Question 4.

Section C. Probabilistic Machine Learning

Detail of question:

- red box: 3 apples + 1 oranges 50%
- blue box: 4 apples + 4 oranges 30%
- yellow box: 5 apples + 3 oranges 20%
- given picked fruit is an orange, find probability it is picked from yellow box Pr(yellow|orange)

Lets denote selected Box as random variable B, red box = r, blue box = b, yellow box = y

Then the following probability is given:

$$\mathrm{P}(B=r)=0.5$$
$$\mathrm{P}(B=b)=0.3$$
$$\mathrm{P}(B=y)=0.2$$

Also denote fruits picked as random variable F, orange = o, apple = a

Then the probability need to find out is:

$$\mathrm{P}(B=y \mid F=\text{ orange })$$

According to Bayes rule:

$$P(A \mid B)=\frac{P(B \mid A) P(A)}{P(B)}$$

Then

$$P(B=y \mid F=o)=\frac{P(F=o \mid B=y) P(B=y)}{P(F=o)}$$

We can find the probability of picking a orange from yellow box:

$$P(F=o \mid B=y)=\frac{3}{3+5}=\frac{3}{8}$$

and already have found:

$$P(B=y)=0.2$$

and need to find:

$$P(F=o)=P(F=o \mid B=r) * P(B=r)+P(F=o \mid B=b) * P(B=b)+P(F=o \mid B=y) * P(B=y)$$
$$P(F=o)=\frac{1}{4} * 0.5+\frac{1}{2} * 0.3+\frac{3}{8} * 0.2=0.35$$

Then

$$P(B=y \mid F=o)=\frac{\frac{3}{8} * 0.2}{0.35}=\frac{3}{14}$$

Question 5 [Ridge Regression, 25 Marks]

1. The error function to minimize for linear regression:

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

which is the sum of square of errors

2. Add ridge regression (L2) to it:

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

where lambda is a constant represents the amount of the penalty.

3. Change from GD to SGD :

$$E(\mathbf{w}) := \sum_n E_n(\mathbf{w}) \text{ where } E_n(\mathbf{w}) := \frac{1}{2}\left(t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)^2\right)$$

which means we will update our weights by each individual data point rather than all together at once.

- Then for En(w) with ridge regression (L2):

$$E_n(\mathbf{w}) := \frac{1}{2}\left(t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)^2\right) + \frac{\lambda}{2}w^2$$

4. To update the weight:

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

need to find

$$\nabla E_n\left(\boldsymbol{w}^{(\tau-1)}\right)$$

5. Take derivative of En(W) respect to (W) to find the weight that minimize the error function:

$$\nabla E_n = \frac{d(E_n)}{d(W)} := \frac{1}{2}\left(t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)^2\right)\frac{d}{d(W)} + \frac{\lambda}{2}w^2\frac{d}{d(W)}$$

$$\frac{d(E_n)}{d(W)} := -\left(t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)\right)\boldsymbol{\phi}(\mathbf{x}_n) + \lambda w$$

6. Then weight update step in SGD with L2 regularisation at one data point each iteration is:

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