Selected Topics in Probabilistic Data Modelling

November 2010

Lehel Csató
Faculty of Mathematics and Informatics
Babeş–Bolyai University, Cluj-Napoca
lehel.csato@cs.ubbcluj.ro
www.cs.ubbcluj.ro/~csatol

1. **Boston Housing**
   Statisticians collected data from the Boston area to search for a **formula** giving the – approximated – price of an estate depending on attributes like: *crime rate* (CRIM), *prop. of resid. land* (ZN), *prop. of business*, *etc.* .[1] Due to problems met when assessing some variables – including the price of the house – the data-set is quite noisy.

   Your task is to build a model that is successful in predicting the house price. You should use a generalised linear model with the following setup:

   $$\boldsymbol{x}_n = [x_{n1}, x_{n2}, \ldots, x_{nd}], \qquad y_n = \text{declared price}, \quad \boldsymbol{y} = [y_1, \ldots, y_N]^T$$
   $$f(x_n|\boldsymbol{\theta}) = \sum_{k=1}^{K} \theta_k \phi_k(\boldsymbol{x}_n) = \boldsymbol{\phi}_n \boldsymbol{\theta} \qquad \text{with } \phi_k \text{ a function base of your choice}$$
   $$\text{with the notations:} \qquad \boldsymbol{\theta} \stackrel{\text{def}}{=} [\theta_1, \ldots, \theta_K]^T, \quad \boldsymbol{\phi}_n = [\phi_1(\boldsymbol{x}_n), \ldots, \phi_K(\boldsymbol{x}_n]$$

   Consider using the M.A.P. procedure where the solution is:

   $$\boldsymbol{\theta}^* = \left(\boldsymbol{\Phi}^T \boldsymbol{\Phi} + \frac{1}{\sigma_0^2} \boldsymbol{I}_K \right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{y}$$

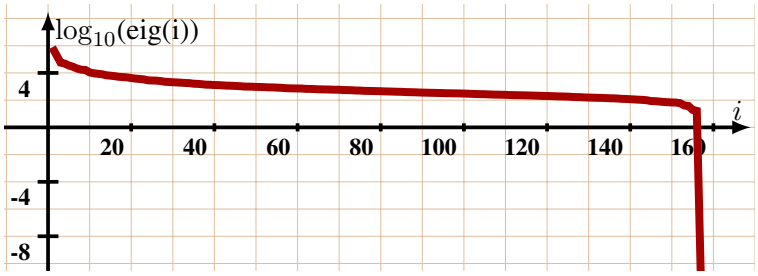
   To assess the **error of the model**, use the *leave-one-out procedure*, *i.e.* design a testing procedure that measures the error by:

   (a) Using a data set with *a single missing element* $(\boldsymbol{x}_n, y_n)$, denote it with $\mathcal{D}^{-n}$;
   (b) Computing the best model $\boldsymbol{\theta}^*_{-n}$ based on $\mathcal{D}^{-n}$;
   (c) Compute the square error for the data left-out: $LOO(n) = (y_n - f(\boldsymbol{x}_n|\boldsymbol{\theta}^*_{-n}))^2$.

   and finally compute $LOO(\mathcal{D}) = \sum_n LOO(n)$

   Analyse the results of the experiments: display the errors as a function of the *number of basis functions* and the *a-priori distribution of the parameters*, σ_0^2.

2. **Yale faces** [2]
   The Yale face dataset contains 165 images of subjects in different circumstances. The images are of size $[243, 320]$, thus the number of attributes for each observation is 77760 and there are altogether 165 images.
   The plotted eigenvectors – on the right – indicate that there are close to 160 basis elements needed to *compactly* represent the data-set.



   (a) Implement procedures to reduce this number by modifying the images with a minimum "loss" of information and such that the eigenvectors will drop much quicker – making possible the implementation of a simple dimensionality reduction technique. (hint: try to remove – algorithmically – "empty" zones from the top, bottom, left, right. Since the regions are shaded, you should detect first the position of the *e.g.* nose-eyes)
   (b) Label the data as **those-who-wear-glasses** and **those-who-not** and use the leave-one-out (LOO) procedure defined before to test whether the **original data-set** or the **reduced one** produces better results.

   You can access the Yale data-set at `http://www.cs.ubbcluj.ro/~csatol/prob_datamin/data/yale.mat` and additional M-files are available (like visualising `vis(yale,ind,[243,320])`) in the `http://www.cs.ubbcluj.ro/~csatol/prob_datamin/matlab` directory.

[1] See `http://lib.stat.cmu.edu/S/Harrell/data/descriptions/boston.html` for the data and the full description of the variables. The data itself is *e.g.* at `http://lib.stat.cmu.edu/datasets/boston`.
[2] Available from: `http://cvc.yale.edu/projects/yalefaces/yalefaces.html`