Машинное обучение в экономике

Деревья

Потанин Богдан Станиславович

доцент, научный сотрудник, кандидат экономических наук

2023–2024

- Методы классификации и регрессионного анализа:
  - Решающее дерево.
  - Регрессионное дерево.
  - Случайный лес.
- Базовые понятия:
  - Энтропия, выборочная энтропия и средняя энтропия.
  - Ререссионный анализ.
  - Соотношение сложности модели с дисперсией и смещением оценок, декомпозиция среднеквадратической ошибки прогноза.
  - Стрижка деревьев (pruning).
  - Бутстрап.
  - Ансамбли.
  - Бэггинг как пример ансамблевого метода.
  - Ранжирование информативности признаков с помощью случайного леса.

Интуиция

Игра в угадывание персонажа

- **Правила игры** – первый игрок загадывает персонажа, а второй игрок пытается его отгадать, задавая вопросы, на которые можно ответить либо да, либо нет.
- **Цель игры** – отгадать персонажа с использованием минимального числа вопросов.
- Для достижения цели игрок пытается каждый раз задать наиболее **информативный** из возможных вопросов.
- Например, первый вопрос скорее всего будет о поле персонажа или о том, существует ли он в реальности, поскольку ответы на эти вопросы гарантированно позволяют существенно сузить область поиска.
- То, насколько удачным будет очередной вопрос, зависит от ответов на предыдущие вопросы. Например, вопрос о том, умел ли персонаж летать, будет удачным, если из ответов на предыдущие вопросы мы знаем, что загаданный персонаж был животным, и неудачным, если известно, что персонаж был древнегреческим философом.
- Прогнозирование значения целевой переменной можно также рассмотреть как подобную игру, в которой вместо вопросов мы проверяем значения признаков.

Мера неопределенности распределения (impurity measure)

Бинарная переменная

- Рассмотрим бернуллиевскую случайную величину $X \sim \text{Ber}(p)$, где $p \in (0, 1)$.
- Если p близко к 1 или к 0, то случайная величина X будет, как правило, принимать одно и то же значение. То есть мера неопределенности в ее распределении будет низкой.
- Чем ближе p к 0.5, тем сложнее предсказать значение X, а значит мера неопределенности распределения возрастает.
- Функцию, измеряющую меру неопределенности в соответствии с изложенной выше интуицией, можно записать, например, как $\min(p, 1-p)$ или $\text{Var}(X) = p(1-p)$.
- Например, мера неопределенности распределения $X \sim \text{Ber}(0.6)$ выше, чем у распределения $Y \sim \text{Ber}(0.2)$, поскольку $\min(0.6, 1-0.6) = 0.4$ и $\min(0.2, 1-0.2) = 0.2$. По аналогии $0.6(1-0.6) = 0.24 > 0.16 = 0.2(1-0.2)$.
- **Вопрос** – как обобщить идею неопределенности на случай, когда случайная величина X может принимать несколько значений?

Мера неопределенности распределения (impurity measure)

Энтропия

- Рассмотрим дискретную случайную величину X.
- С чем меньшей вероятностью X принимает то или иное значение, тем более неожиданным считается его возникновение.
- Определим меру неожиданности значения $x \in \text{supp}(X)$ как $g\left(P\left(X = x\right)\right)$, где $g(.)$ – строго убывающая функция. Обычно полагают $g\left(P\left(X = x\right)\right) = -\log_2\left(P\left(X = x\right)\right)$.
- **Энтропия** отражает **меру неопределенности** в распределении случайной величины как ожидаемую неожиданность значения, принимаемого случайной величиной:

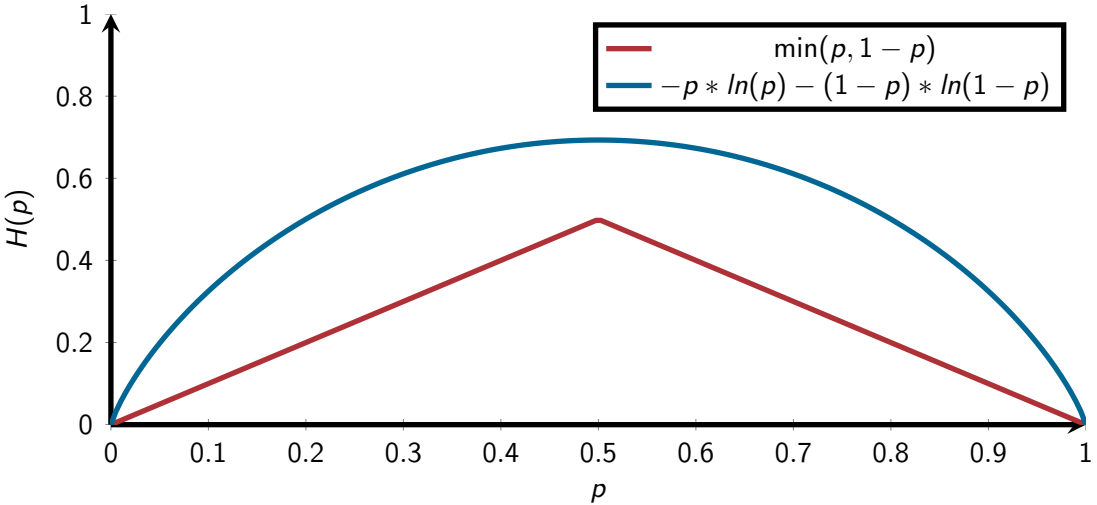
$$H(X) = E(-\log_2\left(P\left(X = x\right)\right)) = -\sum_{x \in \text{supp}(X)} P\left(X = x\right)\log_2\left(P\left(X = x\right)\right)$$

Пример:
Случайная величина X принимает значения 1, 5 и 10 с вероятностями 0.5, 0.3 и 0.2 соответственно. Энтропия этой случайной величины равна:

$$H(X) = -\left(0.5 \times \log_2(0.5) + 0.3 \times \log_2(0.3) + 0.2 \times \log_2(0.2)\right) \approx 1.49$$

Мера неопределенности распределения (impurity measure)

Визуализация для бинарного случая



Мера неопределенности распределения (impurity measure)

Выборочная энтропия

- Рассмотрим выборку $X_1, ..., X_n$ из дискретного распределения и ее реализацию $x_1, ..., x_n$.
- Введем функцию индикатор и среднее значение индикатора, являющееся оценкой вероятности принятия того или иного значения:

$$I(X_i = x) = \begin{cases} 1, \text{ если } X_i = x \\ 0, \text{ в противном случае} \end{cases} \qquad \hat{p}_x = \frac{1}{n} \sum_{i=1}^{n} I(X_i = x)$$

- Обозначим через $y_1, ..., y_m$ все уникальные (без повторов) реализации x_i, встречающиеся в выборке (то есть $m \leq n$) и определим выборочную энтропию следующим образом:

$$\hat{H} = -\sum_{i=1}^{m} \hat{p}_{y_i} \log_2 (\hat{p}_{y_i})$$

- При помощи теоремы Слуцкого нетрудно показать, что выборочная энтропия $\hat{H}$ является состоятельной оценкой энтропии H.

Мера неопределенности распределения (impurity measure)

Пример расчета выборочной энтропии

- Имеется выборка с реализацией $x_1 = 1$, $x_2 = 5$, $x_3 = 1$, $x_4 = 0$, $x_5 = 5$.
- Оставляя лишь уникальные значения, получаем $y_1 = 0$, $y_2 = 1$, $y_3 = 5$.
- Оценим вероятности:

$$\hat{p}_0 = \frac{1}{5}\left(\underbrace{I(1=0)}_{0} + \underbrace{I(5=0)}_{0} + \underbrace{I(1=0)}_{0} + \underbrace{I(0=0)}_{1} + \underbrace{I(5=0)}_{0}\right) = \frac{1}{5} = 0.2$$

$$\hat{p}_1 = \frac{1}{5}\left(\underbrace{I(1=1)}_{1} + \underbrace{I(5=1)}_{0} + \underbrace{I(1=1)}_{1} + \underbrace{I(0=1)}_{1} + \underbrace{I(5=1)}_{0}\right) = \frac{2}{5} = 0.4$$

$$\hat{p}_5 = \frac{1}{5}\left(\underbrace{I(1=5)}_{0} + \underbrace{I(5=5)}_{1} + \underbrace{I(1=5)}_{0} + \underbrace{I(0=5)}_{0} + \underbrace{I(5=5)}_{1}\right) = \frac{2}{5} = 0.4$$

- Посчитаем выборочную энтропию:

$$\hat{H} = -\left(0.4 \times \log_2(0.4) + 0.4 \times \log_2(0.4) + 0.2 \times \log_2(0.2)\right) \approx 1.52$$

Решающее дерево

Первый шаг алгоритма построения решающего дерева

- Через Y обозначим n-мерный вектор значений целевой переменной (target), а через X – матрицу признаков (features) размерности $n \times m$, где m – число признаков, а n – число наблюдений.
- Через X_{i*} будем обозначать i-ю строку матрицы признаков (наблюдение), а через X_{*j} обозначим j-й столбец (признак).
- Через $Y|X_{*j} = k$ обозначим подвектор значений целевой переменной, включающий все наблюдения i такие, что $X_{ij} = k$. Для простоты рассмотрим случай $k \in \{0, 1\}$, когда все признаки бинарные.
- Через n обозначим число наблюдений, то есть элементов вектора Y или строк матрицы X, а через n_j – количество 1 у признака X_{*j}, то есть число наблюдений в векторе $Y|X_{*j} = 1$.
- Определим **среднюю энтропию** (не то же самое, что условная) как:

$$\hat{H}(Y, X_{*j}) = \frac{n_j}{n} \hat{H}(Y|X_{*j} = 1) + \frac{n - n_j}{n} \hat{H}(Y|X_{*j} = 0)$$

  Обозначение $\hat{H}(Y|X_{*i} = k)$ не эквивалентно условной энтропии, а лишь говорит о том, что выборочная энтропия считается по таким Y, что $X_{*j} = k$.
- На первом шаге выбираем признак, обеспечивающий наименьшую среднюю энтропию:

$$j = \underset{j \in \{1, \ldots, m\}}{\operatorname{argmin}} \hat{H}(Y|X_{*j})$$

Пример

Кредитный скоринг

	Дефолт	Работа	Брак	Образование
Клиент 1	1	0	1	1
Клиент 2	1	1	1	1
Клиент 3	1	0	0	0
Клиент 4	1	1	0	0
Клиент 5	1	0	0	0
Клиент 6	0	1	1	0
Клиент 7	0	0	1	0
Клиент 8	0	1	0	0
Клиент 9	0	0	1	1
Клиент 10	0	1	1	0

- Дефолт – факт дефолта по кредиту (1 – наступил дефолт, 0 – не наступил дефолт).
- Работа – статус на рынке труда (1 – работает, 0 – не работает)
- Образование – наличие высшего образования (1 – есть, 0 – нет)
- Брак – брачный статус (1 – в браке, 0 – не в браке).

Таблица: Распределение целевой переменной в зависимости от признаков

	Всего	Дефолт = 1	Доля дефолтов	Энтропия	Средняя энтропия
Работа = 1	5.000	2.000	0.400	0.971	0.971
Работа = 0	5.000	3.000	0.600	0.971	0.971
Брак = 1	6.000	2.000	0.333	0.918	0.875
Брак = 0	4.000	3.000	0.750	0.811	0.875
Образование = 1	3.000	2.000	0.667	0.918	0.965
Образование = 0	7.000	3.000	0.429	0.985	0.965

Пример расчета средней энтропии для брака:

$$\hat{H}(\text{Дефолт}|\text{Брак} = 1) = -\left((2/6) * log_2(2/6) + (4/6) * log_2(4/6)\right) \approx 0.918$$

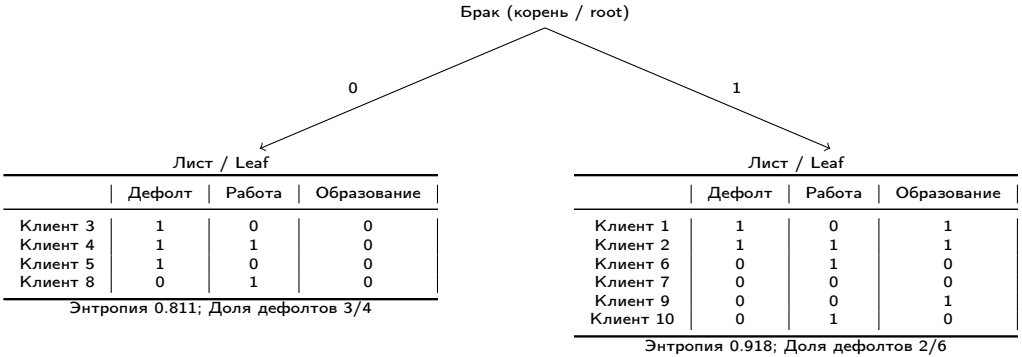
$$\hat{H}(\text{Дефолт}|\text{Брак} = 0) = -\left((3/4) * log_2(3/4) + (1/4) * log_2(1/4)\right) \approx 0.811$$

$$\hat{H}(\text{Дефолт}|\text{Брак}) = (6/10) * 0.918 + (4/10) * 0.811 \approx 0.875$$

Вывод – на первом шаге выбираем **Брак**, поскольку он обладает наименьшей средней энтропией.

Пример

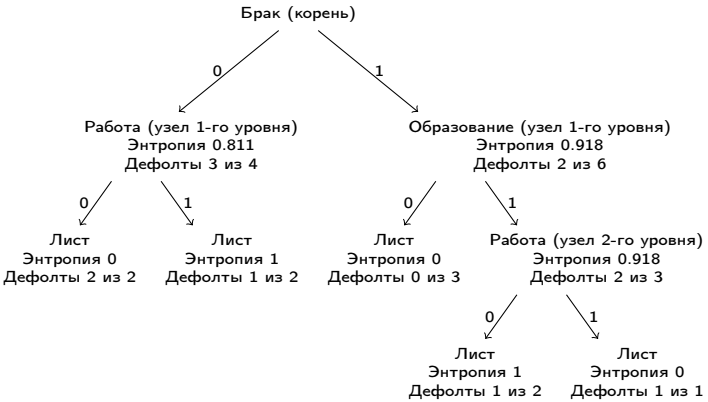
Визуализация первого шага алгоритма построения решающего дерева



- Далее каждое из образовавшихся разбиений независимо друг от друга разделяется по одному из признаков по аналогии с тем, как это происходило на первом шаге.
- Процесс продолжается до тех пор, пока общая энтропия после разбиения не окажется больше, чем до разбиения.

Пример

Визуализация решающего дерева и оценивание вероятностей



В задаче кредитного скоринга вероятность дефолта индивида оценивается как доля дефолтов в листе, в которой попал индивид. Например, вероятность дефолта для женатого безработного индивида с высшим образованием составит $1/2$, а у холостого безработного $2/2 = 1$.

Решающее дерево

Непрерывные, порядковые и категориальные признаки

- Если признак X_{*j} измерен в непрерывной или порядковой шкале, то его перекодируют в бинарную $I(X_{*j} \geq q)$, используя пороговое значение q.
- Для простоты обозначений рассмотрим первый шаг (на остальных по аналогии).
- Подбирается $q \in \left((X_{1j} + X_{2j})/2, ..., \left(X_{(n-1)j} + X_{nj}\right)/2\right)$, минимизирующее среднюю энтропию:

$$\hat{H}(Y, X_{*j}) = \frac{n_j}{n}\hat{H}(Y|I(X_{*j} \geq q) = 1) + \frac{n - n_j}{n}\hat{H}(Y|I(X_{*j} \geq q) = 0)$$

- Узел формируется по критерию $I(X_{*j} \geq q)$, если минимизированная по q средняя энтропия X_{*j} меньше, чем у других признаков.
- Например, узел может разветвляться для людей с доходом не больше 50 тысяч рублей и меньше 50 тысяч рублей.
- В отличие от бинарных переменных, один и тот же порядковый или непрерывный признак может возникать несколько раз в различных узлах, но с разными значениями q.
- Например, тех, кто зарабатывает не больше 50 тысяч рублей, в очередном узле можно разбить на тех, кто зарабатывает до 30 тысяч рублей и от 30 до 50 тысяч рублей.
- Категориальные признаки обычно превращают в дамми-переменные, предварительно объединяя схожие категории.

- Ранее мы рассматривали лишь задачу классификации, то есть предсказания значения категориальных целевых переменных. Однако, в машинном обучении также популярен **регрессионный анализ**, под которым понимается прогнозирование целевых переменных, измерянных в непрерывной шкале: прибыль, расходы, число привлеченных клиентов за рассматриваемый период и т.д.
- В классификационной задаче для получения прогноза сперва, как правило, оценивается условная вероятность $P(Y_i|X_i)$. Затем прогноз строится как функция от этой вероятности, например, в виде $\hat{Y}_i = I(\hat{P}(Y_i = 1|X_i) \geq 0.5)$.
- По аналогии в регрессионном анализе в качестве прогноза Y_i обычно используется оценка условного математического ожидания $E(Y_i|X_i)$, то есть $\hat{Y}_i = \hat{E}(Y_i|X_i)$.

- Если целевая переменная является непрерывной (прибыль, объем продаж и т.д.), то по аналогии с решающим деревом можно построить **регрессионное дерево**.
- **Основная идея** – в случае с непрерывными переменными меру неопределенности естественно измерять как дисперсию. Поэтому вместо средней энтропии минимизируется средняя дисперсия после разбиения.
- Например, если все признаки бинарные, то на первом шаге в качестве признака X_{*j}, использующегося для разбиения, будет выбран тот, что минимизирует **средневшзвешенную дисперсию**:

$$\widehat{Var}(Y, X_{*j}) = \frac{n_j}{n}\widehat{Var}(Y|X_{*j} = 1) + \frac{n - n_j}{n}\widehat{Var}(Y|X_{*j} = 0)$$

- В качестве прогноза целевой переменной, как правило, используется обычное среднее значение этой переменной в листе.

- Метод ближайших соседей можно также применять в задаче регрессионного анализа.
- Аналитическая формула MSE.

- RMSE, MSE и MAPE.

Баланс дисперсии и смещения

Среднеквадратическая ошибка прогноза

- Целевую переменную (в том числе категориальную) можно выразить через ошибку прогноза и условное математическое ожидание:

$$\varepsilon_i = Y_i - E(Y_i|X_i) \implies E(\varepsilon_i|X_i) = 0 \text{ и } Y_i = E(Y_i|X_i) + \varepsilon_i$$

- Рассмотрим функцию $\hat{y}(.)$, которая была оценена на обучающей выборке и используется для прогнозирования целевой переменной y с помощью признаков x на тестовой выборке.
- Часто исследователи ищут модель, минизирующую **среднеквадратическую ошибку прогноза**:

$$\text{MSE} = E\left((y - \hat{y}(x))^2 |x\right) = \text{Var}(E(y|x) + \varepsilon - \hat{y}(x)|x) + (\text{E}(E(y|x) + \varepsilon - \hat{y}(x)|x))^2 =$$

$$= \underbrace{\text{Var}(\hat{y}(x)|x)}_{\text{дисперсия}} + \underbrace{(E(y|x) - E(\hat{y}(x)|x))^2}_{\text{смещение}} + \underbrace{\text{Var}(\varepsilon)}_{\text{шум}}$$

  Где $\text{Cov}(\varepsilon, \hat{y}(x)|x) = 0$, поскольку y не входит в обучающую выборку.
- **Вывод** – мы не можем повлиять на шум, поэтому необходимо искать модель для прогнозирования, минимизирующую дисперсию и смещение прогноза.
- **Важно** – как правило, чем выше (ниже) сложность модели, тем больше (меньше) дисперсия и меньше (больше) смещение ее прогнозов.
- **Примечание** – схожие разложения вомзожны и при иных метриках точности, отличных от MSE.

- Глубина листа определяется уровнем узла, из которого он выходит.
- Чем глубже расположен лист, используемый для прогнозирования, тем, как правило, меньше смещение прогноза (ведь мы используем информацию о большом количестве признаков), но тем больше дисперсия прогноза (поскольку на большой глубине, обычно, остается мало наблюдений).
- В результате среднеквадратическая ошибка прогнозов, полученных с помощью слишком глубоких листов, может оказаться достаточно высокой.
- Таким образом, решающие и регрессионные деревья с большим (относительно числа наблюдений) количеством признаков (сложные модели) склонны к переобучению, поскольку их прогнозы обладают низким смещением, но высокой дисперсией.
- Проблема переобучения возникает и в деревьях с непрерывными признаками, поскольку каждый из них может использоваться большое число раз.

Баланс дисперсии и смещения

Способы поиска оптимального баланса для решающих и регрессионных деревьев

- Наиболее популярные походы к снижению дисперсии прогнозов деревьев:
  - Установить максимальную глубину дерева, то есть запретить формирование узлов более определенного уровня.
  - **Постричь дерево** (pruning), например, заменив узлы на листья в случае, если предсказание из листа на тестовой выборке лучше, чем предсказание из узла.
  - Воспользоваться **ансамблем** методов, например, применив **случайный лес**.
- **Проблема** – обычно снижение дисперсии сопровождается ростом смещения, из-за чего среднеквадратическая ошибка прогноза может возрасти.
- **Решение** – найти оптимальный метод снижения дисперсии. Например, в случае с регрессионными деревьями можно ориентироваться на выборочную среднеквадратичную ошибку, посчитанную по тестовой выборке.

$$\widehat{\text{MSE}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- Идея ансамблей заключается в объединении прогнозов нескольких моделей для получения более точного итогового прогноза.
- Например, итоговый прогноз может быть получен как результат усреднения (необязательно с равными весами) прогнозов нескольких моделей.
- Идея ансамблей схожа с идеей усреднения прогнозов различных экспертов.
- Для построения ансамблей используются различные техники. Наиболее популярными подходами являются **бэггинг** и **бустинг**.

Бэггинг (bagging / bootstrap aggregation)

Алгоритм реализации

- **Бэггинг** можно представить как двухшаговую процедуру обучения модели с использованием бутстрапированных выборок.
- На **первом шаге** формируются k **бутстрапированных выборок** – изначальная выборка из признаков и целевой переменной (X, Y) превращается в новую выборку $(X^{(b)}, Y^{(b)})$, где $b \in \{1, ..., k\}$, с таким же числом наблюдений, за счет **выбора с возвращением**, где каждое наблюдение изначальной выборки может с равной вероятностью попасть в новую.
- Если $k = 3$ и изначальная выборка с одним признаком была $x = (1, 2, 3)$, $y = (4, 5, 6)$, то новые (бутстрапированные) выборки, полученные случайным выбором с возвращением, могут иметь вид, например, $X^{(1)} = (2, 1, 1)$, $Y^{(1)} = (6, 6, 4)$, $X^{(2)} = (3, 1, 2)$, $Y^{(2)} = (5, 6, 4)$ и $X^{(3)} = (3, 3, 3)$, $Y^{(3)} = (5, 5, 5)$.
- На **втором шаге** на каждой выборке $(X^{(b)}, Y^{(b)})$ оценивается модель (например, решающее дерево) и в качестве прогноза выбирается значение, предсказанное наибольшим числом моделей (например, деревьев).
- Если целевая переменная является непрерывной, то в качестве прогноза можно использовать среднее, посчитанное по предсказаниям моделей.

Бэггинг (bagging / bootstrap aggregation)

Ошибка неотобранных элементов (out of bag error)

- Для каждого наблюдения в выборке находят модели, которые не использовали это наблюдение для обучения.
- Строится прогноз для этого наблюдения с использованием лишь соответствующих моделей.
- Ошибка неотобранных элементов (OOB error) считается как средняя ошибка соответствующих прогнозов, например, как доля неверных прогнозов.
- Данный подход дает результаты, схожие с кросс-валидацией, но часто **гораздо** менее ресурсозатратен.

Бэггинг (bagging / bootstrap aggregation)

Почему усреднение повышает точность прогноза?

- Обозначим через $\hat{Y}_i^{(b)}$ прогноз целевой переменной для i-го наблюдения, полученный по b-й модели.
- Поскольку при бэггинге все прогнозы были получены по одной и той же модели, то они одинаково распределены, а значит можем положить $\mathrm{E}(\hat{Y}_i^{(b)}) = \mu$, $\mathrm{Var}(\hat{Y}_i^{(b)}) = \sigma^2$ и $\mathrm{Cov}(\hat{Y}_i^{(b_1)}, \hat{Y}_i^{(b_2)}) = \rho\sigma^2$, где ρ – корреляция между различными прогнозами и $b_1 \neq b_2$.
- Смещение прогнозов от усреднения не изменяется, поскольку:

$$\mathrm{E}\left(\frac{1}{k}\sum_{i=1}^{k}\hat{Y}_i^{(b)}\right) = \frac{n\mu}{n} = \mu$$

- Рассмотрим дисперсию усредненного прогноза:

$$\mathrm{Var}\left(\frac{1}{k}\sum_{i=1}^{k}\hat{Y}_i^{(b)}\right) = \underbrace{\frac{k\sigma^2}{k^2}}_{\text{сумма дисперсий}} + \underbrace{\frac{k(k-1)\rho\sigma^2}{k^2}}_{\text{сумма ковариаций}} = \rho\sigma^2 + \frac{(1-\rho)\sigma^2}{k}$$

  При $k \to \infty$ дисперсия прогноза стремится к $\rho\sigma^2$, что при $\rho \in (0, 1)$ меньше σ^2 – дисперсии оценки одного прогноза. То есть дисперсия прогноза уменьшается.

- **Вывод** – при бэггинге важно снижать корреляцию между прогнозами ρ, чего можно добиться привнося случайность в принцип построения рассматриваемых моделей, например, случайным образом каждый раз отбирая признаки, что, впрочем, может увеличить σ^2.

Случайный лес

Алгоритм построения

- Случайный лес это бэггинг, в котором в качестве модели применяется решающее или регрессионное дерево.
- Каждое отдельное дерево страдает от проблемы переобучения, но поскольку за счет бэггинга деревья обучаются на различных данных, их усредненный прогноз уже не подвержен этой проблеме, если прогнозы не сильно коррелированы (деревья обученные на различных бутстрапированных выборках обычно достаточно сильно отличаются друг от друга).
- В каждом дереве случайного леса используются не все признаки, а лишь **часть** из них, выбираемая случайным образом.
- Выбор случайных признаков **снижает корреляцию прогнозов деревьев**, построенный на различных бутстрапированных выборках, и тем самым уменьшает дисперсию итогового прогноза, но может повышать смещение.
- Число случайно выбираемых признаков является гиперпараметром и часто подбирается на основании кросс-валидации (CV) или ошибки неотобранных элементов (OOB).
- Иногда прогноз получается не за счет усреднения, а с весами, пропорциональными качеству прогнозов соответствующих деревьев, что позволяет снизить негативный эффект наличия малоинформативных признаков на качество прогнозов (tree weighted random forest).

Случайный лес

Ранжирование информативности признаков (подбор оптимальных признаков)

- Случайный лес часто применяют для того, чтобы определить наиболее информативные признаки.
- Существуют различные способы измерения информативности.
- **Перестановочная важность** (permutation importance) – для измерения важности j-го признака можно случайным образом перемешать все его значения в данных. Например, перемешать возраста индивидов в случайном порядке. После этого обучить случайный лес с перемешанными значениями признака и посчитать OOB ошибку, после чего сравнить ее с той, что была до случайного перемешивания. Чем больше окажется разница, тем более существенным является j-й признак.
- **Важность в снижении неопределенности** (impurity importance) – информативными можно считать те признаки, которые обычно существенно снижают меру неопределенности в деревьях.
- Эти методы склонны серьезно переоценивать информативность категориальных переменных, принимающих большое число значений.
- Описанные процедуры чувствительны к качеству исходной модели. Поэтому если исходная модель некачественная, то информативные признаки могут быть отобраны некорректно.
- Отобранные случайным лесом информативные признаки могут применяться для построения других моделей, например, Байесовских сетей.