Метод ближайших соседей и мэтчинг

Метод ближайших соседей

Метод ближайших соседей: план

- Использование соседей в задаче регрессии.

Метод ближайших соседей: план

- Использование соседей в задаче регрессии.
- Кого считать соседом?

Метод ближайших соседей: план

- Использование соседей в задаче регрессии.
- Кого считать соседом?
- Использование соседей в задаче классификации.

Метод k ближайших соседей: регрессия

Цель: хотим спрогнозировать непрерывную величину y.

Метод k ближайших соседей: регрессия

Цель: хотим спрогнозировать непрерывную величину y.

Не предполагаем линейной зависимости y от предикторов.

Скажи мне, кто твой друг, и я скажу, кто ты.

Мигель де Сервантес

wikipedia.org

Ближайшие соседи

Зависимая переменная: y_i.

Предикторы: $x_i = (a_i, b_i, c_i, \ldots)$.

Ближайшие соседи

Зависимая переменная: y_i.

Предикторы: $x_i = (a_i, b_i, c_i, \ldots)$.

Расстояние между двумя наблюдениями:

$$d(x_1, x_2) = \sqrt{(a_1 - a_2)^2 + (b_1 - b_2)^2 + (c_1 - c_2)^2 + \ldots}$$

Ближайшие соседи

Зависимая переменная: y_i.

Предикторы: $x_i = (a_i, b_i, c_i, \ldots)$.

Расстояние между двумя наблюдениями:

$$d(x_1, x_2) = \sqrt{(a_1 - a_2)^2 + (b_1 - b_2)^2 + (c_1 - c_2)^2 + \ldots}$$

Естественное определение

Ближайшими соседями наблюдения x называем те наблюдения, расстояние от него до которых наименьшее.

Прогнозирование

Цель: построить прогноз для $x = (a, b, c, \ldots)$.

Прогнозирование

Цель: построить прогноз для $x = (a, b, c, \ldots)$.

Выбираем $k = 3$ ближайших соседей данного наблюдения.

Прогнозирование

Цель: построить прогноз для $x = (a, b, c, \ldots)$.

Выбираем $k = 3$ ближайших соседей данного наблюдения.

Допустим это оказались наблюдения номер 5, 42 и 100.

Прогнозирование

Цель: построить прогноз для $x = (a, b, c, \ldots)$.

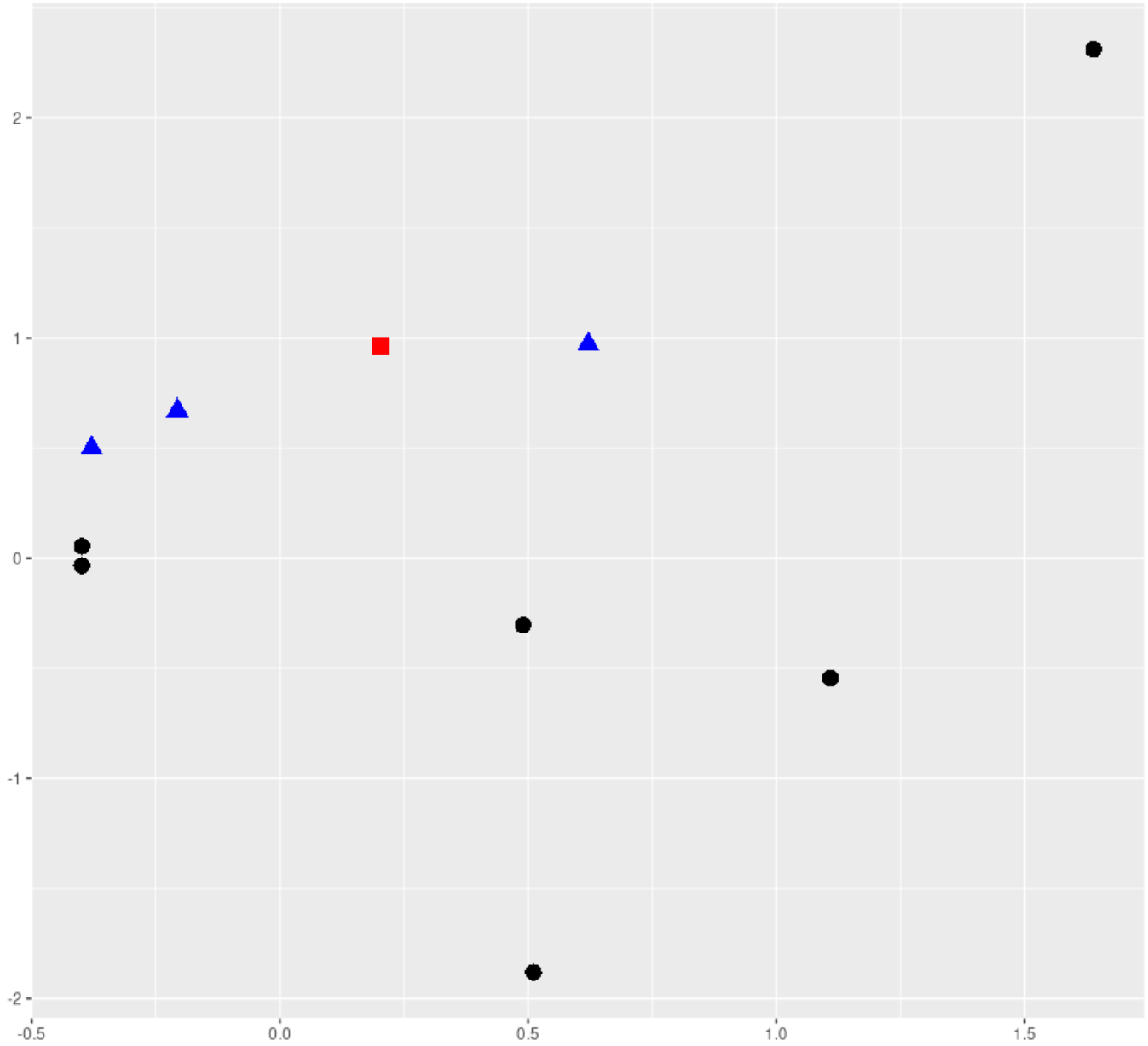
Выбираем $k = 3$ ближайших соседей данного наблюдения.

Допустим это оказались наблюдения номер 5, 42 и 100.

Считаем прогноз для вектора предикторов x как среднее:

$$\hat{y} = \frac{y_5 + y_{42} + y_{100}}{3}.$$

Треугольники — соседи квадрата



Важность нормировки

Расстояние чувствительно к выбору масштаба.

$$d(x_i, x_j) = \sqrt{(a_i - a_j)^2 + (b_i - b_j)^2 + (c_i - c_j)^2 + \dots}$$

Важность нормировки

Расстояние чувствительно к выбору масштаба.

$$d(x_i, x_j) = \sqrt{(a_i - a_j)^2 + (b_i - b_j)^2 + (c_i - c_j)^2 + \dots}$$

Важно избавиться от единиц измерения!

Важность нормировки

Расстояние чувствительно к выбору масштаба.

$$d(x_i, x_j) = \sqrt{(a_i - a_j)^2 + (b_i - b_j)^2 + (c_i - c_j)^2 + \dots}$$

Важно избавиться от единиц измерения!

Способ 1:

$$a_i \to \frac{a_i - \min(a)}{\max(a) - \min(a)}.$$

Важность нормировки

Расстояние чувствительно к выбору масштаба.

$$d(x_i, x_j) = \sqrt{(a_i - a_j)^2 + (b_i - b_j)^2 + (c_i - c_j)^2 + \dots}$$

Важно избавиться от единиц измерения!

Способ 1:

$$a_i \to \frac{a_i - \min(a)}{\max(a) - \min(a)}.$$

После преобразования все переменные лежат в отрезке $[0; 1]$.

Важность нормировки

Расстояние чувствительно к выбору масштаба.

$$d(x_i, x_j) = \sqrt{(a_i - a_j)^2 + (b_i - b_j)^2 + (c_i - c_j)^2 + \dots}$$

Важно избавиться от единиц измерения!

Способ 1:

$$a_i \to \frac{a_i - \min(a)}{\max(a) - \min(a)}.$$

После преобразования все переменные лежат в отрезке $[0; 1]$.

Данное масштабирование чувствительно к выбросам.

Избавиться от единиц измерения!

Способ 2:

$$a_i \to \frac{a_i - \bar{a}}{\sqrt{\frac{\sum (a_j - \bar{a})^2}{n-1}}}.$$

Избавиться от единиц измерения!

Способ 2:

$$a_i \to \frac{a_i - \bar{a}}{\sqrt{\frac{\sum (a_j - \bar{a})^2}{n-1}}}.$$

После преобразования каждая переменная имеет среднее равное нулю и около 95% её значений лежат в отрезке $[-2; 2]$.

Избавиться от единиц измерения!

Способ 2:

$$a_i \to \frac{a_i - \bar{a}}{\sqrt{\frac{\sum (a_j - \bar{a})^2}{n-1}}}.$$

После преобразования каждая переменная имеет среднее равное нулю и около 95% её значений лежат в отрезке $[-2; 2]$.

Данное масштабирование не учитывает выборочную корреляцию между переменными.

Избавиться от единиц измерения!

Способ 3: «метрика Махаланобиса»

Матрицу центрированных переменных умножаем слева на корень из выборочной ковариационной матрицы.

Избавиться от единиц измерения!

Способ 3: «метрика Махаланобиса»

Матрицу центрированных переменных умножаем слева на корень из выборочной ковариационной матрицы.

Каждая новая переменная имеет среднее равное нулю и около 95% её значений лежат в отрезке $[-2; 2]$.

Избавиться от единиц измерения!

Способ 3: «метрика Махаланобиса»

Матрицу центрированных переменных умножаем слева на корень из выборочной ковариационной матрицы.

Каждая новая переменная имеет среднее равное нулю и около 95% её значений лежат в отрезке $[-2; 2]$.

Новые переменные имеют нулевую выборочную корреляцию.

Как выбрать количество соседей?

Основной способ: кросс-валидация.

Как выбрать количество соседей?

Основной способ: кросс-валидация.

Для нескольких разных k, например, для $k \in \{1, 2, 3, 4, 5\}$:

1. Построим прогноз для каждого наблюдения, используя k его ближайших соседей.

Как выбрать количество соседей?

Основной способ: кросс-валидация.

Для нескольких разных k, например, для $k \in \{1, 2, 3, 4, 5\}$:

1. Построим прогноз для каждого наблюдения, используя k его ближайших соседей.
2. Посчитаем сумму квадратов ошибок прогнозов

$$MSE_k = \frac{1}{n} \sum (y_i - \hat{y}_i^{cv})^2.$$

Как выбрать количество соседей?

Основной способ: кросс-валидация.

Для нескольких разных k, например, для $k \in \{1, 2, 3, 4, 5\}$:

1. Построим прогноз для каждого наблюдения, используя k его ближайших соседей.
2. Посчитаем сумму квадратов ошибок прогнозов

$$MSE_k = \frac{1}{n} \sum (y_i - \hat{y}_i^{cv})^2.$$

Как выбрать количество соседей?

Основной способ: кросс-валидация.

Для нескольких разных k, например, для $k \in \{1, 2, 3, 4, 5\}$:

1. Построим прогноз для каждого наблюдения, используя k его ближайших соседей.
2. Посчитаем сумму квадратов ошибок прогнозов

$$MSE_k = \frac{1}{n} \sum (y_i - \hat{y}_i^{cv})^2.$$

Выберем оптимальное k.

Задача классификации

Что изменится, если y — бинарная переменная?

Задача классификации

Что изменится, если y — бинарная переменная?

Кратко: ничего.

Задача классификации

Что изменится, если y — бинарная переменная?

Кратко: ничего.

Зависимая переменная y закодирована как 0 и 1, а x_5, x_{42}, x_{100} — ближайшие наблюдения к x.

Считаем прогноз для вектора предикторов x как среднее:

$$\hat{y} = \frac{y_5 + y_{42} + y_{100}}{3}.$$

Задача классификации

Что изменится, если y — бинарная переменная?

Кратко: ничего.

Зависимая переменная y закодирована как 0 и 1, а x_5, x_{42}, x_{100} — ближайшие наблюдения к x.

Считаем прогноз для вектора предикторов x как среднее:

$$\hat{y} = \frac{y_5 + y_{42} + y_{100}}{3}.$$

Величина $\hat{y}$ — оценка вероятности того, что $y = 1$.

Метод k ближайших соседей: итоги

- Предсказывает непрерывную или дискретную y.

Метод k ближайших соседей: итоги

- Предсказывает непрерывную или дискретную y.
- Не требует явного предположения о виде зависимости от предикторов.

Метод k ближайших соседей: итоги

- Предсказывает непрерывную или дискретную y.
- Не требует явного предположения о виде зависимости от предикторов.
- Важно привести предикторы к общему масштабу.

Метод k ближайших соседей: итоги

- Предсказывает непрерывную или дискретную y.
- Не требует явного предположения о виде зависимости от предикторов.
- Важно привести предикторы к общему масштабу.
- Нет коэффициентов, чтобы интерпретировать.

Метод k ближайших соседей: итоги

- Предсказывает непрерывную или дискретную y.
- Не требует явного предположения о виде зависимости от предикторов.
- Важно привести предикторы к общему масштабу.
- Нет коэффициентов, чтобы интерпретировать.
- Можно комбинировать с другими методами.

Мэтчинг

Мэтчинг: план

- Задача оценки эффекта влияния.

Мэтчинг: план

- Задача оценки эффекта влияния.
- Алгоритм мэтчинга.

Мэтчинг: план

- Задача оценки эффекта влияния.
- Алгоритм мэтчинга.
- Мера склонности к воздействию.

Цель анализа

Данные:

$y_i \in \mathbb{R}$ — значение целевой переменной;

$a_i \in \{0, 1\}$ — индикатор того, что индивид i получил воздействие;

$x_i \in \mathbb{R}$ — прочие характеристики индивида.

Цель анализа

Данные:

$y_i \in \mathbb{R}$ — значение целевой переменной;

$a_i \in \{0, 1\}$ — индикатор того, что индивид i получил воздействие;

$x_i \in \mathbb{R}$ — прочие характеристики индивида.

Хотим оценить влияние бинарной переменной a_i на y_i.

Формализуем понятие «влияет»

Гипотетические значения:

$y_i(0)$ — значение целевой переменной, если бы индивид не получил воздействие;

$y_i(1)$ — значение целевой переменной, если бы индивид получил воздействие;

Формализуем понятие «влияет»

Гипотетические значения:

$y_i(0)$ — значение целевой переменной, если бы индивид не получил воздействие;

$y_i(1)$ — значение целевой переменной, если бы индивид получил воздействие;

Задача

Хотим оценить средний эффект воздействия:

$$ATE = E(y_i(1) - y_i(0)).$$

А в чём проблема?

Задача

Хотим оценить средний эффект воздействия:

$$ATE = E(y_i(1) - y_i(0)).$$

А в чём проблема?

Задача

Хотим оценить средний эффект воздействия:

$$ATE = E(y_i(1) - y_i(0)).$$

Наблюдаем только одну из гипотетических величин!

У тех, кто получил воздействие, видим $y_i(1)$.

У тех, кто не получил воздействие, видим $y_i(0)$.

$$y_i = y_i(a_i).$$

Рандомизированный эксперимент

Рандомизированный эксперимент

Воздействие a_i назначается или выбирается случайно и не зависит от $y_i(0)$ и $y_i(1)$.

$$a_i \perp y_i(0), y_i(1)$$

Рандомизированный эксперимент

Рандомизированный эксперимент

Воздействие a_i назначается или выбирается случайно и не зависит от $y_i(0)$ и $y_i(1)$.

$$a_i \perp y_i(0), y_i(1)$$

При назначении воздействия мы не предпочитаем тех, кто на на него лучше среагирует.

Нет самостоятельного выбора воздействия индивидами, умеющими прогнозировать $y_i(0)$ и $y_i(1)$.

В мире розовых пони…

В прекрасном мире рандомизированного эксперимента:

Любая регрессия даст несмещённую и состоятельную оценку ATE.

В мире розовых пони…

В прекрасном мире рандомизированного эксперимента:

Любая регрессия даст несмещённую и состоятельную оценку ATE.

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_a a_i,$$

В мире розовых пони…

В прекрасном мире рандомизированного эксперимента:

Любая регрессия даст несмещённую и состоятельную оценку ATE.

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_a a_i,$$

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_a a_i + \hat{\beta}_x x_i,$$

В мире розовых пони…

В прекрасном мире рандомизированного эксперимента:

Любая регрессия даст несмещённую и состоятельную оценку ATE.

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_a a_i,$$

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_a a_i + \hat{\beta}_x x_i,$$

В рандомизированном эксперименте:

$$\operatorname{plim}_{n\to\infty} \hat{\beta}_a = ATE, \quad \mathbb{E}(\hat{\beta}_a) = ATE.$$

Отсутствие рандомизации

y_i — результат теста по теории вероятностей;

$a_i \in \{0, 1\}$ — пользовался ли студент шпаргалкой;

x_i — уровень знаний студента.

Отсутствие рандомизации

y_i — результат теста по теории вероятностей;

$a_i \in \{0, 1\}$ — пользовался ли студент шпаргалкой;

x_i — уровень знаний студента.

Предположим, что только слабые студенты пользуются шпаргалкой.

Отсутствие рандомизации

y_i — результат теста по теории вероятностей;

$a_i \in \{0, 1\}$ — пользовался ли студент шпаргалкой;

x_i — уровень знаний студента.

Предположим, что только слабые студенты пользуются шпаргалкой.

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_a a_i$$

Регрессия недооценивает эффект, $E(\hat{\beta}_a) < ATE$.

Отсутствие рандомизации

y_i — результат теста по теории вероятностей;

$a_i \in \{0, 1\}$ — пользовался ли студент продвинутым учебником;

x_i — уровень знаний студента.

Отсутствие рандомизации

y_i — результат теста по теории вероятностей;

$a_i \in \{0, 1\}$ — пользовался ли студент продвинутым учебником;

x_i — уровень знаний студента.

Предположим, что только сильные студенты пользуются продвинутым учебником.

Отсутствие рандомизации

y_i — результат теста по теории вероятностей;

$a_i \in \{0, 1\}$ — пользовался ли студент продвинутым учебником;

x_i — уровень знаний студента.

Предположим, что только сильные студенты пользуются продвинутым учебником.

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_a a_i$$

Регрессия переоценивает эффект, $E(\hat{\beta}_a) > ATE$.

Суровая реальность…

Узнав о возможности получить воздействие, индивиды сами решают, получать ли воздействие.

Значение a_i связано с $y_i(0)$ и $y_i(1)$.

Суровая реальность…

Узнав о возможности получить воздействие, индивиды сами решают, получать ли воздействие.

Значение a_i связано с $y_i(0)$ и $y_i(1)$.

В общем случае всё пропало, найти состоятельные и несмещённые оценки ATE невозможно.

Суровая реальность…

Узнав о возможности получить воздействие, индивиды сами решают, получать ли воздействие.

Значение a_i связано с $y_i(0)$ и $y_i(1)$.

В общем случае всё пропало, найти состоятельные и несмещённые оценки ATE невозможно.

Последняя надежда!

Воздействие не зависит от гипотетических исходов при фиксированном x_i.

$$a_i \perp y_i(0), y_i(1) \mid x_i$$

Мэтчинг и регрессия

1. Создаём кластеры исходных наблюдений.
   В один кластер входят наблюдения с близкими характеристиками x_i и разными значениями a_i.
   Внутри кластера a_i не зависит от потенциальных исходов $y_i(0)$ и $y_i(1)$.

Мэтчинг и регрессия

1. Создаём кластеры исходных наблюдений.
   В один кластер входят наблюдения с близкими характеристиками x_i и разными значениями a_i.
   Внутри кластера a_i не зависит от потенциальных исходов $y_i(0)$ и $y_i(1)$.
2. Строим регрессию y_i на a_i с использованием весов наблюдений и с робастными стандартными ошибками.
   Веса нивелируют различие в x_i для получивших и не получивших воздействие.
   Робастные стандартные ошибки учтут гетероскедастичность и корреляцию внутри кластера.

Создаём кластеры наблюдений!

Кластеры бывают двух типов:

- Один индивид, получивший воздействие, и несколько близких к нему индвидов, не получивших воздействия.

Создаём кластеры наблюдений!

Кластеры бывают двух типов:

- Один индивид, получивший воздействие, и несколько близких к нему индвидов, не получивших воздействия.
- Один индивид, не получивший воздействия, и несколько близких к нему индвидов, получивших воздействие.

Создаём кластеры наблюдений!

Кластеры бывают двух типов:

- Один индивид, получивший воздействие, и несколько близких к нему индвидов, не получивших воздействия.
- Один индивид, не получивший воздействия, и несколько близких к нему индвидов, получивших воздействие.

При создании кластеров наблюдения получат веса.

Создаём кластеры наблюдений!

Кластеры бывают двух типов:

- Один индивид, получивший воздействие, и несколько близких к нему индвидов, не получивших воздействия.
- Один индивид, не получивший воздействия, и несколько близких к нему индвидов, получивших воздействие.

При создании кластеров наблюдения получат веса.

Как выбрать близких индивидов к данному?

- Расстояние Махаланобиса для прочих характеристик x_i.

Создаём кластеры наблюдений!

Кластеры бывают двух типов:

- Один индивид, получивший воздействие, и несколько близких к нему индвидов, не получивших воздействия.
- Один индивид, не получивший воздействия, и несколько близких к нему индвидов, получивших воздействие.

При создании кластеров наблюдения получат веса.

Как выбрать близких индивидов к данному?

- Расстояние Махаланобиса для прочих характеристик x_i.
- Мера склонности к воздействию (propensity score), $\widehat{ps}(x_i)$.

Мера склонности

Мера склонности

Оценка вероятности получить воздействие, полученная с помощью прочих характеристик индивида x_i.

$$\widehat{ps}(x_i) = \hat{P}(a_i = 1 \mid x_i).$$

Мера склонности

Мера склонности

Оценка вероятности получить воздействие, полученная с помощью прочих характеристик индивида x_i.

$$\widehat{ps}(x_i) = \hat{P}(a_i = 1 \mid x_i).$$

Например, можно оценить логистическую регрессию

$$\hat{P}(a_i = 1 \mid x_i) = \Lambda(\hat{\beta}_1 + \hat{\beta}_x x_i),$$

где $\Lambda(t) = \exp(t)/(1 + \exp(t))$.

Мера склонности и веса

Мера склонности

Оценка вероятности получить воздействие, полученная с помощью прочих характеристик индивида x_i.

$$\widehat{ps}(x_i) = \hat{P}(a_i = 1 \mid x_i).$$

Мера склонности и веса

Мера склонности

Оценка вероятности получить воздействие, полученная с помощью прочих характеристик индивида x_i.

$$\widehat{ps}(x_i) = \hat{P}(a_i = 1 \mid x_i).$$

Вес наблюдения

$$w_i = \begin{cases} 1/\widehat{ps}(x_i), \text{ если } a_i = 1; \\ 1/(1 - \widehat{ps}(x_i)), \text{ если } a_i = 0; \end{cases}$$

Мера склонности и веса

Мера склонности

Оценка вероятности получить воздействие, полученная с помощью прочих характеристик индивида x_i.

$$\widehat{ps}(x_i) = \hat{P}(a_i = 1 \mid x_i).$$

Вес наблюдения

$$w_i = \begin{cases} 1/\widehat{ps}(x_i), \text{ если } a_i = 1; \\ 1/(1 - \widehat{ps}(x_i)), \text{ если } a_i = 0; \end{cases}$$

Вес наблюдения — насколько редко встречаются наблюдения с данным a_i среди похожих по x_i.

Рецепт и большая наука

Краткий рецепт

1. Создаём кластеры наблюдений отличающихся только воздействием.
2. Оцениваем регрессию и получаем доверительный интервал для эффекта влияния.

Важно: наличие воздействия a_i должно не зависеть от потенциальных исходов $y_i(1)$ и $y_i(0)$ при фиксированных x_i.

Рецепт и большая наука

Краткий рецепт

1. Создаём кластеры наблюдений отличающихся только воздействием.
2. Оцениваем регрессию и получаем доверительный интервал для эффекта влияния.

Важно: наличие воздействия a_i должно не зависеть от потенциальных исходов $y_i(1)$ и $y_i(0)$ при фиксированных x_i.

Много интересного!

1. ATE, ATT, ATM, непараметрические методы, …
2. Многие методы ещё не имеют глубокого теоретического обоснования.

Соседи и мэтчинг: итоги

Общая идея

Метод ближайших соседей: используем похожие наблюдения, чтобы прогнозировать y_i.

Мэтчинг: используем похожие наблюдения, чтобы оценить эффект a_i.

Соседи и мэтчинг: итоги

Общая идея

Метод ближайших соседей: используем похожие наблюдения, чтобы прогнозировать y_i.

Мэтчинг: используем похожие наблюдения, чтобы оценить эффект a_i.

- Мэтчинг требует предположения $a_i \perp y_i(1), y_i(0) \mid x_i$.

Соседи и мэтчинг: итоги

Общая идея

Метод ближайших соседей: используем похожие наблюдения, чтобы прогнозировать y_i.

Мэтчинг: используем похожие наблюдения, чтобы оценить эффект a_i.

- Мэтчинг требует предположения $a_i \perp y_i(1), y_i(0) \mid x_i$.
- Мэтчинг — область активных исследований.