Logistic Regression

Kairit Sirts

28.03.2014

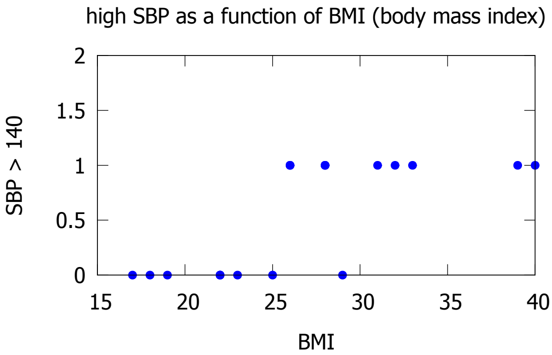
Example:

Set of health data:

Age	BMI	high SBP
30	26,3	no
31	27,1	no
32	27,6	yes
32	24,1	no
32	24,4	yes
...	...	...

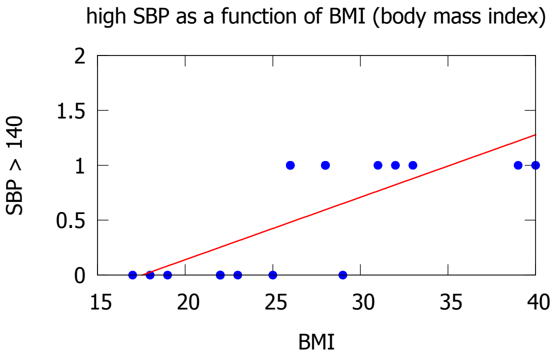
Goal: Learn to predict whether the person has a risk for high systolic blood pressure (SBP > 140) based on the age and body mass index (BMI).

Example



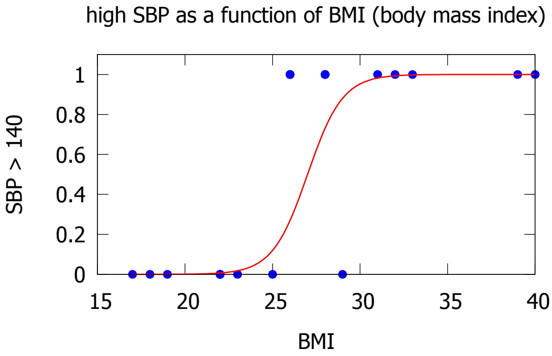
high SBP as a function of BMI (body mass index)
2
1.5
1
0.5
0
15
20
25
30
35
40
BMI
SBP > 140

Example: fitting linear regression



Example

Instead of the straight line we would like to fit a curve with range between 0 and 1.



Logistic function

Logistic function is a sigmoid function and has the formula:

$$g(z) = \frac{1}{1 + e^{-z}}$$

Note that:

$$g(z) = 0.5, \text{ if } z = 0$$
$$g(z) > 0.5, \text{ if } z > 0$$
$$g(z) < 0.5, \text{ if } z < 0$$

Derivative of logistic function

Logistic function derivative has a nice form:

$$g'(z) = g(z)(1 - g(z))$$

How do we get it?

$$\begin{aligned}
g(z) &= \frac{1}{1+e^{-z}} \\
g'(z) &= -\frac{1}{(1+e^{-z})^2}(1+e^{-z})' \quad \text{because } \left(\frac{1}{x}\right)' = -\frac{1}{x^2} \\
&= \frac{e^{-z}}{(1+e^{-z})^2} = \frac{1}{1+e^{-z}}\frac{e^{-z}}{1+e^{-z}} \quad \text{because } \left(e^{-x}\right)' = -e^{-x} \\
&= g(z)\frac{1+e^{-z}-1}{1+e^{-z}} = g(z)\left(\frac{1+e^{-z}}{1+e^{-z}} - \frac{1}{1+e^{-z}}\right) \\
&= g(z)(1-g(z))
\end{aligned}$$

Hypothesis for logistic regression

Let's change the hypothesis by using the logistic function:

$$h_{\boldsymbol{\theta}}(\mathbf{x}) = g(\boldsymbol{\theta}^T\mathbf{x}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T\mathbf{x}}}$$

where:

$$\boldsymbol{\theta}^T\mathbf{x} = \sum_{j=0}^{n} \theta_j x_j \quad \text{and by convention } x_0 = 1$$

Again, note that:

$$g(\boldsymbol{\theta}^T\mathbf{x}) = 0.5, \text{ if } \boldsymbol{\theta}^T\mathbf{x} = 0$$
$$g(\boldsymbol{\theta}^T\mathbf{x}) > 0.5, \text{ if } \boldsymbol{\theta}^T\mathbf{x} > 0$$
$$g(\boldsymbol{\theta}^T\mathbf{x}) < 0.5, \text{ if } \boldsymbol{\theta}^T\mathbf{x} < 0$$

Probabilistic interpretation

We can again give the model the probabilistic interpretation and then use the maximum likelihood principle to find the parameters:

$$P(y=1|\mathbf{x};\boldsymbol{\theta}) = h_{\boldsymbol{\theta}}(\mathbf{x}) = g(\boldsymbol{\theta}^T\mathbf{x})$$
$$P(y=0|\mathbf{x};\boldsymbol{\theta}) = 1 - h_{\boldsymbol{\theta}}(\mathbf{x}) = 1 - g(\boldsymbol{\theta}^T\mathbf{x})$$

It is possible to write these two equations compactly with a single formula:

$$P(y|\mathbf{x};\boldsymbol{\theta}) = h_{\boldsymbol{\theta}}(\mathbf{x})^y(1 - h_{\boldsymbol{\theta}}(\mathbf{x}))^{1-y}$$

When $y = 1$ then the second factor is equal to one and only the first factor counts. When $y = 0$ then the first factor becomes equal to one and only the second factor counts.

Meaning of $\boldsymbol{\theta}^T\mathbf{x}$ in logistic regression

We can take the logistic function and express it in terms of $\boldsymbol{\theta}^T\mathbf{x}$:

$$
\begin{aligned}
g(\boldsymbol{\theta}^T\mathbf{x}) &= \frac{1}{1+e^{-\boldsymbol{\theta}^T\mathbf{x}}} = \frac{1}{1+\frac{1}{e^{\boldsymbol{\theta}^T\mathbf{x}}}} = \frac{e^{\boldsymbol{\theta}^T\mathbf{x}}}{1+e^{\boldsymbol{\theta}^T\mathbf{x}}} \\
e^{\boldsymbol{\theta}^T\mathbf{x}} &= g(\boldsymbol{\theta}^T\mathbf{x})(1+e^{\boldsymbol{\theta}^T\mathbf{x}}) = g(\boldsymbol{\theta}^T\mathbf{x}) + g(\boldsymbol{\theta}^T\mathbf{x})e^{\boldsymbol{\theta}^T\mathbf{x}} \\
g(\boldsymbol{\theta}^T\mathbf{x}) &= e^{\boldsymbol{\theta}^T\mathbf{x}} - g(\boldsymbol{\theta}^T\mathbf{x})e^{\boldsymbol{\theta}^T\mathbf{x}} = e^{\boldsymbol{\theta}^T\mathbf{x}}(1-g(\boldsymbol{\theta}^T\mathbf{x})) \\
e^{\boldsymbol{\theta}^T\mathbf{x}} &= \frac{g(\boldsymbol{\theta}^T\mathbf{x})}{1-g(\boldsymbol{\theta}^T\mathbf{x})} \\
\boldsymbol{\theta}^T\mathbf{x} &= \log\frac{g(\boldsymbol{\theta}^T\mathbf{x})}{1-g(\boldsymbol{\theta}^T\mathbf{x})}
\end{aligned}
$$

This is called **log-odds**, where **odds** refers to the value where the probability of an event occurring is divided by the probability of not occurring $\left(\frac{p}{1-p}\right)$.

Likelihood

We first write down the formula for the probability of the whole data set (likelihood of the parameters):

$$\mathcal{L}(\boldsymbol{\theta}) = P(Y|\mathbf{X};\boldsymbol{\theta}) = \prod_{i=1}^{m} h_{\boldsymbol{\theta}}(\mathbf{x}_i)^{y_i}(1 - h_{\boldsymbol{\theta}}(\mathbf{x}_i))^{1-y_i}$$

As usual, we will prefer operating on log-likelihood:

$$\begin{aligned}
\ell(\boldsymbol{\theta}) = \log \mathcal{L}(\boldsymbol{\theta}) &= \log \prod_{i=1}^{m} h_{\boldsymbol{\theta}}(\mathbf{x}_i)^{y_i}(1 - h_{\boldsymbol{\theta}}(\mathbf{x}_i))^{1-y_i} \\
&= \sum_{i=1}^{m} \log h_{\boldsymbol{\theta}}(\mathbf{x}_i)^{y_i}(1 - h_{\boldsymbol{\theta}}(\mathbf{x}_i))^{1-y_i} \\
&= \sum_{i=1}^{m} \left(\log h_{\boldsymbol{\theta}}(\mathbf{x}_i)^{y_i} + \log (1 - h_{\boldsymbol{\theta}}(\mathbf{x}_i))^{1-y_i}\right) \\
&= \sum_{i=1}^{m} (y_i \log h_{\boldsymbol{\theta}}(\mathbf{x}_i) + (1 - y_i) \log (1 - h_{\boldsymbol{\theta}}(\mathbf{x}_i)))
\end{aligned}$$

Maximizing likelihood

- Now we can use the already familiar method of gradient descent to minimize the negative log-likelihood
- Or we can use the method of **gradient ascent** to maximise the log-likelihood
- The difference between gradient ascent and gradient descent is in the sign of the update step
  - For gradient descent we subtract the update:

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} \ell(\boldsymbol{\theta})$$

  - For gradient ascent we add the update:

$$\theta_j = \theta_j + \alpha \frac{\partial}{\partial \theta_j} \ell(\boldsymbol{\theta})$$

Derivative for the gradient method

- Take the derivative from the log-likelihood:

$$
\begin{aligned}
\frac{\partial}{\partial \theta_j}\ell(\boldsymbol{\theta}) &= \frac{\partial}{\partial \theta_j}\sum_{i=1}^{m}\left(y_i \log h_{\boldsymbol{\theta}}(\mathbf{x}_i) + (1-y_i)\log\left(1-h_{\boldsymbol{\theta}}(\mathbf{x}_i)\right)\right) \\
&= \sum_{i=1}^{m}\Bigg(y_i \frac{1}{h_{\boldsymbol{\theta}}(\mathbf{x}_i)}\frac{\partial}{\partial \theta_j}h_{\boldsymbol{\theta}}(\mathbf{x}_i) \\
&+ (1-y_i)\frac{1}{1-h_{\boldsymbol{\theta}}(\mathbf{x}_i)}\frac{\partial}{\partial \theta_j}(1-h_{\boldsymbol{\theta}}(\mathbf{x}_i))\Bigg) \\
&= \sum_{i=1}^{m}\Bigg(\frac{y_i h_{\boldsymbol{\theta}}(\mathbf{x}_i)(1-h_{\boldsymbol{\theta}}(\mathbf{x}_i))}{h_{\boldsymbol{\theta}}(\mathbf{x}_i)} \\
&- \frac{(1-y_i)h_{\boldsymbol{\theta}}(\mathbf{x}_i)(1-h_{\boldsymbol{\theta}}(\mathbf{x}_i))}{1-h_{\boldsymbol{\theta}}(\mathbf{x}_i)}\Bigg)\frac{\partial}{\partial \theta_j}\boldsymbol{\theta}^T\mathbf{x}_i
\end{aligned}
$$

Derivative continued ...

$$
\begin{aligned}
\frac{\partial}{\partial \theta_j}\ell(\boldsymbol{\theta}) &= \sum_{i=1}^{m} \left(y_i(1 - h_{\boldsymbol{\theta}}(\mathbf{x}_i)) - (1 - y_i)h_{\boldsymbol{\theta}}(\mathbf{x}_i)\right) x_{ij} \\
&= \sum_{i=1}^{m} \left(y_i - h_{\boldsymbol{\theta}}(\mathbf{x}_i)y_i - h_{\boldsymbol{\theta}}(\mathbf{x}_i) + h_{\boldsymbol{\theta}}(\mathbf{x}_i)y_i\right) x_{ij} \\
&= \sum_{i=1}^{m} \left(y_i - h_{\boldsymbol{\theta}}(\mathbf{x}_i)\right) x_{ij}
\end{aligned}
$$

Gradient ascent update

So the gradient ascent update for logistic regression will be:

$$\theta_j^{k+1} = \theta_j^k + \alpha \sum_{i=1}^{m} \left(y_i - h_{\boldsymbol{\theta}}(\mathbf{x}_i)\right) x_{ij}$$

for each θ_j, $j = 0 \ldots n$ simultaneously.

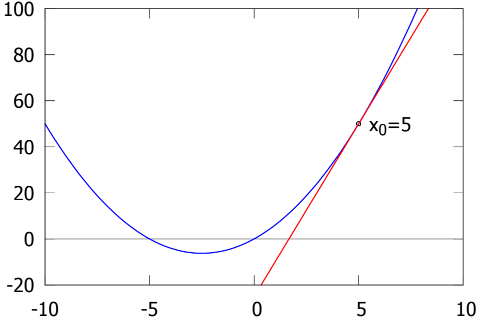
Newton's method

- Another iterative method in calculus for finding the zeroes of real-valued functions.

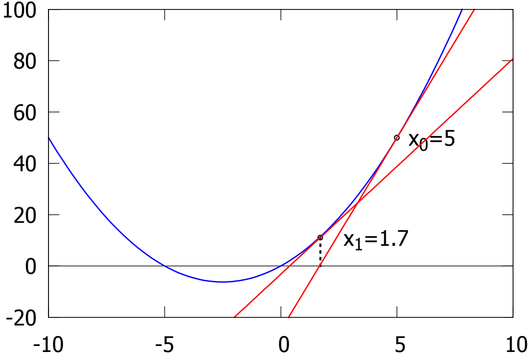
$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

- For example:

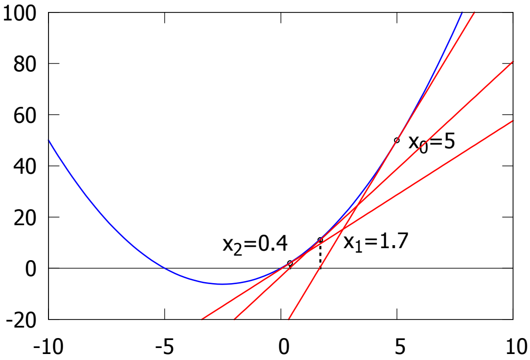
$$y = x^2 + 5x \qquad y' = 2x + 5 \qquad x_0 = 5$$



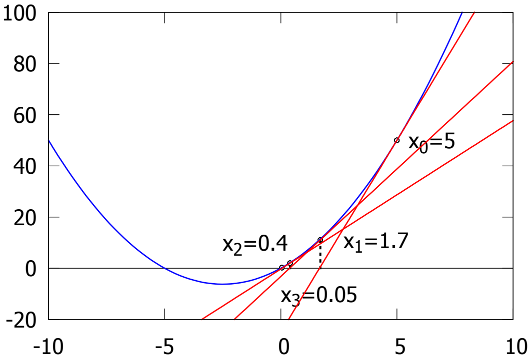
Newton's method



Newton's method



Newton's method



Newton's method in optimization

- A function is minimized if it's derivatives are 0.
- So in optimization we apply Newton's method to the derivative function:

$$\theta^{(k+1)} = \theta^{(k)} - \frac{\ell'(\theta)}{\ell''(\theta)}$$

- This is **second order** method, because it uses second derivatives.

Newton's method update rule

When $\boldsymbol{\theta}$ is a vector as we previously had:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - H^{-1}\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}),$$

where $\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})$ is the vector of partial derivatives and H is called **Hessian** and is the $(n+1)\times(n+1)$ matrix of second partial derivatives:

$$H = \begin{bmatrix} \frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\theta_0\partial\theta_0} & \cdots & \frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\theta_0\partial\theta_n} \\ \cdots & \cdots & \cdots \\ \frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\theta_n\partial\theta_0} & \cdots & \frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\theta_n\partial\theta_n} \end{bmatrix}$$

Newton's method in optimization

- Hessian must be positive definite
- This is true when the optimized objective function is convex.
- A matrix A is positive definite if $\mathbf{x}^T A\mathbf{x}$ is positive for any nonzero vector $\mathbf{x}$
- If Hessian is not positive definite then the objective function is not convex and the Newton step might not point to a decent direction.

Newton's method for logistic regression

- For Hessian we need to compute second partial derivatives:

$$
\begin{aligned}
\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} &= \frac{\partial}{\partial \theta_k} \sum_{i=1}^{m} \left(y_i - h_{\boldsymbol{\theta}}(\mathbf{x}_i)\right) x_{ij} \\
&= -\sum_{i=1}^{m} h_{\boldsymbol{\theta}}(\mathbf{x}_i)(1 - h_{\boldsymbol{\theta}}(\mathbf{x}_i)) x_{ij} x_{ik}
\end{aligned}
$$

Regularized logistic regression

- When data is linearly separable then maximum likelihood can lead to severe overfitting.
- This is because the MLE solution is obtained when $\|\boldsymbol{\theta}\| \to \infty$
- In this case the logistic sigmoid function will approach Heaviside step function and each point is classified as 0 or 1 with probability 1.
- Overfitting can be prevented by adding regularization:

$$\ell_{reg}(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$$