Machine Learning
Exercise 3

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

May 7, 2014

1 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 03:15.)

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 $\frac{\partial^2}{\partial \beta^2} L(\beta)$.

2 Logistic Regression

On the course webpage there is a data set `data2Class.txt` for a binary classification problem. Each line contains a data entry (x, y) with $x \in \mathbb{R}^2$ and $y \in \{0, 1\}$.

a) Compute the optimal parameters β and the mean neg-log-likelihood ($\frac{1}{n} L(\beta)$) of logistic regression using linear features. Plot the probability $p(y = 1 \,|\, x)$ over a 2D grid of test points.

Useful gnuplot commands:

```
splot [-2:3][-2:3][-3:3.5] 'model' matrix \
   us ($1/20-2):($2/20-2):3 with lines notitle
plot [-2:3][-2:3] 'data2Class.txt' \
   us 1:2:3 with points pt 2 lc variable title 'train'
```

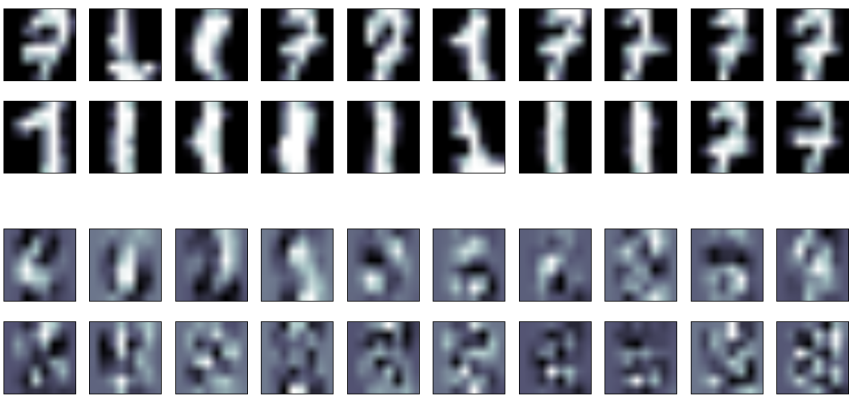

Figure 1: The top figure displays data examples of two classes of digits. The bottom figure displays the 20 first principle components of the whole (10 digit) data set – these were uses as inputs in the file `xdigit_pcs.txt`

b) Compute and plot the same for quadratic features.

3 Handwritten Digit Classification (optional)

On the course webpage there is a data set in two files, `digit_pcs.txt` and `digit_label.txt`, the first containing the inputs x_i in each row, the second the label $y \in \{0, 1\}$ in each row. This data are handwritten digits, encoded using PCA components (explained later in the lecture), as illustrated in Figure 1.

Use the same code as above to learn a binary classifier on this data. What is the mean neg-log-likelihood you achieve with linear and with quadratic features? What the correct classification rate?

For further information on how this data was generated, see
`http://..../teaching/data/LogReg_digits_PCA_by_Stefan_Otte.pdf`.