Machine Learning HW 1

Victor Manuel Garca Rosales

22/08/2017

1. Pen and Paper

Problem 1

Let us consider the regression model

$$y_n = \theta^T \mathbf{x}_n + \eta_n, \quad n = 1, 2, ..., N$$

Where the noise samples $\eta = [\eta_1, ..., \eta_N]^T$ come from a zero mean Gaussian random vector, with covariance matrix Σ_η. If $X = [\mathbf{x}_1, ..., \mathbf{x}_N]^T$ stands for the input matrix, and $\mathbf{y} = [y_1, ..., y_N]^T$, then show that,

$$\hat{\theta} = (X^T \Sigma_\eta^{-1} X) X^T \Sigma_\eta^{-1} \mathbf{y}$$

is an efficient estimate.

Notice, here, that the previous estimate coincides with the ML one. Moreover, bear in mind that in the case where $\Sigma_\eta = \sigma^2 \mathbf{I}$ then the ML estimate becomes equal to the LS one.

Result:

Our regression model could be expressed as, considering a regression matrix model and a probability previously defined:

$$p(\mathbf{y} - \mathbf{X}\theta) = \mathcal{N}(\mathbf{y} - \mathbf{X}\theta | 0, \Sigma) = \frac{1}{2\pi^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\big\{ -\frac{1}{2}(\mathbf{y} - \mathbf{X}\theta)^T \Sigma^{-1} (\mathbf{y} - \mathbf{X}\theta) \big\}$$

$$\mathbf{y} = \mathbf{X}\theta + \eta$$

$$p(\eta) = \mathcal{N}(\eta | \mu, \Sigma) \rightarrow \mathcal{N}(\eta | 0, \Sigma)$$

By using the logarithm we can simplify the optimization problem and therefore we are going to have the following problem to minimize

$$\min_\theta -\ln\big\{\mathcal{N}(\mathbf{y} - \mathbf{X}\theta | 0, \Sigma)\big\} = \min_\theta \big[\frac{1}{2}(\mathbf{y} - \mathbf{X}\theta)^T \Sigma^{-1} (\mathbf{y} - \mathbf{X}\theta)\big]$$

We can minimize it by taking the derivative and then equaling it to zero and solving for θ ,

$$\frac{\partial}{\partial\theta}\ln\{\mathcal{N}(\mathbf{y}-\mathbf{X}\theta|0,\Sigma)\} = \frac{\partial}{\partial\theta}\left[\frac{1}{2}(\mathbf{y}-\mathbf{X}\theta)^T\Sigma^{-1}(\mathbf{y}-\mathbf{X}\theta)\right] = 0$$

$$\frac{1}{2}\left(-2\mathbf{X}^T\mathbf{\Sigma^{-1}}(\mathbf{y}-\mathbf{x}\theta)\right) = 0$$

$$\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X}\theta = \mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{y}$$

$$\theta = (\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{y}$$

Problem 2

Show that solving the ask

$$\textit{minimize} \quad L(\theta,\lambda) = \sum_{n=1}^{N}\left(y_n - \theta_0 - \sum_{i=1}^{l}\theta_i x_{ni}\right)^2 + \lambda\sum_{i=1}^{l}|\theta_i|^2,$$

is equivalent with minimizing

$$\textit{minimize} \quad L(\theta,\lambda) = \sum_{n=1}^{N}\left((y_n - \bar{y}) - \sum_{i=1}^{l}\theta_i(x_{ni} - \bar{x}_i)\right)^2 + \lambda\sum_{i=1}^{l}|\theta_i|^2,$$

and the estimate of θ_0 is given by

$$\hat{\theta}_0 = \bar{y} - \sum_{i=1}^{l}\hat{\theta}_i\bar{x}_i$$

Result:

First we need to differentiate w.r.t. θ_0, afterwards we will equal it to zero, to get the maximum.

$$\sum_{n=1}^{N} y_n + N\theta + \sum_{i=1}^{l}\theta_i\sum_{n=1}^{N}x_{ni} = 0$$

$$\theta_0 = \frac{1}{N}\left(\sum_{n=1}^{N} y_n - \sum_{i=1}^{l}\theta_i\sum_{n=1}^{N}x_{ni}\right)$$

Finally we get the estimate of θ_0 and thus solving the above stated problem

$$\theta_0 = \bar{y} - \sum_{i=1}^{l}\theta_i\bar{x}$$

Problem 3

A classifier is said to be a piecewise linear machine if its discriminant functions have the form

$$g_i(\mathbf{x}) = \max_{j=1,\ldots,n_i} g_{ij}(\mathbf{x}),$$

where

$$g_{ij}(\mathbf{x}) = \mathbf{w}_{ij}^T \mathbf{x} + \omega_{ij0},$$

1. Indicate how a piecewise linear machine can be viewed in terms of a linear machine for classifying subclasses of patterns.
2. Show that the decision regions of a piecewise linear machine can be nonconvex and even multiply connected.
3. Sketch a plot of $g_{ij}(x)$ for a one-dimensional example in which $n_1 = 2$ and $n_2 = 1$ to illustrate your answer to part (b)

Where the components are linear

$$g_{ij}(\mathbf{x}) = \mathbf{w}_{ij}^T \mathbf{x} + \omega_{ij0},$$

and the decision rule is to classify x in category i, if maxgk(x) = gi(x), k for j = 1, . . . , c. We can expand this discriminant function as max gk (x) = max max gkj (x), k k j =1,...,nk Where gkj(x) = wktjx+wkj0, are linear functions. Thus our decision rule implies max max gkj (x) = max gij (x) = gij(i)(x). k j =1,...,nk 1j ni Therefore, the piecewise linear machine can be viewed as a linear machine for classifying subclasses of patterns, as follows: Classify x in category i if max max gkj (x) = gij (x). k j =1,...,nk (b) Consider the following two categories in one dimension:

$$\mathbf{w_1 = x : |x| > 1, w_2 = x : |x| < 1}$$

which clearly are not linearly separable. However, a classifier that is a piecewise linear machine with the following discriminant functions

$$\mathbf{g}_{11}(x) = 1xg12(x) = 1 + xg1(x) = \max_{j=1,2} \mathbf{g}_{1j}(x)g2(x) = 2$$

can indeed classify the patterns.

Problem 4

Consider a data set in which each data point t_n is associated with a weighting factor $r_n > 0$, so that the sum-of-squares error function becomes

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} r_n \left\{ t_n - \mathbf{w}^T \phi(\mathbf{x}_n) \right\}^2$$

Find an expression for the solution $\mathbf{w}*$ that minimizes this error function. Give two alternative interpretations of the weighted sum-of-squares error function in terms of (i) data dependent noise variance and (ii) replicated data points.