Линейная регрессия

Антон Коробейников

2 сентября 2014 г.

Линейная регрессия

- Простейшая модель обучения с учителем, предполагающая *линейную* зависимость Y от $X_1, \ldots, X_p$.
- Настоящие линии регресии редко бывают линейными <image>
- Несмотря на кажующуюся простоту, линейная регрессия очень полезна как концептуально, так и на практике

Данные о рекламе

<image here>

Данные о рекламе

Возможные вопросы:

- Есть ли связь между бюджетом на рекламу и продажами
- Насколько сильна эта связь
- Какой тип места размещения рекламы больше всего влияет на продажи
- Насколько точно мы можем предсказывать будущие продажи
- Линейна ли наблюдаемая зависимость
- Есть ли значимый совместный эффект от разных источников размещения рекламы

Обыкновенная линейная регрессия

- Мы предполагаем модель

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

где β_0 и β_1 — неизвестные константы, отвечающие за сдвиг (intercept) и наклон (slope). Также могут называться коэффициентами и параметрами. ϵ — ошибки измерений.

- При наличии некоторых оценок параметров $\hat{\beta}_0$ и $\hat{\beta}_1$ мы можем «предсказывать» будущие продажи как

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Метод наименьших квадратов

- Обозначим $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ — предсказание Y на основе X. Тогда $e_i = y_i - \hat{y}_i$ — i-й *остаток* (residual)
- Обозначим через

$$RSS = e_1^2 + \cdots + e_N^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2$$

остаточную сумму квадратов (RSS)

- Метод наименьших квадратов:

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} RSS(\beta_0, \beta_1)$$

- Можно показать, что:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Метод наименьших квадратов

- Обозначим $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ — предсказание Y на основе X. Тогда $e_i = y_i - \hat{y}_i$ — i-й *остаток* (residual)
- Обозначим через

$$RSS = e_1^2 + \cdots + e_N^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2$$

остаточную сумму квадратов (RSS)

- Метод наименьших квадратов:

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} RSS(\beta_0, \beta_1)$$

- Можно показать, что:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Метод наименьших квадратов

- Обозначим $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ — предсказание Y на основе X. Тогда $e_i = y_i - \hat{y}_i$ — i-й *остаток* (residual)
- Обозначим через

$$RSS = e_1^2 + \cdots + e_N^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2$$

остаточную сумму квадратов (RSS)

- Метод наименьших квадратов:

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} RSS(\beta_0, \beta_1)$$

- Можно показать, что:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Метод наименьших квадратов

- Обозначим $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ — предсказание Y на основе X. Тогда $e_i = y_i - \hat{y}_i$ — i-й *остаток* (residual)
- Обозначим через

$$RSS = e_1^2 + \cdots + e_N^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2$$

остаточную сумму квадратов (RSS)

- Метод наименьших квадратов:

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} RSS(\beta_0, \beta_1)$$

- Можно показать, что:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Данные о рекламе

<image> В целом линейная зависимость достаточно явно выражена. Однако, есть некоторые проблемы в левой части рисунка.

О точности оценок параметров

- Можно подсчитать дисперсию оценок параметров:

$$SE(\hat{\beta}_0)^2 = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, SE(\hat{\beta}_1)^2 = \sigma^2 \left[\frac{1}{N} + \frac{\bar{x}^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right]$$

где

$$\sigma^2 = Var(\epsilon)$$

- Дисперсию оценок можно использовать для построения доверительного интервала. Например, 95% доверительный интервал выглядит как:

$$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$$

- Откуда появилась мировая константа 2 ?

О точности оценок параметров

- Можно подсчитать дисперсию оценок параметров:

$$SE(\hat{\beta}_0)^2 = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, SE(\hat{\beta}_1)^2 = \sigma^2 \left[\frac{1}{N} + \frac{\bar{x}^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right]$$

где

$$\sigma^2 = Var(\epsilon)$$

- Дисперсию оценок можно использовать для построения доверительного интервала. Например, 95% доверительный интервал выглядит как:

$$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$$

- Откуда появилась мировая константа 2 ?

О точности оценок параметров

- Можно подсчитать дисперсию оценок параметров:

$$SE(\hat{\beta}_0)^2 = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, SE(\hat{\beta}_1)^2 = \sigma^2 \left[\frac{1}{N} + \frac{\bar{x}^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right]$$

где

$$\sigma^2 = Var(\epsilon)$$

- Дисперсию оценок можно использовать для построения доверительного интервала. Например, 95% доверительный интервал выглядит как:

$$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$$

- Откуда появилась мировая константа 2 ?

Проверка статистических гипотез

- Дисперсия оценок позволяет также проверять статистические гипотезы, сформулированные относительно коэффициентов.
- Одна из наиболее часто встречающихся гипотез — гипотеза о значимости регрессии.

  H_0: Не существует зависимости между Y и X
  H_1: Существует зависимость между Y и X
- В терминах коэффициентов:

  H_0: $\beta_1 = 0$,
  H_1: $\beta_1 \neq 0$,

  так как при $\beta_1 = 0$ модель превращается в $Y = \beta_0 + \epsilon$ и от X не зависит.

Проверка статистических гипотез

- Дисперсия оценок позволяет также проверять статистические гипотезы, сформулированные относительно коэффициентов.
- Одна из наиболее часто встречающихся гипотез — гипотеза о значимости регрессии.

  H_0: Не существует зависимости между Y и X
  H_1: Существует зависимость между Y и X
- В терминах коэффициентов:

  H_0: $\beta_1 = 0$,
  H_1: $\beta_1 \neq 0$,

  так как при $\beta_1 = 0$ модель превращается в $Y = \beta_0 + \epsilon$ и от X не зависит.

Проверка статистических гипотез

- Дисперсия оценок позволяет также проверять статистические гипотезы, сформулированные относительно коэффициентов.
- Одна из наиболее часто встречающихся гипотез — гипотеза о значимости регрессии.

  H_0: Не существует зависимости между Y и X
  H_1: Существует зависимость между Y и X
- В терминах коэффициентов:

  H_0: $\beta_1 = 0$,
  H_1: $\beta_1 \neq 0$,

  так как при $\beta_1 = 0$ модель превращается в $Y = \beta_0 + \epsilon$ и от X не зависит.

Проверка статистических гипотез

- Статистика критерия:

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)},$$

- В случае, когда верна H_0 (т.е. $\beta_1 = 0$), имеет распределение Стьюдента с $N - 2$ степенями свободы.
- Можно легко вычислить *p-значение* — вероятность случайно встретить значение статистики критерия, равное $|t|$ или больше.

Результаты для данных рекламы

Оценка точности модели

- *Остаточная ошибка* (Residual Standard Error):

$$RSE = \sqrt{\frac{1}{N-2}RSS} = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2},$$

где $RSS = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$.

- R^2 или *доля объясненной дисперсии*:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS},$$

где $TSS = \sum_{i=1}^{N}(y_i - \bar{y})^2$ — *полная сумма квадратов*.

- Можно показать, что для обыкновенной линейной регресии $R^2 = r^2$, где r — коэффициент корреляции Y и X.

Оценка точности модели

- *Остаточная ошибка* (Residual Standard Error):

$$RSE = \sqrt{\frac{1}{N-2}RSS} = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2},$$

где $RSS = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$.

- R^2 или *доля объясненной дисперсии*:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS},$$

где $TSS = \sum_{i=1}^{N}(y_i - \bar{y})^2$ — *полная сумма квадратов*.

- Можно показать, что для обыкновенной линейной регресии $R^2 = r^2$, где r — коэффициент корреляции Y и X.

Оценка точности модели

- *Остаточная ошибка* (Residual Standard Error):

$$RSE = \sqrt{\frac{1}{N-2}RSS} = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2},$$

где $RSS = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$.

- R^2 или *доля объясненной дисперсии*:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS},$$

где $TSS = \sum_{i=1}^{N}(y_i - \bar{y})^2$ — *полная сумма квадратов*.

- Можно показать, что для обыкновенной линейной регресии $R^2 = r^2$, где r — коэффициент корреляции Y и X.

Результаты для данных рекламы

Множественная регрессия

- Усложним модель, добавляя несколько независимых переменных:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

- Интерпретация β_j: средний вклад X_j в Y при условии, что все остальные независимые переменные зафиксированы.
- Данные по рекламе:

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper} + \epsilon$$

Множественная регрессия

- Усложним модель, добавляя несколько независимых переменных:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

- Интерпретация β_j: средний вклад X_j в Y при условии, что все остальные независимые переменные зафиксированы.
- Данные по рекламе:

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper} + \epsilon$$

Множественная регрессия

- Усложним модель, добавляя несколько независимых переменных:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

- Интерпретация β_j: средний вклад X_j в Y при условии, что все остальные независимые переменные зафиксированы.
- Данные по рекламе:

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper} + \epsilon$$

Интерпретация коэффициентов

- В идеальном мире независимые переменные некоррелированы: *сбалансированный дизайн*:
  - Каждый коэффициент может оцениваться и проверяться на значимость независимо от других
  - Возможна интерпретация коэффициентов в виде «изменение в признаке X_j на 1 попугай влечет за собой изменение Y на β_j попугаев в то время, как все остальные переменные зафиксированы».
- Зависимые признаки приводят к проблемам:
  - Увеличивается дисперсия оценок параметров. Иногда очень значительно.
  - Становится сложно делать какие-либо выводы: при изменении X_j меняется и все остальное.

Интерпретация коэффициентов

- В идеальном мире независимые переменные некоррелированы: *сбалансированный дизайн*:
  - Каждый коэффициент может оцениваться и проверяться на значимость независимо от других
  - Возможна интерпретация коэффициентов в виде «изменение в признаке X_j на 1 попугай влечет за собой изменение Y на β_j попугаев в то время, как все остальные переменные зафиксированы».
- Зависимые признаки приводят к проблемам:
  - Увеличивается дисперсия оценок параметров. Иногда очень значительно.
  - Становится сложно делать какие-либо выводы: при изменении X_j меняется и все остальное.

The woes

Оценивание параметров для множественной регрессии

Все аналогично одномерному случаю:

- При наличии оценок $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ мы можем строить предсказанное значение $\hat{y}$:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p.$$

- Параметры $\beta_0, \beta_1, \ldots, \beta_p$ оцениваются методом наименьших квадратов:

$$RSS = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \cdots - \hat{\beta}_p x_p \right)^2.$$

$$\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p \right) = \underset{\beta_0, \ldots, \beta_p}{\operatorname{argmin}} RSS(\beta_0, \ldots, \beta_p)$$

Оценивание параметров для множественной регрессии

Все аналогично одномерному случаю:

- При наличии оценок $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ мы можем строить предсказанное значение $\hat{y}$:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p.$$

- Параметры $\beta_0, \beta_1, \ldots, \beta_p$ оцениваются методом наименьших квадратов:

$$RSS = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \cdots - \hat{\beta}_p x_p \right)^2.$$

$$\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p \right) = \underset{\beta_0, \ldots, \beta_p}{\operatorname{argmin}} RSS(\beta_0, \ldots, \beta_p)$$

Image

Результаты множественной регресии

Некоторые важные вопросы

- Полезен ли хотя бы один из признаков $X_1, \ldots, X_p$ для предсказания результата?
- Нужны ли все переменные $X_1, \ldots, X_p$ для хорошего предсказания Y, или же можно обойтись каким-то подмножеством?
- Насколько хорошо модель объясняет данные?
- Имея набор значений признаков, какой результат регрессии мы получим и насколько он будет точен?

Некоторые важные вопросы

- Полезен ли хотя бы один из признаков $X_1, \ldots, X_p$ для предсказания результата?
- Нужны ли все переменные $X_1, \ldots, X_p$ для хорошего предсказания Y, или же можно обойтись каким-то подмножеством?
- Насколько хорошо модель объясняет данные?
- Имея набор значений признаков, какой результат регрессии мы получим и насколько он будет точен?

Некоторые важные вопросы

- Полезен ли хотя бы один из признаков $X_1, \ldots, X_p$ для предсказания результата?
- Нужны ли все переменные $X_1, \ldots, X_p$ для хорошего предсказания Y, или же можно обойтись каким-то подмножеством?
- Насколько хорошо модель объясняет данные?
- Имея набор значений признаков, какой результат регрессии мы получим и насколько он будет точен?

Некоторые важные вопросы

- Полезен ли хотя бы один из признаков $X_1, \ldots, X_p$ для предсказания результата?
- Нужны ли все переменные $X_1, \ldots, X_p$ для хорошего предсказания Y, или же можно обойтись каким-то подмножеством?
- Насколько хорошо модель объясняет данные?
- Имея набор значений признаков, какой результат регрессии мы получим и насколько он будет точен?

О значимости регрессии

О значимости признака

Отбор информативных признаков

Forward Selection

Backward Selection

Еще раз об отборе информативных признаков

Категориальные признаки в регрессии

Credit Card Data

Категориальные признаки в регрессии

Результаты регресии

Что делать, если градаций > 2?

Результаты регресии

Расширения линейных моделей

Взаимодействие

Взаимодействие в данных о рекламе

Взаимодействие и модель

Интерпретация результатов

Интерпретация результатов

Принцип иерархичности

Взаимодействие между числовыми и категориальными переменными

Нелинейности

Результаты

Обощение линейных моделей

Далее мы будем рассматривать всевозможные обобщения линейных моделей в разных вариантах и комбинациях:

Задачи классификации: Логистическая регрессия, support vector machines

Нелинейности: Сглаживание с ядрами, сплайны, обобщенные аддитивные модели. Методы ближайшего соседа

Взаимодействие: Деревья, bagging, boosting, случайные деревья

Отбор информативных признаков и регуляризация: LASSO, ridge регрессия

Обощение линейных моделей

Далее мы будем рассматривать всевозможные обобщения линейных моделей в разных вариантах и комбинациях:

Задачи классификации: Логистическая регрессия, support vector machines

Нелинейности: Сглаживание с ядрами, сплайны, обобщенные аддитивные модели. Методы ближайшего соседа

Взаимодействие: Деревья, bagging, boosting, случайные деревья

Отбор информативных признаков и регуляризация: LASSO, ridge регрессия

Обощение линейных моделей

Далее мы будем рассматривать всевозможные обобщения линейных моделей в разных вариантах и комбинациях:

Задачи классификации: Логистическая регрессия, support vector machines

Нелинейности: Сглаживание с ядрами, сплайны, обобщенные аддитивные модели. Методы ближайшего соседа

Взаимодействие: Деревья, bagging, boosting, случайные деревья

Отбор информативных признаков и регуляризация: LASSO, ridge регрессия

Обощение линейных моделей

Далее мы будем рассматривать всевозможные обобщения линейных моделей в разных вариантах и комбинациях:

Задачи классификации: Логистическая регрессия, support vector machines

Нелинейности: Сглаживание с ядрами, сплайны, обобщенные аддитивные модели. Методы ближайшего соседа

Взаимодействие: Деревья, bagging, boosting, случайные деревья

Отбор информативных признаков и регуляризация: LASSO, ridge регрессия