FIT5201-ASSESSMENT 1

Question 4 [Bayes Rule, 20 Marks]

Suppose we have one red, one blue, and one yellow box. In the red box we have 3 apples and 5 oranges, in the blue box we have 4 apples and 4 orange, and in the yellow box we have 3 apples and 1 orange. Now suppose we randomly selected one of the boxes and picked a fruit. If the picked fruit is an apple, what is the probability that it was picked from the yellow box?
Note that the chances of picking the red, blue, and yellow boxes are 50%, 30%, and 20% respectively and the selection chance for any of the pieces from a box is equal for all the pieces in that box.

Solution

Probability of Red Box is **P(Red)** $=\frac{50}{100}$ = 0.5

Probability of Blue Box is **P(Blue)** = $\frac{30}{100}$ = 0.3

Probability of Yellow Box is **P(Yellow)** $=\frac{20}{100}$ =0.2

Probability of Apple is **P(Apple)** = $\frac{3}{8}\frac{50}{100}+\frac{4}{8}\frac{30}{100}+\frac{3}{4}\frac{20}{100}$ = 0.1875+0.15+0.15 = 0.4875

Probability of Apple picked from Yellow box is

By using Bayes Theorem,

$$\textbf{P(Yellow/Apple)} = \frac{P(apple/Yellow).P(Yellow)}{P(Apple)}$$

$$= \frac{(3/4).(0.2)}{0.4875} = 0.3076$$

Thus the probability of picking apple from a yellow box is 30.76%

Question 5 [Ridge Regression, 25 Marks]

Given the gradient descent algorithms for linear regression (discussed in Chapter 2 of Module 2), derive weight update steps of stochastic gradient descent (SGD) for linear regression with L2 regularisation norm.

Solution:

The error function is added with a regularization term in order to control over-fitting, so that the total error function to be minimized and takes the form

$$E_{\mathcal{D}}(\boldsymbol{w}) + \lambda\Omega(\boldsymbol{w})$$

where λ is the regularisation parameter that controls the relative importance of the data-dependent error E_D and the regularisation function Ω

the error function to minimise is:

$$E(\mathbf{w}) := \frac{1}{2}\sum_{n=1}^{N}[t_n - \mathbf{w}\cdot\boldsymbol{\phi}(\mathbf{x}_n)]^2$$

where $D = \{(x_n, t_n)\}_{n=1}^{N} D := \{(x_n, t_n)\}_{n=1}^{N}$ is the training data.

The gradient of the training objective is:

$$\nabla E(\mathbf{w}) = -\sum_{n=1}^{N}[t_n - \mathbf{w}\cdot\boldsymbol{\phi}(\mathbf{x}_n)]\boldsymbol{\phi}(\mathbf{x}_n)$$

the stochastic gradient descent algorithm updates the parameter vector ω using

$$\boldsymbol{w}^{(\tau)} := \boldsymbol{w}^{(\tau-1)} - \eta^{(\tau)}\nabla E_n(\boldsymbol{w}^{(\tau-1)})$$

For the case of the sum-of-squares error function, the stochastic gradient descent algorithm gives

$$\mathbf{w}^{(\tau)} := \boldsymbol{w}^{(\tau-1)} + \eta^{(\tau)}(t_n - \mathbf{w}^{(\tau-1)}\cdot\boldsymbol{\phi}(\mathbf{x}_n))\boldsymbol{\phi}(\mathbf{x}_n)$$

where τ refers to visiting a *single* train datapoint in Stochastic Gradient Descent

$$\Omega(\boldsymbol{w}) := \tfrac{1}{2}\sum_{j=0}^{M-1}|w_j|^q$$

$$E_{\mathcal{D}} := \tfrac{1}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x}_n))^2$$

When the error function and the regularisation term are added together, the resulting training objective will be

$$\frac{1}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x}_n))^2 + \frac{\lambda}{2}\sum_{j=0}^{M-1}|w_j|^q$$

when q=2

$$\frac{1}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x}_n))^2 + \frac{\lambda}{2}\sum_{j=0}^{M-1}w_j^2$$

Error Term +L2 Normalisation term

$$\frac{1}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x}_n))^2 + \frac{\lambda}{2}\sum_{j=0}^{M-1}w_j^2$$

Differentiation we get

$$\nabla E(\mathbf{w}) = -\sum_{n=1}^{N}[t_n - \mathbf{w}\cdot\boldsymbol{\phi}(\mathbf{x}_n)]\boldsymbol{\phi}(\mathbf{x}_n) + \lambda\, w_{\vec{j}}$$

stochastic gradient descent algorithm updates the parameter vector ω using

$$\boldsymbol{w}^{(\tau)} := \boldsymbol{w}^{(\tau-1)} - \eta^{(\tau)}\nabla E_n(\boldsymbol{w}^{(\tau-1)})$$

$$\mathbf{w}^{(\tau)} := \boldsymbol{w}^{(\tau-1)} + \left(\eta^{(\tau)}(t_n - \mathbf{w}^{(\tau-1)}\cdot\boldsymbol{\phi}(\mathbf{x}_n))\boldsymbol{\phi}(\mathbf{x}_n) - \lambda\,\boldsymbol{w}^{(\tau-1)}\right)$$

Thus, the error term is added with the regularisation term