# Linear Regression and Logistic Regression

Viktor Hansen
Department of Computer Science
University of Copenhagen

December 3, 2016

**Abstract**

This is my submission for the second third assignment for the Machine Learning course offered at The Department of Computer Science, Uni. Copenhagen.

# Preface

Solutions for all tasks in the programming part were completed in MATLAB, and generic solutions were attempted where possible. The implementation was compiled and tested with the R2015b distribution of MATLAB. The main script is found in `code/main.m`, and running it will reproduce the results presented in this report. The output of running the program can be found in.

References to locations of source files in this report will be assumed to have `code/` implicitly prepended; that is `functions/my_function.m` is to be found in `code/function/my_function.m` etc.

# 1 Logistic Regression

## 1.1 Cross-entropy error measure

Recall that the likelihood function is defined as

$$\mathbb{P}\{\, y \,|\, \mathbf{x} \,\} = \begin{cases} h(\mathbf{x}) & \text{for } y = +1 \\ 1 - h(\mathbf{x}) & \text{for } y = -1 \end{cases} \tag{1}$$

If we rewrite (**??**) in terms of indicator functions we get

$$\mathbb{P}\{\, y \,|\, \mathbf{x} \,\} = \mathbb{1}_{y \in \{+1\}} h(\mathbf{x}) + \mathbb{1}_{y \in \{-1\}} (1 - h(\mathbf{x})) \tag{2}$$

Next, recall that maximizing the simultaneous likelihood is equivalent to minimizing the quantity on the following RHS

$$E_{\text{in}}(\mathbf{w}) = -\frac{1}{N} \ln \left( \prod_{i=1}^{N} \mathbb{P}\{\, y_i \,|\, \mathbf{x}_i \,\} \right) = \frac{1}{N} \sum_{i=1}^{N} \ln \left( \frac{1}{\mathbb{P}\{\, y_i \,|\, \mathbf{x}_i \,\}} \right) \tag{3}$$

Now, in the case that $y_i = +1$, then $\mathbb{P}\{\, y_i \,|\, \mathbf{x}_i \,\} = h(\mathbf{x}_i)$ and similarly, when $y_i = -1$, then $\mathbb{P}\{\, y_i \,|\, \mathbf{x}_i \,\} = 1 - h(\mathbf{x}_i)$. Substituting (**??**) into (**??**) means that each subterm of the summation in (**??**) can be written as

$$\ln\left(\frac{1}{\mathbb{P}\{\, y_i \,|\, \mathbf{x}_i \,\}}\right) = \ln\left(\frac{1}{\mathbb{1}_{y\in\{+1\}}h(\mathbf{x}_i) + \mathbb{1}_{y\in\{-1\}}(1-h(\mathbf{x}_i))}\right) \tag{4}$$

$$= \mathbb{1}_{y\in\{+1\}}\ln\left(\frac{1}{h(\mathbf{x}_i)}\right) + \mathbb{1}_{y\in\{-1\}}\ln\left(\frac{1}{1-h(\mathbf{x}_i)}\right) \tag{5}$$

which shows part (a), i.e.

$$\begin{aligned} E_{\text{in}}(\mathbf{w}) &= \frac{1}{N}\sum_{i=1}^{N}\ln\left(\frac{1}{\mathbb{P}\{\, y_i \,|\, \mathbf{x}_i \,\}}\right) \\ &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}_{y\in\{+1\}}\ln\left(\frac{1}{h(\mathbf{x}_i)}\right) + \mathbb{1}_{y\in\{-1\}}\ln\left(\frac{1}{1-h(\mathbf{x}_i)}\right) \end{aligned}$$

To show (b) we let $h(\mathbf{x}) = \theta(\mathbf{w}^{\mathrm{T}}\mathbf{x}) = e^{\mathbf{w}^{\mathrm{T}}\mathbf{x}}/(1+e^{\mathbf{w}^{\mathrm{T}}\mathbf{x}})$ and substitute $h$ into (**??**) so

$$\mathbb{1}_{y\in\{+1\}}\ln\left(\frac{1}{\theta(\mathbf{w}^{\mathrm{T}}\mathbf{x})}\right) + \mathbb{1}_{y\in\{-1\}}\ln\left(\frac{1}{1-\theta(\mathbf{w}^{\mathrm{T}}\mathbf{x})}\right)$$

As $1-\theta(s) = \theta(-s)$ we get

$$\mathbb{1}_{y\in\{+1\}}\ln\left(\frac{1}{\theta(\mathbf{w}^{\mathrm{T}}\mathbf{x})}\right) + \mathbb{1}_{y\in\{-1\}}\ln\left(\frac{1}{\theta(-\mathbf{w}^{\mathrm{T}}\mathbf{x})}\right) = \ln\left(\frac{1}{\theta(y_i\ \mathbf{w}^{\mathrm{T}}\mathbf{x})}\right)$$

And hence

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}\ln\left(\frac{1}{\theta(y_i\ \mathbf{w}^{\mathrm{T}}\mathbf{x})}\right) = \frac{1}{N}\sum_{i=1}^{N}\ln\left(1+e^{-y_i\mathbf{w}^{\mathrm{T}}\mathbf{x}_i}\right) \tag{6}$$

Equation (**??**) shows that minimizing the in-sample error of part a is equivalent to minimizing the one in 3.9.

## 1.2 Logistic Regression Loss Gradient

We first determine the gradient of the in-sample loss function

$$\nabla E_{\text{in}}(\mathbf{w}) = \frac{\partial}{\partial\mathbf{w}}\left[\frac{1}{N}\sum_{i=1}^{N}\ln\left(1+e^{-y_i\mathbf{w}^{\mathrm{T}}\mathbf{x}_i}\right)\right] = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\mathbf{w}}\left[\ln\left(1+e^{-y_i\mathbf{w}^{\mathrm{T}}\mathbf{x}_i}\right)\right]$$

Letting $f(x) = \ln(x)$ and $g(\mathbf{w}) = 1+e^{-y_i\mathbf{w}^{\mathrm{T}}\mathbf{x}_i}$ and applying the chain rule for gradients, we get

$$\frac{\partial}{\partial\mathbf{w}}\left[\ln\left(1+e^{-y_i\mathbf{w}^{\mathrm{T}}\mathbf{x}_i}\right)\right] = \nabla(f\circ g)(\mathbf{w}) = f'(g(\mathbf{w}))\nabla g(\mathbf{w})$$

We have $f'(x) = \frac{\mathrm{d}}{\mathrm{d}x}(\ln(x)) = \frac{1}{x}$, so $f'(g(\mathbf{w})) = 1/(1 + e^{-y_i \mathbf{w}^{\mathrm{T}}\mathbf{x}_i})$ and

$$\nabla g(\mathbf{w}) = \frac{\partial}{\partial \mathbf{w}}\left[1 + e^{-y_i \mathbf{w}^{\mathrm{T}}\mathbf{x}_i}\right] = \frac{\partial}{\partial \mathbf{w}}\left[e^{-y_i \mathbf{w}^{\mathrm{T}}\mathbf{x}_i}\right] = e^{-y_i \mathbf{w}^{\mathrm{T}}\mathbf{x}_i} \cdot (-y_i \mathbf{x}_i)$$

The last step follows from the chain rule. Thus

$$\begin{aligned} f'(g(\mathbf{w}))\nabla g(\mathbf{w}) &= \frac{-y_i \mathbf{x}_i e^{-y_i \mathbf{w}^{\mathrm{T}}\mathbf{x}_i}}{1 + e^{-y_i \mathbf{w}^{\mathrm{T}}\mathbf{x}_i}} \\ &= \frac{-y_i \mathbf{x}_i (e^{-y_i \mathbf{w}^{\mathrm{T}}\mathbf{x}_i})/(e^{-y_i \mathbf{w}^{\mathrm{T}}\mathbf{x}_i})}{(1 + e^{-y_i \mathbf{w}^{\mathrm{T}}\mathbf{x}_i})/(e^{-y_i \mathbf{w}^{\mathrm{T}}\mathbf{x}_i})} \\ &= \frac{-y_i \mathbf{x}_i}{1 + e^{y_i \mathbf{w}^{\mathrm{T}}\mathbf{x}_i}} \end{aligned}$$

And

$$\nabla E_{\mathrm{in}}(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N} \frac{-y_i \mathbf{x}_i}{1 + e^{y_i \mathbf{w}^{\mathrm{T}}\mathbf{x}_i}} = \frac{1}{N}\sum_{i=1}^{N} -y_i \mathbf{x}_i \, \theta(-y_i \mathbf{w}^{\mathrm{T}}\mathbf{x}_i)$$

To show that a misclassified sample contributes more to the gradient than, we consider any sample $(\mathbf{x}, y)$. The contribution to the gradient from this sample is given by the term $\|-y\, \mathbf{x}\, \theta(-y\mathbf{w}^{\mathrm{T}}\mathbf{x})\| = \|x\| \cdot |-y\, \theta(-y\, \mathbf{w}^{\mathrm{T}}\mathbf{x})|$. Note that the magnitude of the contribution to the gradient is proportional to $|-y\, \theta(-y\, \mathbf{w}^{\mathrm{T}}\mathbf{x})| = |\theta(-y\, \mathbf{w}^{\mathrm{T}}\mathbf{x})|$. We thus need to consider the cases in which a sample is misclassified as a -1 and a +1, and show that the contribution to the gradient is larger than those of the correctly classified ones in both cases. That is

$$|\theta(\mathbf{w}^{\mathrm{T}}\mathbf{x})| > |\theta(-\mathbf{w}^{\mathrm{T}}\mathbf{x})| = |1 - \theta(\mathbf{w}^{\mathrm{T}}\mathbf{x})|$$

when the sample is misclassified as $-1$ and

$$|\theta(\mathbf{w}^{\mathrm{T}}\mathbf{x})| = |1 - \theta(\mathbf{w}^{\mathrm{T}}\mathbf{x})| > |\theta(\mathbf{w}^{\mathrm{T}}\mathbf{x})|$$

In the first case, the sample is misclassified as $+1$, meaning that $\theta(\mathbf{w}^{\mathrm{T}}\mathbf{x}) > \frac{1}{2}$, and thus the contribution is greater than that of the correct classification, as $|\theta(\mathbf{w}^{\mathrm{T}}\mathbf{x})| > |1 - \theta(\mathbf{w}^{\mathrm{T}}\mathbf{x})|$.

In the second case, the sample is misclassified as $-1$ meaning that $\theta(\mathbf{w}^{\mathrm{T}}\mathbf{x}) < \frac{1}{2}$, and by the same logic as before, the contribution is greater than of the correct classification.

## 1.3 Logistic Regression Implementation

The logistic regression method was implemented in `logistic_regression.m`. An initial choice of $\eta_0 = 1.0$ is updated so $\eta_{t+1} = \eta_t \|\nabla E_{\mathrm{in}}\|$ for at most 10000 iterations of gradient descent. The gradient descent loop terminates when $\nabla E_{\mathrm{in}}(\mathbf{w}) < 0.0001$. Initial values for $\mathbf{w}$ were drawn randomly from a normal distribution with $\mu = 0$ and $\sigma = 0.01$ as recommended by the book. Classification with the weight vector $\mathbf{w} = [8.9595, -159.8700]^{\mathrm{T}}$ yields a training error reported by the program of 0.0161 and a test error of 0.0385.

# 2 Linear Least Squares

We wish to fit the model, $f(x) = ax^2 + bx$, of a parabola passing through the origin.

1. Linear least squares was used in order to determine the projection of $\mathbf{w} = [a\ b]^{\mathrm{T}}$ onto the column space of $X$. The estimated model parameters were $[a\ b]^{\mathrm{T}} = [-0.96\ \ 9.76]^{\mathrm{T}}$.

2. In order to estimate the distance from the cannon where Baron von Mnchhausen will fall down, we solve $0 = -0.96x^2 + 9.76x = x \cdot (-0.96x + 9.76)$. That is, the equation is satisfied when $x = 0 \ \vee\ -0.96x + 9.76 = 0$. We are only interested in the second solution, however, as this is the one that determines where the Baron will land. Solving for $x$ yields $x = \frac{9.76}{0.96} = 10.17$.

3. The plot of the Barons trajectory can be seen in Figure **??**.

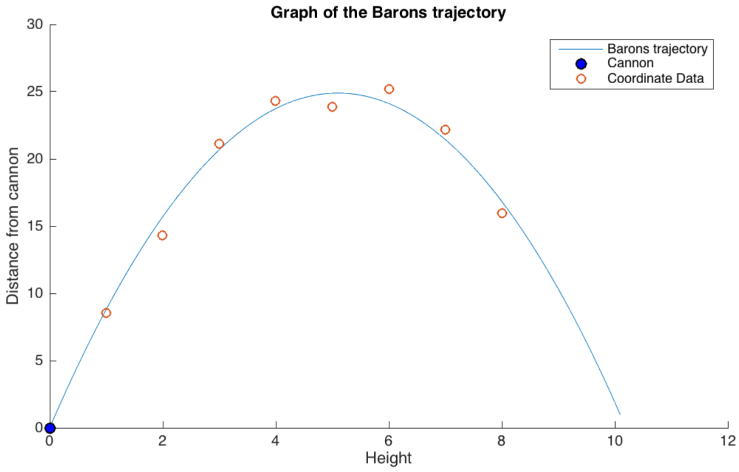

Figure 1: Plot of the Baron's trajectory, reported data and position of the cannon.