Логистическая регрессия

Логистическая регрессия — частный случай обобщенной линейной регрессии. Предполагается, что зависимая переменная принимает два значения и имеет биномиальное распределение.

На практике логистическая регрессия используется для решения задач классификации с линейно-разделяемыми классами.

1. Постановка задачи восстановления логистической регрессии

Задана выборка — множество m пар $(\mathbf{x}_i, y_i)$, в которых описание i-го элемента $\mathbf{x}_i \in \mathbb{R}^n$, и значения зависимой переменной $y \in \{0, 1\}$.

Принята модель логистической регрессии, согласно которой свободные переменные $\mathbf{x}$ и зависимая переменная y связаны зависимостью

$$y = \text{logit}^{-1}(z) + \varepsilon = \frac{1}{1 + \exp(-z)} + \varepsilon,$$

где $z = b_0 + \sum_{j=1}^{n} b_j x_j$.

Примем обозначения $p_i = f(\mathbf{b}, \mathbf{x}_i)$, вектор $\mathbf{b} = [b_0, \ldots, b_n]^T$. Для удобства дальнейшего изложения обозначим выборку свободных переменных как

$$X = \begin{bmatrix} 1 & \mathbf{x}_1^T \\ \vdots & \vdots \\ 1 & \mathbf{x}_m^T \end{bmatrix}.$$

Требуется найти такое значение вектора параметров $\mathbf{b}$, которое бы доставляло минимум норме вектора невязок

$$S = \|\mathbf{y} - \mathbf{p}\|^2 = \sum_{i=1}^{m} (y_i - p_i)^2.$$

2. Алгоритм отыскания оптимальных параметров

Оптимальные параметры отыскиваются последовательно с помощью итерационного метода наименьших квадратов с использованием взвешивания элементов выборки. Приведенный ниже алгоритм основан на алгоритме Ньютона-Рафсона.

В начале работы алгоритма задаются параметры начального приближения: скаляр $b_0 = \log \frac{\tilde{y}}{1-\tilde{y}}$, где $\tilde{y} = \frac{1}{m} \sum_{i=1}^{m} y_i$ — среднее значение выборки зависимой переменной и значения $b_j = 0$ для $j = 1, \ldots, n$.

Далее итеративно повторяется следующая процедура.

1. С использованием вектора параметров $\mathbf{b}$ вычисляется переменная

$$\mathbf{z} = X\mathbf{b}.$$

2. Вычисляется восстановленное значение выборки зависимой переменной

$$p = \frac{1}{1 + \exp(\mathbf{z})}.$$

3. Вычисляется вектор значений зависимой переменной для текущего шага линейной регрессии

$$\mathbf{u} = \mathbf{z} + \frac{\mathbf{y} - \mathbf{p}}{\mathbf{w}},$$

где $\mathbf{w} = \mathbf{p}(1 - \mathbf{p})$ — вектор весов значений зависимой переменной.

4. Решается задача наименьших квадратов с взвешиванием элементов выборки. При этом больший вес приобретают те элементы, которые имеют большую невязку

$$\mathbf{b} = (X^T W X)^{-1} X^T W \mathbf{z},$$

где диагональная матрица весов $W = \mathrm{diag}(\mathbf{w})$.

Процедура останавливается после того, как норма разности векторов весов на каждой итерации не будет превышать заданную константу: $\|\mathbf{w}^{\text{next}} - \mathbf{w}^{\text{previuos}}\|^2 \leqslant \Delta_w$.

3. Пример на модельных данных

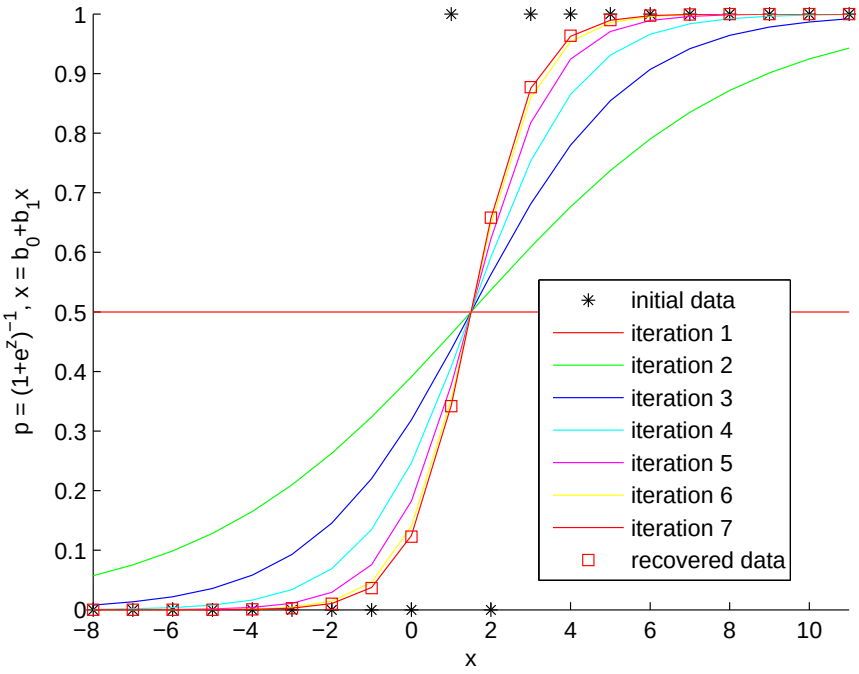
Перед началом работы алгоритма задаются начальные значения параметров

```
% 1st element, function of the mean value of y's
b0 = log(mean(y)/(1-mean(y)));
% column-vector of parameters
b = [b0 zeros(size(X,2)-1)]';
```

Итерационное вычисление параметров логистической регрессии

```
while 1==1
    % the logit^-1 variable is function of parameters
    z = X*b;
    % recover the regression
    p = 1./(1+exp(-z));
    % calculate the weights of the samples
    w = p.*(1-p);
    % calculate the dependent variable for this step of least squares
    u = z + (y-p)./w;
    % store old parameters
```

```
    b_old = b;
    % calculate new parameters with least squares
    b = inv( X'*diag(w)*X ) * X' * diag(w) * u;
    % termanate the iterations if changes of the parameters are small
if sumsqr(b - b_old) <= TolFun, break; end
end
```



4. Смотри также

- Категория «Регрессионный анализ» на http://machinelearning.ru.
- Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. Springer, 2001. 533 pages.
- Исходный код данного примера demo_logistic_regression.m.