	3.33	?
0.53	7.53	3.33	?
3.32	8.52	1.38	?
9.1	3.33	0.53	?

Предсказываем
на B



Получаем 3-й
столбец
 B_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.1	0.02	0.34			0
0.25	0.34	0.5			0
0.56	0.45	0.49			1
0.89	0.68	0.30			1

B_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.33	0.29				?
0.28	0.78				?
0.57	0.4				?
0.99	0.66				?

C_{meta}

Блендинг (Blending)

A

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1

B

x_1	x_2	x_3	y
2.24	1.24	0.53	0
8.52	7.01	0.9	0
0.53	0.24	1.3	1
0.54	3.33	3.41	1

C

x_1	x_2	x_3	y
1.12	5.5	3.33	?
0.53	7.53	3.33	?
3.32	8.52	1.38	?
9.1	3.33	0.53	?

Предсказываем
на C

Модель b_3

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.1	0.02	0.34			0
0.25	0.34	0.5			0
0.56	0.45	0.49			1
0.89	0.68	0.30			1

B_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.33	0.29	0.67			?
0.28	0.78	0.56			?
0.57	0.4	0.33			?
0.99	0.66	0.56			?

C_{meta}

Получаем 3-й
столбец
 C_{meta}

Блендинг (Blending)

A

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1

B

x_1	x_2	x_3	y
2.24	1.24	0.53	0
8.52	7.01	0.9	0
0.53	0.24	1.3	1
0.54	3.33	3.41	1

C

x_1	x_2	x_3	y
1.12	5.5	3.33	?
0.53	7.53	3.33	?
3.32	8.52	1.38	?
9.1	3.33	0.53	?

Учим на A



x_{b1}	x_{b2}	x_{b3}	...	x_{bM}	y
0.1	0.02	0.34	...		0
0.25	0.34	0.5	...		0
0.56	0.45	0.49	...		1
0.89	0.68	0.30	...		1

B_{meta}

x_{b1}	x_{b2}	x_{b3}	...	x_{bM}	y
0.33	0.29	0.67	...		?
0.28	0.78	0.56	...		?
0.57	0.4	0.33	...		?
0.99	0.66	0.56	...		?

C_{meta}

Блендинг (Blending)

A

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1

B

x_1	x_2	x_3	y
2.24	1.24	0.53	0
8.52	7.01	0.9	0
0.53	0.24	1.3	1
0.54	3.33	3.41	1

C

x_1	x_2	x_3	y
1.12	5.5	3.33	?
0.53	7.53	3.33	?
3.32	8.52	1.38	?
9.1	3.33	0.53	?

Предсказываем
на B

Модель b_M

Получаем M-й
столбец
 B_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.1	0.02	0.34	...	0.34	0
0.25	0.34	0.5	...	0.89	0
0.56	0.45	0.49	...	0.25	1
0.89	0.68	0.30	...	0.45	1

B_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.33	0.29	0.67	...		?
0.28	0.78	0.56	...		?
0.57	0.4	0.33	...		?
0.99	0.66	0.56	...		?

C_{meta}

Блендинг (Blending)

A

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1

B

x_1	x_2	x_3	y
2.24	1.24	0.53	0
8.52	7.01	0.9	0
0.53	0.24	1.3	1
0.54	3.33	3.41	1

C

x_1	x_2	x_3	y
1.12	5.5	3.33	?
0.53	7.53	3.33	?
3.32	8.52	1.38	?
9.1	3.33	0.53	?

Учим на A



X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.1	0.02				0
0.25	0.34				0
0.56	0.45				1
0.89	0.68				1

B_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.33	0.29				?
0.28	0.78				?
0.57	0.4				?
0.99	0.66				?

C_{meta}

Блендинг (Blending)

A

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1

B

x_1	x_2	x_3	y
2.24	1.24	0.53	0
8.52	7.01	0.9	0
0.53	0.24	1.3	1
0.54	3.33	3.41	1

C

x_1	x_2	x_3	y
1.12	5.5	3.33	?
0.53	7.53	3.33	?
3.32	8.52	1.38	?
9.1	3.33	0.53	?

Предсказываем
на B



Получаем 3-й
столбец
 B_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.1	0.02	0.34			0
0.25	0.34	0.5			0
0.56	0.45	0.49			1
0.89	0.68	0.30			1

B_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.33	0.29				?
0.28	0.78				?
0.57	0.4				?
0.99	0.66				?

C_{meta}

Блендинг (Blending)

A

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1

B

x_1	x_2	x_3	y
2.24	1.24	0.53	0
8.52	7.01	0.9	0
0.53	0.24	1.3	1
0.54	3.33	3.41	1

C

x_1	x_2	x_3	y
1.12	5.5	3.33	?
0.53	7.53	3.33	?
3.32	8.52	1.38	?
9.1	3.33	0.53	?

Предсказываем
на C

Модель b_3

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.1	0.02	0.34			0
0.25	0.34	0.5			0
0.56	0.45	0.49			1
0.89	0.68	0.30			1

B_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.33	0.29	0.67			?
0.28	0.78	0.56			?
0.57	0.4	0.33			?
0.99	0.66	0.56			?

C_{meta}

Получаем 3-й
столбец
 C_{meta}

Блендинг (Blending)

A

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1

B

x_1	x_2	x_3	y
2.24	1.24	0.53	0
8.52	7.01	0.9	0
0.53	0.24	1.3	1
0.54	3.33	3.41	1

C

x_1	x_2	x_3	y
1.12	5.5	3.33	?
0.53	7.53	3.33	?
3.32	8.52	1.38	?
9.1	3.33	0.53	?

Учим на A



x_{b1}	x_{b2}	x_{b3}	...	x_{bM}	y
0.1	0.02	0.34	...		0
0.25	0.34	0.5	...		0
0.56	0.45	0.49	...		1
0.89	0.68	0.30	...		1

B_{meta}

x_{b1}	x_{b2}	x_{b3}	...	x_{bM}	y
0.33	0.29	0.67	...		?
0.28	0.78	0.56	...		?
0.57	0.4	0.33	...		?
0.99	0.66	0.56	...		?

C_{meta}

Блендинг (Blending)

A

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1

B

x_1	x_2	x_3	y
2.24	1.24	0.53	0
8.52	7.01	0.9	0
0.53	0.24	1.3	1
0.54	3.33	3.41	1

C

x_1	x_2	x_3	y
1.12	5.5	3.33	?
0.53	7.53	3.33	?
3.32	8.52	1.38	?
9.1	3.33	0.53	?

Предсказываем
на B



Получаем M-й
столбец
 B_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.1	0.02	0.34	...	0.34	0
0.25	0.34	0.5	...	0.89	0
0.56	0.45	0.49	...	0.25	1
0.89	0.68	0.30	...	0.45	1

B_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.33	0.29	0.67	...		?
0.28	0.78	0.56	...		?
0.57	0.4	0.33	...		?
0.99	0.66	0.56	...		?

C_{meta}

Блендинг (Blending)

A

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1

B

x_1	x_2	x_3	y
2.24	1.24	0.53	0
8.52	7.01	0.9	0
0.53	0.24	1.3	1
0.54	3.33	3.41	1

C

x_1	x_2	x_3	y
1.12	5.5	3.33	?
0.53	7.53	3.33	?
3.32	8.52	1.38	?
9.1	3.33	0.53	?

Предсказываем
на C

Модель b_M

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.1	0.02	0.34	...	0.34	0
0.25	0.34	0.5	...	0.89	0
0.56	0.45	0.49	...	0.25	1
0.89	0.68	0.30	...	0.45	1

B_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.33	0.29	0.67	...	0.28	?
0.28	0.78	0.56	...	0.4	?
0.57	0.4	0.33	...	0.66	?
0.99	0.66	0.56	...	0.67	?

C_{meta}

Блендинг (Blending)



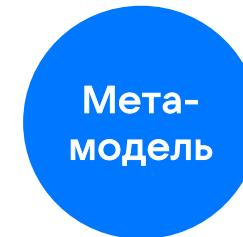
X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.1	0.02	0.34	...	0.34	0
0.25	0.34	0.5	...	0.89	0
0.56	0.45	0.49	...	0.25	1
0.89	0.68	0.30	...	0.45	1

B_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.33	0.29	0.67	...	0.28	?
0.28	0.78	0.56	...	0.4	?
0.57	0.4	0.33	...	0.66	?
0.99	0.66	0.56	...	0.67	?

C_{meta}

Блендинг (Blending)



Предсказываем
итоговые ответы
на C_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.1	0.02	0.34	...	0.34	0
0.25	0.34	0.5	...	0.89	0
0.56	0.45	0.49	...	0.25	1
0.89	0.68	0.30	...	0.45	1

B_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.33	0.29	0.67	...	0.28	0
0.28	0.78	0.56	...	0.4	0
0.57	0.4	0.33	...	0.66	1
0.99	0.66	0.56	...	0.67	1

C_{meta}

Алгоритм

- Даны обучающая (X, y) и тестовая выборки (X_{test}, y_{test})
- Хотим использовать M моделей: $b_1(x), b_2(x), \dots, b_m(x)$
- Поделим (X, y) на 2 части: (X_{train}, y_{train}), (X_{meta}, y_{meta})
- Назовем (X_{train}, y_{train}) = А, (X_{meta}, y_{meta}) = В, (X_{test}, y_{test}) = С
- Для каждой модели b_i :
 - Обучим модель b_i на подвыборке А
 - Для каждого объекта из В сделаем предсказание с помощью b_i , получим i -й столбец матрицы “мета-признаков В”
 - Для каждого объекта из С сделаем предсказание с помощью b_i , получим i -й столбец матрицы “мета-признаков С”
- Получим новую матрицу “мета-признаков” B_{meta} (размера $N_{obj_B} \times M$), составленную из предсказаний моделей b_i для объектов из В, и матрицу “мета-признаков” C_{meta} (размера $N_{obj_C} \times M$), составленную из предсказаний моделей b_i для объектов из С
- Обучим мета-алгоритм b_{meta} на подборке B_{meta}
- Для каждого из объектов из C_{meta} сделаем предсказание с помощью b_{meta} - это и будут ответы блендинга

Базовые алгоритмы

Желательно, максимально разные:

- Случайный лес
- Линейная модель
- SVM
- KNN
- Нейросеть
- Бустинг
- ...

Мета-модель

Мета-модель необязательно сложна - иногда достаточно линейной модели (которая буквально “взвесит” предсказания всех базовых моделей

Hint: к выборке мета-модели можно докинуть часть или все базовые фичи. Получится смесь экспертов (Mixture of Experts)

Минусы блендинга

Базовые алгоритмы учатся только на части трейна.

Мета-алгоритм — тоже только на части трейна, только другой.

Stacking

Стекинг (Stacking)

A			
x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1
8.52	0.46	0.62	1
1.1	0.78	0.24	0

A_1	x_1	x_2	x_3	y
	0.15	0.11	3.41	1
	2.42	2.5	0.12	0
A_2	x_1	x_2	x_3	y
	5.14	3.41	7.62	0
	1.22	0.86	1.38	1
A_3	x_1	x_2	x_3	y
	8.52	0.46	0.62	1
	1.1	0.78	0.24	0

- Пример:
- 3 фолда
- M моделей

C

x_1	x_2	x_3	y
1.12	5.5	3.33	?
0.53	7.53	3.33	?
3.32	8.52	1.38	?
9.1	3.33	0.53	?

Стекинг (Stacking)

A

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1
8.52	0.46	0.62	1
1.1	0.78	0.24	0

A₁

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0

A₂

x_1	x_2	x_3	y
5.14	3.41	7.62	0
1.22	0.86	1.38	1

A₃

x_1	x_2	x_3	y
8.52	0.46	0.62	1
1.1	0.78	0.24	0



На этом фолде обучаем M моделей



Для этого фолда делаем M предсказаний

A_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
					1
					0
					0
					1
					1
					0

Стекинг (Stacking)

A			
A			
x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1
8.52	0.46	0.62	1
1.1	0.78	0.24	0

A_1	x_1	x_2	x_3	y
	0.15	0.11	3.41	1
	2.42	2.5	0.12	0

A_2	x_1	x_2	x_3	y
	5.14	3.41	7.62	0
	1.22	0.86	1.38	1

A_3	x_1	x_2	x_3	y
	8.52	0.46	0.62	1
	1.1	0.78	0.24	0

Шаг

1

:

:



На этом фолде обучаем M
моделей

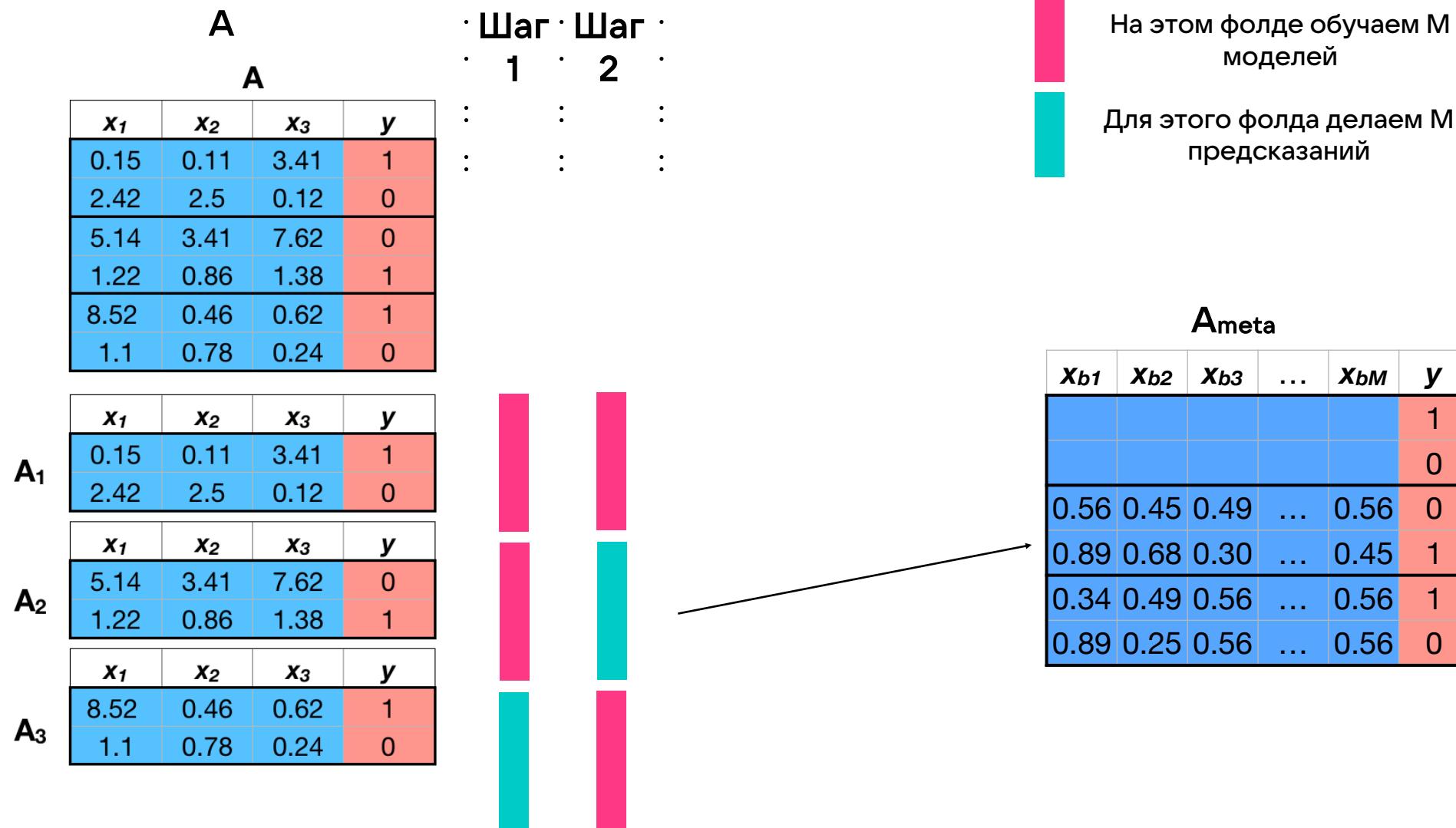


Для этого фолда делаем M
предсказаний

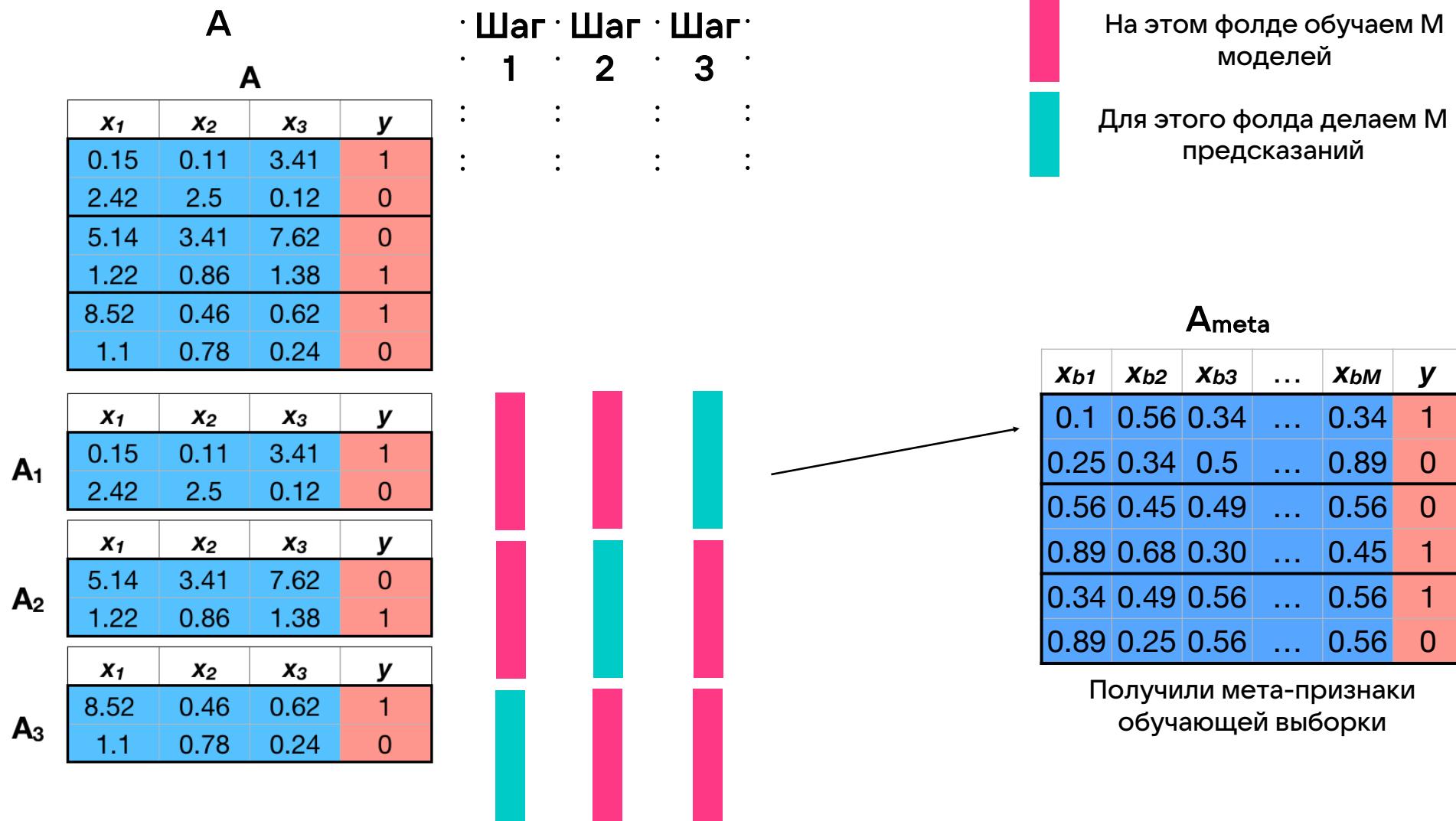
A_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
					1
					0
					0
					1
0.34	0.49	0.56	...	0.56	1
0.89	0.25	0.56	...	0.56	0

Стекинг (Stacking)



Стекинг (Stacking)



Стекинг (Stacking)

C

A

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1
8.52	0.46	0.62	1
1.1	0.78	0.24	0

A_{meta}

x_{b1}	x_{b2}	x_{b3}	...	x_{bM}	y
0.1	0.56	0.34	...	0.34	1
0.25	0.34	0.5	...	0.89	0
0.56	0.45	0.49	...	0.56	0
0.89	0.68	0.30	...	0.45	1
0.34	0.49	0.56	...	0.56	1
0.89	0.25	0.56	...	0.56	0

Мета-признаки обучающей
выборки

x_1	x_2	x_3	y
1.12	5.5	3.33	?
0.53	7.53	3.33	?
3.32	8.52	1.38	?
9.1	3.33	0.53	?

C_{meta}

x_{b1}	x_{b2}	x_{b3}	...	x_{bM}	y
					?
					?
					?
					?

Мета-признаки тестовой
выборки

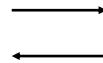
Стекинг (Stacking)

A

A

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1
8.52	0.46	0.62	1
1.1	0.78	0.24	0

Учим на A



C

x_1	x_2	x_3	y
1.12	5.5	3.33	?
0.53	7.53	3.33	?
3.32	8.52	1.38	?
9.1	3.33	0.53	?

A_{meta}

x_{b1}	x_{b2}	x_{b3}	...	x_{bM}	y
0.1	0.56	0.34	...	0.34	1
0.25	0.34	0.5	...	0.89	0
0.56	0.45	0.49	...	0.56	0
0.89	0.68	0.30	...	0.45	1
0.34	0.49	0.56	...	0.56	1
0.89	0.25	0.56	...	0.56	0

Мета-признаки обучающей
выборки

C_{meta}

x_{b1}	x_{b2}	x_{b3}	...	x_{bM}	y
					?
					?
					?
					?

Мета-признаки тестовой
выборки

Стекинг (Stacking)

A

A

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1
8.52	0.46	0.62	1
1.1	0.78	0.24	0



Предсказываем
на C

C

x_1	x_2	x_3	y
1.12	5.5	3.33	?
0.53	7.53	3.33	?
3.32	8.52	1.38	?
9.1	3.33	0.53	?

A_{meta}

x_{b1}	x_{b2}	x_{b3}	...	x_{bM}	y
0.1	0.56	0.34	...	0.34	1
0.25	0.34	0.5	...	0.89	0
0.56	0.45	0.49	...	0.56	0
0.89	0.68	0.30	...	0.45	1
0.34	0.49	0.56	...	0.56	1
0.89	0.25	0.56	...	0.56	0

Мета-признаки обучающей
выборки

C_{meta}

x_{b1}	x_{b2}	x_{b3}	...	x_{bM}	y
0.33	0.29	0.67	...	0.28	?
0.28	0.78	0.56	...	0.4	?
0.57	0.4	0.33	...	0.66	?
0.99	0.66	0.56	...	0.67	?

Мета-признаки тестовой
выборки

Стекинг (Stacking)

A

A

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1
8.52	0.46	0.62	1
1.1	0.78	0.24	0

C

x_1	x_2	x_3	y
1.12	5.5	3.33	?
0.53	7.53	3.33	?
3.32	8.52	1.38	?
9.1	3.33	0.53	?

A_{meta}

x_{b1}	x_{b2}	x_{b3}	...	x_{bM}	y
0.1	0.56	0.34	...	0.34	1
0.25	0.34	0.5	...	0.89	0
0.56	0.45	0.49	...	0.56	0
0.89	0.68	0.30	...	0.45	1
0.34	0.49	0.56	...	0.56	1
0.89	0.25	0.56	...	0.56	0

C_{meta}

x_{b1}	x_{b2}	x_{b3}	...	x_{bM}	y
0.33	0.29	0.67	...	0.28	?
0.28	0.78	0.56	...	0.4	?
0.57	0.4	0.33	...	0.66	?
0.99	0.66	0.56	...	0.67	?

Мета-признаки обучающей
выборки

Мета-признаки тестовой
выборки

Стекинг (Stacking)

A

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1
8.52	0.46	0.62	1
1.1	0.78	0.24	0

A_{meta}

x_{b1}	x_{b2}	x_{b3}	...	x_{bM}	y
0.1	0.56	0.34	...	0.34	1
0.25	0.34	0.5	...	0.89	0
0.56	0.45	0.49	...	0.56	0
0.89	0.68	0.30	...	0.45	1
0.34	0.49	0.56	...	0.56	1
0.89	0.25	0.56	...	0.56	0

Мета-признаки обучающей
выборки

C_{meta}

x_{b1}	x_{b2}	x_{b3}	...	x_{bM}	y
0.33	0.29	0.67	...	0.28	?
0.28	0.78	0.56	...	0.4	?
0.57	0.4	0.33	...	0.66	?
0.99	0.66	0.56	...	0.67	?

Мета-признаки тестовой
выборки

Стекинг (Stacking)

A

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1
8.52	0.46	0.62	1
1.1	0.78	0.24	0



A_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.1	0.56	0.34	...	0.34	1
0.25	0.34	0.5	...	0.89	0
0.56	0.45	0.49	...	0.56	0
0.89	0.68	0.30	...	0.45	1
0.34	0.49	0.56	...	0.56	1
0.89	0.25	0.56	...	0.56	0

Мета-признаки обучающей
выборки

C_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.33	0.29	0.67	...	0.28	?
0.28	0.78	0.56	...	0.4	?
0.57	0.4	0.33	...	0.66	?
0.99	0.66	0.56	...	0.67	?

Мета-признаки тестовой
выборки

Стекинг (Stacking)

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1
8.52	0.46	0.62	1
1.1	0.78	0.24	0

A



Учим на A_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.1	0.56	0.34	...	0.34	1
0.25	0.34	0.5	...	0.89	0
0.56	0.45	0.49	...	0.56	0
0.89	0.68	0.30	...	0.45	1
0.34	0.49	0.56	...	0.56	1
0.89	0.25	0.56	...	0.56	0

A_{meta}

C_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.33	0.29	0.67	...	0.28	?
0.28	0.78	0.56	...	0.4	?
0.57	0.4	0.33	...	0.66	?
0.99	0.66	0.56	...	0.67	?

Мета-признаки обучающей
выборки

Мета-признаки тестовой
выборки

Стекинг (Stacking)

A

x_1	x_2	x_3	y
0.15	0.11	3.41	1
2.42	2.5	0.12	0
5.14	3.41	7.62	0
1.22	0.86	1.38	1
8.52	0.46	0.62	1
1.1	0.78	0.24	0



Предсказываем
итоговые ответы
на C_{meta}

A_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.1	0.56	0.34	...	0.34	1
0.25	0.34	0.5	...	0.89	0
0.56	0.45	0.49	...	0.56	0
0.89	0.68	0.30	...	0.45	1
0.34	0.49	0.56	...	0.56	1
0.89	0.25	0.56	...	0.56	0

C_{meta}

X_{b1}	X_{b2}	X_{b3}	...	X_{bM}	y
0.33	0.29	0.67	...	0.28	0
0.28	0.78	0.56	...	0.4	0
0.57	0.4	0.33	...	0.66	1
0.99	0.66	0.56	...	0.67	1

Мета-признаки обучающей
выборки

Мета-признаки тестовой
выборки

Алгоритм

- Идея стекинга является обобщением идеи блендинга на случай, когда число разбиений обучающей выборки для построения мета-алгоритма больше 2
- Даны обучающая (X, y) и тестовая выборки (X_{test}, y_{test})
- Хотим использовать M моделей: $b_1(x), b_2(x), \dots, b_m(x)$
- Поделим (X, y) на N равных частей (фолдов, как в кросс-валидации)
- Назовем $(X_{train,i}, y_{train,i}) = A_i$ ($i = 1, 2, 3, \dots, N$), $(X_{test}, y_{test}) = C$
- Для каждого фолда A_i :
 - Обучим M моделей на остальных $N-1$ фолдах (то есть на $A_1, A_2, \dots, A_{i-1}, A_{i+1}, \dots, A_N$)
 - Для каждого объекта из A_i сделаем предсказание
 - Получим новую матрицу “мета-признаков” для данного фолда (размера $N_{obj_A_i} \times M$)
- Каждую из базовых моделей обучаем на всей выборке A и делаем предсказания для C , получаем матрицу “мета-признаков” C_{meta} (размера $N_{obj_C} \times M$), составленную из предсказаний моделей b_i для объектов из C
- Обучим мета-алгоритм b_{meta} на подвыборке A_{meta}
- Для каждого из объектов из C_{meta} сделаем предсказание с помощью b_{meta} - это и будут ответы стекинга

Минусы стекинга

Мета-признаки обучающей и тестовой выборки сделаны разными моделями

Стекинг и блэндинг

Преимущества

- Позволяют очень “дешево” повысить качество
- Хорошо аппроксимируют данные благодаря взвешиванию базовых алгоритмов с разными сильными сторонами
- Можно распараллелить (на фолды или модели)

Недостатки

- Требуют большого количества данных
- Долго учатся и долго работают (в зависимости от времени работы базовых моделей)

Применимость

На первый взгляд, только Kaggle

Если присмотреться к ML-системам, ансамблирование в своей неповторимой форме есть почти везде где больше одной модели.

Спасибо!
Задавайте
ваши вопросы,
ну
пожалуйста :)

Алексей Ярошенко
t.me/yaroshenko