Machine Learning, Tutorial 4
Universität Bern

Simon Jenni (simon.jenni@inf.unibe.ch)

Logistic Regression

In Logistic Regression we want to maximize the following log likelihood w.r.t. the model parameters θ

$$\ell(\theta) = \sum_{i=1}^{m} y^{(i)} \log h_\theta\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_\theta\left(x^{(i)}\right)\right), \tag{1}$$

where the hypothesis function h is given by $h_\theta(x) = g\left(\theta^T x\right) = \frac{1}{1+e^{-\theta^T x}}$.
Compute the Hessian of the log likelihood.

Solution:
The Hessian matrix is defined as $H_{ij}(\theta) = \frac{\partial^2 \ell(\theta)}{\partial\theta_i \partial\theta_j}$, i.e., it is the matrix of second partial derivatives. Let us first compute the first partial derivates of $\ell(\theta)$. Remember that $g(z)$ is the sigmoid function with derivative $g'(z) = g(z)(1 - g(z))$.
The first partial derivates of $\ell(\theta)$ are given by (see lecture notes):

$$\frac{\partial}{\partial\theta_j}\ell(\theta) = \sum_{i=1}^{m}(y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)}. \tag{2}$$

Let us now compute the Hessian, i.e., the second partial derivates of $\ell(\theta)$:

$$\begin{aligned}
\frac{\partial^2}{\partial\theta_k \partial\theta_j}\ell(\theta) &= \frac{\partial}{\partial\theta_k}\left(\sum_{i=1}^{m}(y^{(i)} - g(\theta^T x^{(i)}))x_j^{(i)}\right) \\
&= -\sum_{i=1}^{m} g(\theta^T x^{(i)})(1 - g(\theta^T x^{(i)}))\frac{\partial}{\partial\theta_k}(\theta^T x^{(i)})x_j^{(i)} \\
&= -\sum_{i=1}^{m} h_\theta(x^{(i)})(1 - h_\theta(x^{(i)}))x_k^{(i)} x_j^{(i)}
\end{aligned} \tag{3}$$

We can right this result more succinctly in vector form:

$$H(\theta) = \sum_{i=1}^{m} h_\theta(x^{(i)})(h_\theta(x^{(i)}) - 1)x^{(i)}(x^{(i)})^T \tag{4}$$

Generalized Linear Models

Consider the Laplace distribution with PDF

$$p(y; \mu, \lambda) = \frac{1}{2\lambda} \exp\left(-\frac{|y-\mu|}{\lambda}\right), \tag{5}$$

where $\mu \in \mathbb{R}$ is a location parameter and $\lambda > 0$ is a scale parameter.
Show that the Laplace distribution, if parametrized only on λ (i.e., with fixed μ), is part of the exponential family.

Solution:
The exponential family is defined as all the distributions that can be written in the form

$$p(y; \eta) = b(y) \exp(\eta^T T(y) - a(\eta)) \tag{6}$$

where η is called natural parameter, $T(y)$ is called sufficient statistic and $a(\eta)$ is called log partition function.

The Laplace distribution, if parametrized only on λ, with μ fixed, can be written in the above form:

$$p(y; \lambda) = \exp\left(-\frac{1}{\lambda}|y-\mu|\right)\frac{1}{2\lambda} = \exp\left(-\frac{1}{\lambda}|y-\mu| - \log(2\lambda)\right), \tag{7}$$

that is, $b(y) = 1$, $\eta = -\frac{1}{\lambda}$, $T(y) = |y-\mu|$ and $a(\eta) = \log\left(-\frac{2}{\eta}\right) = \log(2\lambda)$.

Naive Bayes

Consider a binary classification with one binary output y and two binary features x_1 and x_2. The Naive Bayes classifier assumes the following distribution for a pair:

$$p\left(y, x_1, x_2\right) = p\left(x_1|y\right) p\left(x_2|y\right) p(y). \tag{8}$$

Let the values of the probabilities be:

$$\begin{aligned} p(y=0) &= 0.5 & p(y=1) &= 0.5 \\ p\left(x_1=1|y=0\right) &= 0.9 & p\left(x_1=1|y=1\right) &= 0.2 \\ p\left(x_2=1|y=0\right) &= 0.5 & p\left(x_2=1|y=1\right) &= 0.5 \end{aligned}$$

1. What are the simplifying assumptions in the Naive Bayes model?

   Solution:
   The features x_i are assumed to be conditionally independent given the target y. For example $p(x_1|y, x_2) = p(x_1|y)$.

2. Compute the posterior values $p(y=1|x_1=1, x_2=1)$ and $p(y=0|x_1=1, x_2=1)$. How would you classify an example with $x_1 = 1$ and $x_2 = 1$?

   Solution:
   Using Bayes rule we have:

   $$\begin{aligned} p(y=1|x_1=1, x_2=1) &\propto p\left(x_1=1|y=1\right) p\left(x_2=1|y=1\right) p(y=1) = 0.2 * 0.5 * 0.5 \\ p(y=0|x_1=1, x_2=1) &\propto p\left(x_1=1|y=0\right) p\left(x_2=1|y=0\right) p(y=0) = 0.9 * 0.5 * 0.5 \end{aligned}$$

   Therefore we get $p(y=1|x_1=1, x_2=1) = 0.2/1.1 \approx 0.18$

   The example should be classified as $y = 0$