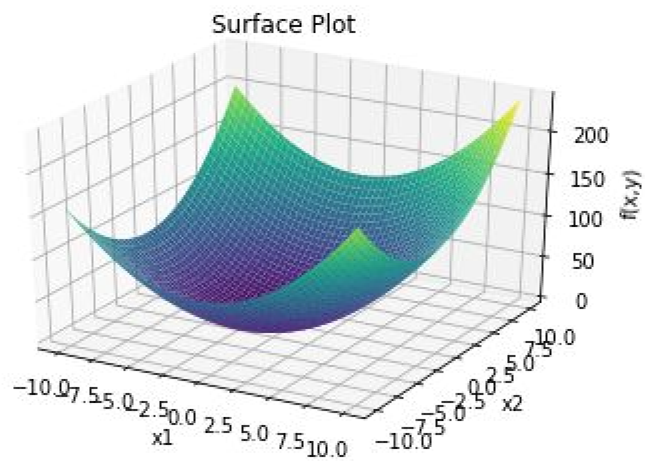
# PRML Mid Semester Solution

**Note:** Marking Scheme is at the end.

1(a)

(i)

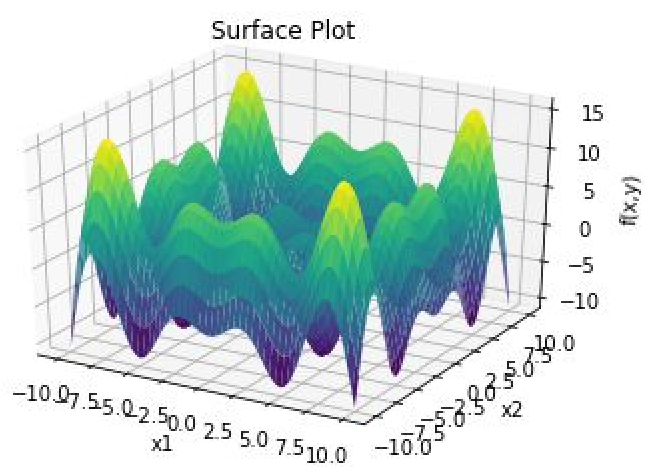
Surface plot of $f_1(x,y) = x_1^2 + x_2^2 + 2x_1 + 2x_2$




$Figure\ :\ 1$

It is a convex surface.

Surface plot of $f_2(x,y) = x_1 sin(x_1) + x_2 sin(x_2)$




$Figure\ :\ 2$

It is a non-convex surface.

(iv)

For $f_1(x, y)$

| Initialization | Learning rate | location of min | Comment |
|---|---|---|---|
| -9,-9 | 0.001 | -1.00, -1.00 | Late convergence compared to l.r =0.1 |
| -5,-5 | 0.001 | -1.00, -1.00 | |
| 7,-8 | 0.001 | -1.00, -1.00 | |
| -9,-9 | 0.1 | -1.00, -1.00 | Converge to global min irrespective of initialization. |
| -5,-5 | 0.1 | -1.00, -1.00 | |
| 7,-8 | 0.1 | -0.99, -1.00 | |
| -9,-9 | 1.1 | -1.21e+80, -1.21e+80 | Does not converge when l.r. is high. |
| -5,-5 | 1.1 | -6.07e+79, -6.07e+79 | |
| 7,-8 | 1.1 | 1.21e+80, -1.06e+80 | |

For $f_2(x, y)$

| Initialization | Learning rate | location of min | Comment |
|---|---|---|---|
| -1,-1 | 0.001 | 0,0 | Late convergence as the l.r. is low. Difference point of convergence when initialization point is different. |
| -5,-5 | 0.001 | -4.91, -4.91 | |
| 7,-8 | 0.001 | 4.92, -10 | |
| -1,-1 | 0.1 | 0,0 | Difference point of convergence when initialization point is different. This is because the gradient decent algorithm gets stuck at local minima. |
| -5,-5 | 0.1 | -4.91, -4.91 | |
| 7,-8 | 0.1 | 4.91, -10 | |
| -1,-1 | 2.5 | 7.11e+107 7.11e+107 | Does not converge when l.r is high. |
| -5,-5 | 2.5 | 1.82e+105 1.82e+105 | |
| 7,-8 | 2.5 | -1.04e+120 -5.92e+107 | |

(ii)

From Figure 1 we can see that their exists a minima here :

$$f_1(x_1, x_2) = x_1^2 + x_2^2 + 2x_1 + 2x_2 .$$

For finding location of minima :-

$$\frac{\partial f_1(x_1, x_2)}{\partial x_1} = 2x_1 + 2 = 0 , \quad \frac{\partial f_1(x_1, x_2)}{\partial x_2} = 2x_2 + 2 = 0$$

$$\Rightarrow x_1 = -1 , x_2 = -1$$

Hence, $(x_1, x_2) = (-1, -1)$ is the location of global ~~local~~ minima.

From Figure 2 it is clear that their exists multiple local maxima & local ~~minima~~ minima. Hence, for finding ~~minim~~ location of minimum value we need to find value at all possible local minima's & extremes ~~values~~

$$f_2(x_1, x_2) = x_1 \sin x_1 + x_2 \sin x_2 .$$

$$\frac{\partial f_2(x_1, x_2)}{\partial x_1} = x_1 \cos x_1 + \sin x_1$$

$x_1 \in [-10, 10]$

$x_2 \in [-10, 10]$.

$$\frac{\partial f_2(x_1, x_2)}{\partial x_2} = x_2 \cos x_2 + \sin x_2 .$$

$$\frac{\partial f_2(x_1, x_2)}{\partial x_1} = 0 = x_1 \cos x_1 + \sin x_1$$

$$\Rightarrow x_1 = -\tan x_1$$

$$\Rightarrow x_1 = 0, \pm 7.725, \pm 4.493$$

$$\frac{\partial f_2(x_1, x_2)}{\partial x_2} = x_2 \cos x_2 + x_2 = 0$$

$$\Rightarrow x_2 = -\tan x_2$$

$$\Rightarrow x_2 = 0, \pm 4.493, \pm 7.725.$$

Therefore, there are 25 points where local maxima or local minima ~~tan~~ exists.

We also need to check at the ~~f~~ 4 extremas $(-10,-10)$, $(-10,10)$, $(+10,-10)$, $(10,10)$.

On checking on all 29 points we get the minimum at extremum i.e $(\pm 10, \pm 10)$.

(iii) Gradient Descent

| Pros | Cons |
| --- | --- |
| (a) Always converges to global minima when surface is convex and learning rate is small. | When the surface is non-convex then reaching ~~void~~ global min depends upon initialization. |
| (b) Rate of convergence depends upon the choice of learning rate. | When learning rate is large then there is divergence. When l.r is very low the it takes long time to converge to the min. |

Normal Equation

| Pros | Cons |
|---|---|
| (a) Works ~~for~~ for small dataset | Fails for larger dataset. |
| (b) Gives parameters values in one step | For parameter computation in one step, inverse of $n \times n$ matrix is computed and this is computationally expensive. |

(b)

| Regression | k-means Clustering |
| --- | --- |
| (i) It predicts continuous values and their output | Groups the data according to the similar data points. |
| (ii) It is supervised learning technique. | It is an unsupervised learning technique. |
| (iii) Eg:- Predicting person's income based on various features | Predicting which ~~point~~ group/cluster new data point belongs to. |

Regression

Cost fn :- $\frac{1}{2M} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$

$h_\theta(x) \rightarrow$ hypothesis fn.

$y \rightarrow$ output.

$M \rightarrow$ # Examples.

As the surface is convex it will reach the minima provided $h_\theta(x)$ is a linear function.

## K-means Clustering

Cost fn: (Error)

$$\frac{1}{M} \sum_{i=1}^{K} \sum_{j \in i^{th} \text{ cluster}} \| c_i - x_j \|_2$$

The second summation in the above equation is like identity function and it is one only when the point j belongs to the $i^{th}$ cluster. This identity brings non linearity in cost fn. Hence, it is a non-convex fn.

b)

Performing clustering task for
Data ÷ 0.5, 0.8, 0.9, 1.0, 1.1, 1.2 & k = 2 clusters
Let's take 0.8 and 0.9 as cluster centers.

Initial centers → ∴ $\mu_1^{(i)} = 0.8$, $\mu_2^{(i)} = 0.9$

Calculating distance

| Data | 0.5 | 0.8 | 0.9 | 1.0 | 1.1 | 1.2 |
|---|---|---|---|---|---|---|
| $\mu_1^{(i)} \to 0.8$ | 0.3 | 0 | 0.1 | 0.2 | 0.3 | 0.4 |
| $\mu_2^{(i)} \to 0.9$ | 0.4 | 0.1 | 0 | 0.1 | 0.2 | 0.3 |

Selecting points to the clusters which are at min distance.

$\Rightarrow$ $\mu_1$ $[0.5, 0.8]$ $\mu_2$ $[0.9, 1.0, 1.1, 1.2]$

New cluster center

$\mu_1^{new} = \frac{0.5+0.8}{2} = 0.65$, $\mu_2^{new} = \frac{0.9+1.1+1.1+1.2}{4}$

$= 1.05.$

Cost (Error) $= \frac{1}{M} \sum_{i=1}^{k} \sum_{j \in i^{th} \text{ cluster}} \| c_i - x_j \|_2^2$

$= (0.5-0.65)^2 + (0.8-0.65)^2 + (0.9-1.05)^2 + (1.0-1.05)^2 + (1.1-1.05)^2 + (1.2-1.05)^2$

$= 0.095$ — (1)

Distance Metric w.r.t $\mu_1^{new}$, $\mu_2^{new}$.

| Data | 0.5 | 0.8 | 0.9 | 1.0 | 1.1 | 1.2 |
|---|---|---|---|---|---|---|
| $\mu_1^{new} = 0.65$ | 0.15 | 0.15 | 0.25 | 0.35 | 0.45 | 0.55 |
| $\mu_2^{new} = 1.05$ | 0.55 | 0.25 | 0.15 | 0.05 | 0.05 | 0.15 |

~~New~~

$\Rightarrow$ $\mu_1$ (0.5, 0.8) $\qquad$ $\mu_2$ ($\cancel{0.8}$, 0.9, 1.0, 1.1, 1.2)

New cluster center $\Rightarrow$

$\mu_1^{new} = \frac{0.5 + 0.8}{2} = 0.65$

$\mu_2^{new} = \frac{0.9 + 1.0 + 1.1 + 1.2}{4} = 1.05$

Stopping criteria →

As the mean do not change we will stop at this step.

∴ Cluster 1 → (0.5, 0.8)

Cluster 2 → (0.9, 1.0, 1.1, 1.2)

Clustering using initial centers as

$\mu_1^{(i)} = 0.5$ , $\mu_2^{(i)} = 1.0$.

| Data | 0.5 | 0.8 | 0.9 | 1.0 | 1.1 | 1.2 |
|---|---|---|---|---|---|---|
| $\mu_1^{(i)} = 0.5$ | 0 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 |
| $\mu_2^{(i)} = 1.0$ | 0.5 | 0.2 | 0.1 | 0 | 0.1 | 0.2 |

$\mu_1$ $\mu_2$

$(0.5)$ $(0.8, 0.9, 1.0, 1.1, 1.2)$

New cluster centers :-

$\mu_1^{(new)} = 0.5$

$\mu_2^{(new)} = \frac{0.8 + 0.9 + 1.0 + 1.1 + 1.2}{5} = 1.0$

Cost (Error) $= (0.5-0.5)^2 + (0.8-1.0)^2 + (0.9-1.0)^2 + (1.0-1.0)^2 + (1.1-1.0)^2 + (1.2-1.0)^2$

$= 0.10$ — (1)

The cluster center did not change hence, we stop at this step.

Cluster 1 → 0.5 , Cluster 2 → (0.8, 0.9, 1, 1.1, 1.2)

We can observe that with different initialization we get different cluster.

Hence, the fn is not a convex function & reaches different local minima when initialized differently.

So, for finding the best cluster we do multiple random initialization & find the cluster with minimum Cost/Error.

Here In this example clusters with initial centers $\mu_1 = 0.5$, $\mu_2 = 1.0$ are chosen.

2) (i) From the t-SNE plot of the balanced dataset we see that two proper clusters are formed

Clustering can be done by better by k-means ~~because~~ when compared to GMM because the circular clusters will ~~easyly~~ easily cluster the two cases. More over, GMM is computationally extensive when compared to k-means.

Aglomerative Clustering is bottom-top approach. Hence from # datapoints to 2 clusters it will take large no. of iterations. Hence, k-means ~~a~~ is better than aglomerative clustering.

~~Since,~~ We do not have outliers in the final set (as outliers are removed by IQR threshold). Hence, DBSCAN will not outperform k-means clustering in this case.

2) ii) For evaluating the clustering algorithm we can use ~~the~~ which of the algorithm (k-means, GMM, hierarchical, DBSCAN) has made a tighter cluster i.e cost function is minimum.

$$\frac{1}{M} \sum_{i=1}^{K} \sum_{\substack{j \in j^{th} \\ \text{cluster}}} \| c_i - x_j \|_2^2 .$$

Note:- Multiple solutions are also allowed for this question.

2)(iii) Multiple solutions are allowed

(iv) Multiple solutions are allowed.

# MARKING SCHEME:

1). a)

(i) 1 Marks

(ii) 1 Marks

(iii) 1 Marks for finding minimum using code.
1 Marks for pros & cons

(iv) 1 Marks.

1) b) (1 Marks) for difference between regression & k-means

(1 Marks) for comment on cost fn.

(2 Marks) for illustrating using code.

(1 Marks) Propose method & justification.

2) (i) (1 Marks) t-SNE (0.5) + preprocessing (0.5)
1 Marks for Justification.

(ii) (1 Marks) for evaluation metric
1 Marks for coding

(iii) Preprocessing - (1 Marks)
To show degradation - (1 Marks)
Final proposal - (1 Marks)

(iv) Block diagram & equation (2 Marks)
Code validation (1 Marks)