Exercises for
Pattern Recognition
Peter Fischer, Shiyang Hu
Assignment 2, 28/31.10.2014

SOLUTION

General Information:

Exercises (1 SWS):	Tue 12:15 – 13:45 (0.154-115) and Fri 08:15 – 09:45 (0.151-115)
Certificate:	Oral exam at the end of the semester
Contact:	peter.fischer@fau.de, shiyang.hu@fau.de

Maximum Likelihood Estimation

Exercise 1 Let $x_1 \dots x_k$ be a set of observations according to the exponential density

$$p(x;\lambda) = \lambda \exp(-\lambda x) \text{ for } x > 0.$$

The observed samples are considered i.i.d. (independent and identically distributed).
Draw exponential distribution on the board
Applications of the pdf: time between events that happen with constant rate (e.g. radioactive decay, failures in a technical system, ...)

(a) Derive the log-likelihood function $L(\lambda)$ for the parameter λ based on a given set of observations.
$L(\lambda) = \log\left(\prod_{i=1}^{k} p(x_i)\right) = \sum_{i=1}^{k} \log\left(p(x_i)\right) = \sum_{i=1}^{k} \log\left(\lambda\right) - \lambda \sum_{i=1}^{k} x_i = k\log(\lambda) - \lambda \sum_{i=1}^{k} x_i$

(b) Determine the Maximum Likelihood estimate for λ.
Necessary condition for a maximum $\frac{dL}{d\lambda} = 0$
$\frac{dL}{d\lambda} = \frac{d}{d\lambda}\left(k\log(\lambda) - \lambda\sum_{i=1}^{k} x_i\right) = \frac{k}{\lambda} - \sum_{i=1}^{k} x_i \stackrel{!}{=} 0$
$\lambda = \frac{k}{\sum_{i=1}^{k} x_i}$

Exercise 2 Create a logistic regression classifier for the toolbox. Assume a decision boundary that is affine in the original variables $F(\boldsymbol{x}) = \boldsymbol{\theta}^T \boldsymbol{x}$, where $\boldsymbol{x} = (x_1, x_2, \dots, 1)^T$. Create a new `m`-file, and modify `Classification.txt` and `contents.m`.
Repetition: Logistic Regression
Discriminative model $\rightarrow$ model posterior probabilities directly

Sigmoid function: $g(x) = \frac{1}{1+e^{-x}}$ and $1 - g(x) = \frac{1}{1+e^{x}}$

Assume $p(y = 1|\boldsymbol{x})$ can be modeled as $p(y = 1|\boldsymbol{x}) = \frac{1}{1+e^{-F(\boldsymbol{x})}} = g(F(\boldsymbol{x}))$ with $F(\boldsymbol{x})$ as defined above.

Thus $p(y=0|\boldsymbol{x}) = 1 - p(y=1|\boldsymbol{x}) = 1 - g(F(\boldsymbol{x})) = \frac{1}{1+e^{F(\boldsymbol{x})}}$
For classification, this means $y = 0$ iff $p(y=0|\boldsymbol{x}) > 0.5$

(a) What are the training formulas for the logistic regression?
Iterative maximization of the log-likelihood function

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \left(\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \mathcal{L}(\boldsymbol{\theta}^{(k)}) \right)^{-1} \frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L}\left(\boldsymbol{\theta}^{(k)}\right) \tag{1}$$

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left(y_i - g(\boldsymbol{\theta}^T \boldsymbol{x}_i) \right) \boldsymbol{x}_i \tag{2}$$

$$\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \mathcal{L}(\boldsymbol{\theta}) = -\sum_{i=1}^{N} g(\boldsymbol{\theta}^T \boldsymbol{x}_i) \left(1 - g(\boldsymbol{\theta}^T \boldsymbol{x}_i)\right) \boldsymbol{x}_i \boldsymbol{x}_i^T \tag{3}$$

(b) Implement the training step using the Newton-Raphson algorithm. Use the modeled posterior probabilities to compute the classification result.
See LinearLogisticRegression.m

(c) The shape of the decision boundary is linear. What does this imply for the class-conditional densities? How can you achieve nonlinear decision boundaries?
It implies nothing for the TRUE class-conditional densities. There are many distributions that can be linearly separated.
In logistic regression, we do not model the class-conditional densities. Instead we directly model the posterior distribution. This is a discriminative approach. This approach can shown to be optimal for some class-conditional densities, e.g. exponential family distribution with equal dispersion.