Линейная регрессия

Тыщенко А.Д.
Конькова А.И.

Линейная регрессия. Введение

Линейная модель:

$f(\mathbf{x}, \mathbf{w}) = w_1 x_1 + w_2 x_2 + \cdots + w_m x_m + w_0$

где:

$f(\mathbf{x}, \mathbf{w})$ – целевая функции,

$x_1, \ldots, x_m$ – признаки объекта $\mathbf{x}$,

$w_1, \ldots, w_m$ – коэффициенты регрессии,

w_0 – смещение (свободный коэффициент).

Оценки

- Обучающая выборка:
  $\mathbf{S} = \{(\mathbf{x_1}, \hat{y}_1), (\mathbf{x_2}, \hat{y}_2), \ldots, (\mathbf{x_n}, \hat{y}_n)\}$,
  $\mathbf{x_j} = (x_{j,1}, \ldots, x_{j,m})$
- Задача: найти наилучшую линейную функцию y, аппроксимирующую $\mathbf{S}$
  или найти $\mathbf{w} = (w_0, w_1, \ldots, w_m)^T$

Фиктивный признак

Часто предполагают, что объект $\mathbf{x}$ содержит в себе фиктивный признак равный 1 для представления свободного члена w_0. В этом случае формула принимает простой вид:

$$y = \langle \mathbf{w}, \mathbf{x} \rangle,$$

где $\langle \cdot, \cdot \rangle$ – скалярное произведение векторов $\mathbf{w}$, $\mathbf{x} \in \mathbb{R}^n$.

Многомерный случай

В многомерном случае формулу линейной регрессии можно переписать следующим образом:

$$\mathbf{Y} = X\mathbf{w},$$

где:
$\mathbf{Y}$ – столбец размера m (количество объектов),
X – матрица признаков размера $n \times m$ (каждая строка матрицы является вектором признаков объекта),
$\mathbf{w}$ – вектор весов размера m.

Линейная регрессионная модель - матричная форма

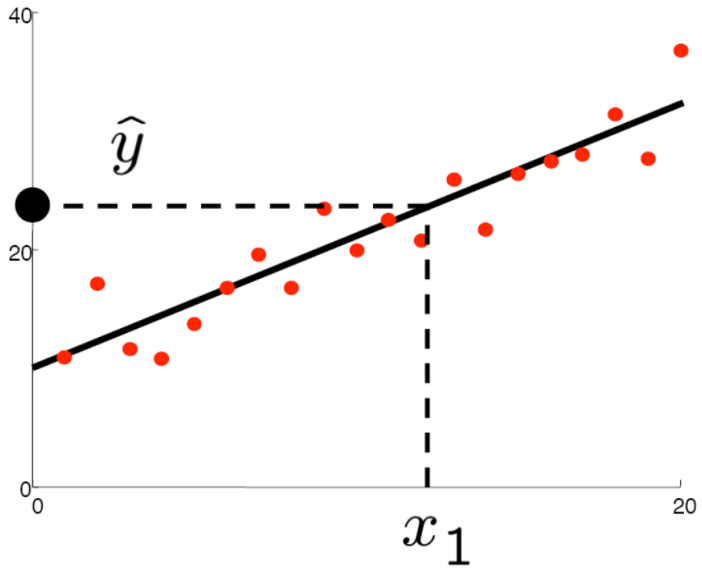
$$\hat{y}_1 = w_0 + w_1 x_{1,1} + \ldots + w_m x_{1,m} + \epsilon_1$$
$$\hat{y}_2 = w_0 + w_1 x_{2,1} + \ldots + w_m x_{2,m} + \epsilon_2 \Rightarrow \quad \mathbf{Y} = \mathbf{X}\mathbf{w} + \epsilon$$
$$\ldots$$
$$\hat{y}_n = w_0 + w_1 x_{n,1} + \ldots + w_m x_{n,m} + \epsilon_n$$

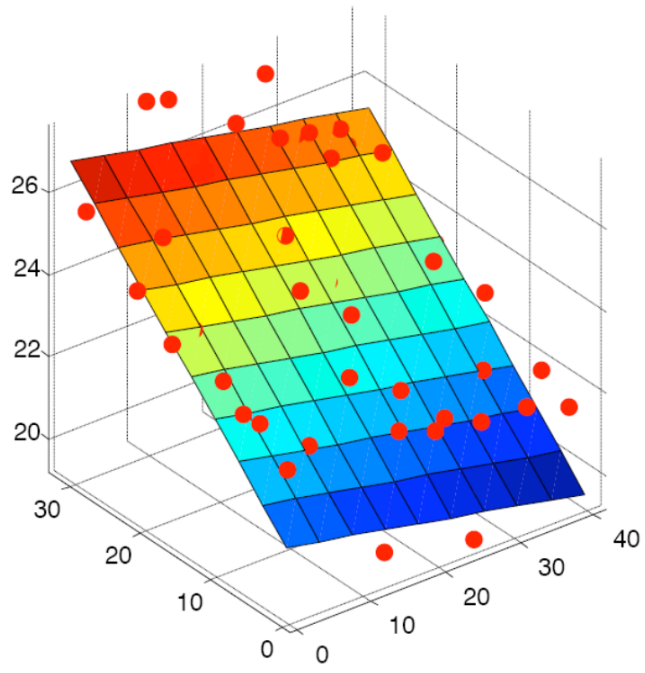
$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} \quad ,X = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,m} \\ 1 & x_{2,1} & \ldots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \ldots & x_{n,m} \end{pmatrix} \quad ,w = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_m \end{pmatrix}$$

Линейная регрессионная модель

$$\hat{y}_i = w_0 + w_1 x_{i,1} + \epsilon$$

$$\hat{y}_i = w_0 + w_1 x_{i,1} + w_1 x_{i,2} + \epsilon$$





Решение через нормальное уравнение

Для квадратичной функции потерь минимум можно найти аналитически.

Минимум достигается при:

$$\mathbf{w} = (X^T X)^{-1} X^T Y,$$

если матрица $X^T X$ обратима.

Это решение называют **нормальным уравнением**. Оно даёт оптимальные веса без итерационных методов.

Эмпирический функционал риска

$$E(\mathbf{w}) = \frac{1}{2N}\sum_{i=1}^{n}(y_i - f(x_i, w))^2 =$$

$$= \frac{1}{2N}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_{i,1} + \ldots + w_m x_{i,m}))^2 \to \min_{\mathrm{b}}$$

Матричная форма:

$$E(\mathbf{w}) = \frac{1}{2N}(\mathbf{Y} - \mathbf{X}\mathbf{w})^T(\mathbf{Y} - \mathbf{X}\mathbf{w}) \to \min_{\mathrm{b}}$$

Как найти коэффициенты $\mathbf{w}$?

Производный по всем $w_0, w_1, \ldots, w_m$

Градиентный спуск

Градиентый спуск заключается в следующих шагах:

1. Расчёт $\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}_k}$ – градиента функции потерь $E(\mathbf{w})$ – от значения параметров модели (весов $\mathbf{w}$);
2. Шаг спуска – изменение весов $\mathbf{w}$ в сторону антиградиента с некоторым коэффициентом α (он же learning rate):

$$\mathbf{w} := \mathbf{w} - \alpha \frac{\partial E(\mathbf{w})}{\partial \mathbf{w}_k};$$

3. Повторение п.1 и п.2 пока не наблюдается сходимость (изменения ошибки малы или отсутсвуют).

Роль learning rate α

- Определяет величину шага обновления весов;
- Слишком большой $\alpha \rightarrow$ дивергенция;
- Слишком маленький $\alpha \rightarrow$ медленная сходимость.

Метрики качества модели

Для оценки качества линейной регрессии используют:

- **MSE** (Mean Squared Error):

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

- **RMSE** (Root Mean Squared Error):

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$$

- **MAE** (Mean Absolute Error):

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Предпосылки линейной регрессии

При использовании линейной регрессии важно учитывать следующие допущения:

1. Линейность зависимости
2. Отсутствие мультиколлинеарности
3. Гомоскедастичность
4. Нормальное распределение ошибок

Нарушение этих предположений может привести к смещённым или нестабильным результатам.

Список литературы

1. Жерон, О. *Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow*. — М.: ДМК Пресс, 2018.
2. Пашинина, П. А., Родионов, Д. Г., Конников, Е. А. *Системное моделирование технологий анализа и обработки данных с использованием интегрированных алгоритмов машинного обучения: Учебное пособие*. — Санкт-Петербург: Университет Петра Великого, 2024.

Спасибо за внимание!