Machine Learning
Exercise 3

Marc Toussaint
Machine Learning & Robotics lab, U Stuttgart
Universitätsstraße 38, 70569 Stuttgart, Germany

April 21, 2016

1 Max-likelihood estimator

Consider data $D = \{(y_i)\}_{i=1}^n$ that consists only of labels $y_i \in \{1, .., M\}$. You want to find a probability distribution $p \in \mathbb{R}^M$ with $\sum_{k=1}^M p_k = 1$ that maximizes the data likelihood. The solution is really simple and intuitive: the optimal probabilities are

$$p_k = \frac{1}{n} \sum_{i=1}^n [y_i = k]$$

where $[expr]$ equals 1 if $expr = true$ and 0 otherwise. So the sum just counts the occurances of $y_i = k$, which is then normalized. Derive this from first principles:

a) Understand (=be able to explain every step) that under i.i.d. assumptions

$$P(D|p) = \prod_{i=1}^n P(y_i|p) = \prod_{i=1}^n p_{y_i}$$

and

$$\log P(D|p) = \sum_{i=1}^n \log p_{y_i} = \sum_{i=1}^n \sum_{k=1}^M [y_i = k] \log p_k = \sum_{k=1}^M n_k \log p_k \ , \quad n_k = \sum_{i=1}^n [y_i = k]$$

b) Provide the derivative of

$$\log P(D|p) + \lambda \Big(1 - \sum_{k=1}^M p_k\Big)$$

w.r.t. p and λ

c) Set both derivatives equal to zero to derive the optimal parameter p^*.

2 Log-likelihood gradient and Hessian

Consider a binary classification problem with data $D = \{(x_i, y_i)\}_{i=1}^n$, $x_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$. We define

$$f(x) = x^\top \beta \tag{1}$$

$$p(x) = \sigma(f(x)) \ , \quad \sigma(z) = 1/(1 + e^{-z}) \tag{2}$$

$$L(\beta) = -\sum_{i=1}^n \Big[y_i \log p(x_i) + (1 - y_i) \log[1 - p(x_i)] \Big] \tag{3}$$

where $\beta \in \mathbb{R}^d$ is a vector. (NOTE: the $p(x)$ we defined here is a short-hand for $p(y = 1|x)$ on slide 03:10.)

a) Compute the derivative $\frac{\partial}{\partial \beta} L(\beta)$. Tip: use the fact $\frac{\partial}{\partial z} \sigma(z) = \sigma(z)(1 - \sigma(z))$.

b) Compute the 2nd derivative $\frac{\partial^2}{\partial \beta^2} L(\beta)$.