Линейная регрессия

Для заданного множества из m пар (x_i, y_i), $i = 1, \ldots, m$, значений свободной и зависимой переменной требуется построить зависимость. Назначена линейная модель

$$y_i = f(\mathbf{w}, x_i) + \varepsilon_i$$

с аддитивной случайной величиной ε. Переменные x, y принимают значения на числовой прямой $\mathbb{R}$. Предполагается, что случайная величина распределена нормально с нулевым матожиданием и фиксированной дисперсией σ_ε^2, которая не зависит от переменных x, y. При таких предположениях параметры $\mathbf{w}$ регрессионной модели вычисляются с помощью метода наименьших квадратов.

Например, требуется построить зависимость цены нарезного хлеба от времени. В таблице регрессионной выборки первая колонка — зависимая переменная (цена батона хлеба), вторая — свободная (время). Всего данные содержат 195 пар значений переменных. Данные нормированы.

1. Одномерная регерссия

Определим модель зависимости как

$$y_i = w_1 + w_2 x_i + \varepsilon_i.$$

Согласно методу наименьнах квадратов, искомый вектор параметров $\mathbf{w} = (w_1, w_2)^T$ есть решение нормального уравнения

$$\mathbf{w} = (A^T A)^{-1} A^T \mathbf{y},$$

где $\mathbf{y}$ — вектор, состоящий из значений зависимой переменной, $\mathbf{y} = (y_1, \ldots, y_m)$. Столбцы матрицы A есть подстановки значений свободной переменной $x_i^0 \mapsto a_{i1}$ и $x_i^1 \mapsto a_{i2}$, $i = 1, \ldots, m$. Матрица имеет вид

$$A = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \ldots & \ldots \\ 1 & x_m \end{pmatrix}.$$

Зависимая переменная восстанавливается по полученным весам и заданным значениям свободной переменной

$$y_i^* = w_1 + w_2 x_i,$$

иначе

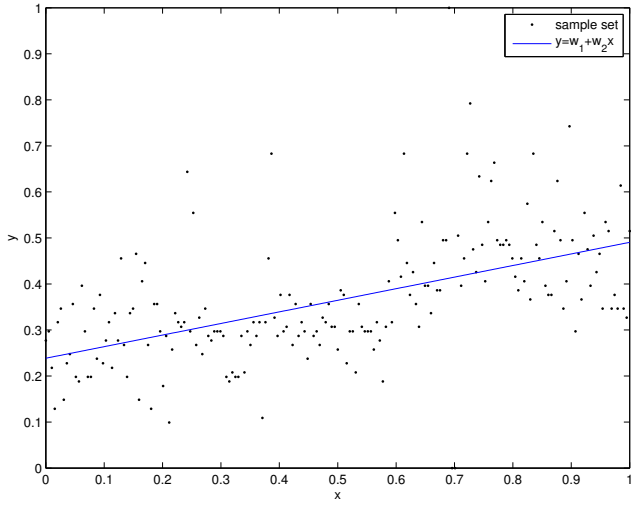
$$\mathbf{y}^* = A\mathbf{w}.$$

Для оценки качества модели используется критерий суммы квадратов регрессионых остатков, SSE — Sum of Squared Errors.

$$SSE = \sum_{i=1}^{m} (y_i - y_i^*)^2 = (\mathbf{y} - \mathbf{y}^*)^T (\mathbf{y} - \mathbf{y}^*).$$

Пример нахождения параметров модели и восстановления линейной регрессии.

```
A = [x.^0, x];          % построить матрицу подстановок, x - (m,1)-вектор
w = inv(A'*A)*(A'*y); % решить нормальное уравнение, y - (m,1)-вектор
y1 = w(1)+w(2)*x;       % восстановить зависимую переменную при заданных x
r = y-y1;               % найти вектор регрессионных остатков
err = r'*r;             % подсчитать ошибку
```



2. Полиномиальная регрессия

Пусть регрессионная модель — полином заданной степени p,

$$y_i = \sum_{j=1}^{p} w_j x_i^{j-1} + \varepsilon_i.$$

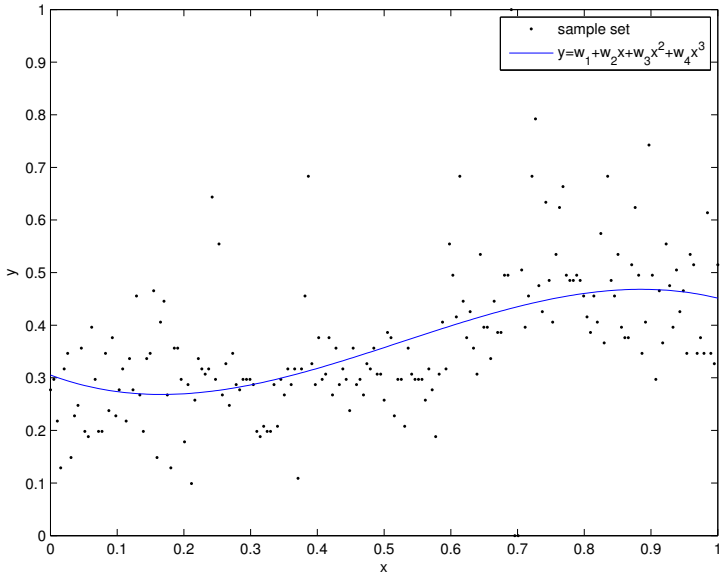
Матрица A в случае полиномиальной регрессии называется матрицей Вандермонда и принимает вид

$$A = \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^p \\ 1 & x_2 & x_2^2 & \dots & x_2^p \\ \dots & \dots & \dots & \dots & \\ 1 & x_m & x_m^2 & \dots & x_m^p \end{pmatrix}.$$

Одномерная регрессия — частный случай полиномиальной регерссии.

Пример нахождения параметров модели и восстановления полиномиальной регрессии.

```
                        % функция для построения матрицы подстановок
f = inline('[x.^0, x, x.^2, x.^3]','x');
A = f(x);               % матрица подстановок - функция свободной переменой
w = inv(A'*A)*(A'*y); % решить нормальное уравнение
y2 = A*w;               % восстановить зависимую переменную
r = y-y2;               % найти вектор регрессионных остатков
SSE = r'*r              % подсчитать ошибку
```



3. Криволинейная регрессия

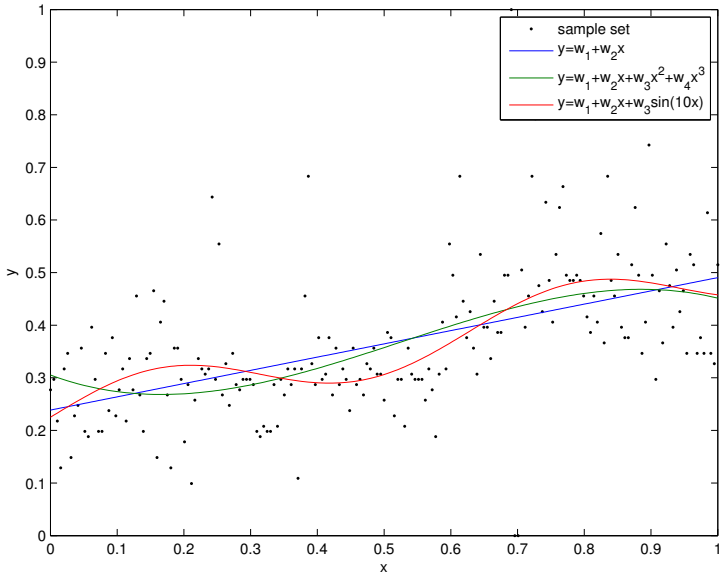
Пусть исходные признаки преобразованы с помощью некоторых заданных, в общем случае нелинейных функций $g_1, \ldots, g_n$. При этом функции не должны содержать дополнительных параметров. Функции должны быть определены на всей числовой прямой, либо, по крайней мере, на всех значениях, которые принимает свободная переменная. Матрица A в случае полиномиальной регрессии называется обобщенной матрицей Вандермонда и принимает вид

$$A = \begin{pmatrix} g_1(x_1) & \ldots & g_n(x_1) \\ g_1(x_2) & \ldots & g_n(x_2) \\ \ldots & \ldots & \ldots \\ g_1(x_m) & \ldots & g_n(x_m) \end{pmatrix}.$$

Полиномиальная регрессия — частный случай криволинейной регрессии.

Пример нахождения параметров модели и восстановления криволинейной регрессии.

```
                        % функция для построения матрицы подстановок
f = inline('[x.^0, x, sin(10*x)]','x');
A = f(x);               % матрица подстановок - функция свободной переменой
w = inv(A'*A)*(A'*y); % решить нормальное уравнение
y3 = A*w;               % восстановить зависимую переменную
r = y-y3;               % найти вектор регрессионных остатков
SSE = r'*r              % подсчитать ошибку
```



4. Смотри также

- Категория «Регрессионный анализ» на http://machinelearning.ru.
- Дрейпер Н., Смит Г. Прикладной регрессионный анализ. Издательский дом "Вильямс". 2007. 912 с.
- Стрижов В. В. Методы индуктивного порождения регрессионных моделей. М.: ВЦ РАН. 2008. 55 с. strijov08ln.pdf.
- Исходный код данного примера wiki_demo_least_squares_fit.m, регрессионная выборка bread_narez_norm.csv, вспомогательный файл plot_regression_2d.m.