

T1

Using direct proof

From $\nabla_A \text{tr} AB$

Using the definition of matrix multiplication, we get: $\nabla \text{tr} \sum_{i,j} \sum_{m=1}^n A_{i,m} \cdot B_{m,j}$

Where n is the dimension of the matrix.

However, we considering only the trace function of the matrix: $\nabla \sum_{k=1}^n \sum_{m=1}^n A_{k,m} \cdot B_{m,k}$

let Y be the resultant matrix obtained from taking derivative with respect to A of $\text{tr} AB$

$$Y_{i,j} = \frac{\partial \sum_{k=1}^n \sum_{m=1}^n A_{k,m} \cdot B_{m,k}}{\partial A_{i,j}} = B_{j,i}$$

Which is the definition of B^T

$$\therefore \nabla \text{tr} AB = B^T$$

T2

Using direct proof

Let $f(A)$ be some function that operate on matrix A

From $\nabla_{A^T} f(A)$

Using the derivative of a matrix definition: $Y_{i,j} = \nabla_{A^T} f(A) = \frac{\partial f(A)}{\partial A_{j,i}}$

Where Y is the resultant matrix obtained from taking derivative with respect to A^T

From $\nabla_A f(A)$

Using the derivative of a matrix definition: $X_{j,i} = \nabla_A f(A) = \frac{\partial f(A)}{\partial A_{j,i}}$

Where X is the resultant matrix obtained from taking derivative with respect to A

Hence, we obtain that $Y_{i,j} = X_{j,i}$

From the definition of matrix transpose, $Y = X^T$

$$\therefore \nabla_{A^T} f(A) = (\nabla_A f(A))^T$$

T3

$$\nabla_A \text{tr} ABA^T C$$

$$(ABA^T C)_{i,j} = \sum_{m=1}^n A_{i,m} (BA^T C)_{m,j}$$

$$(BA^T C)_{m,j} = \sum_{p=1}^n B_{m,p} (A^T C)_{p,j}$$

$$(A^T C)_{p,j} = \sum_{q=1}^n A_{p,q}^T C_{q,j}$$

$$\text{Hence, } (ABA^T C)_{i,j} = \sum_{m=1}^n A_{i,m} \sum_{p=1}^n B_{m,p} \sum_{q=1}^n A_{p,q}^T C_{q,j}$$

$$\text{The trace of } ABA^T C \text{ is } \text{tr}(ABA^T C) = \sum_{k=1}^n \sum_{m=1}^n A_{k,m} \sum_{p=1}^n B_{m,p} \sum_{q=1}^n A_{p,q}^T C_{q,k}$$

Let Y be the resultant matrix obtained by taking derivative with respect to A of $\text{tr} ABA^T C$

$$Y_{s,t} = \frac{\partial \sum_{k=1}^n \sum_{m=1}^n A_{k,m} \sum_{p=1}^n B_{m,p} \sum_{q=1}^n A_{p,q}^T C_{q,k}}{\partial A_{s,t}}$$

$$= \sum_{k=1}^n \sum_{m=1}^n \frac{\partial A_{k,m} \sum_{p=1}^n B_{m,p} \sum_{q=1}^n A_{p,q}^T C_{q,k}}{\partial A_{s,t}}$$

$$\begin{aligned}
&= \sum_{k=1}^n \sum_{m=1}^n (A_{k,m} \frac{\partial \sum_{p=1}^n B_{m,p} \sum_{q=1}^n A_{p,q}^T C_{q,k}}{A_{s,t}} + \sum_{p=1}^n B_{m,p} \sum_{q=1}^n A_{p,q}^T C_{q,k} \frac{\partial A_{k,m}}{\partial A_{s,t}}) \\
&= \sum_{k=1}^n \sum_{m=1}^n A_{k,m} \frac{\partial \sum_{p=1}^n B_{m,p} \sum_{q=1}^n A_{p,q}^T C_{q,k}}{A_{s,t}} + \sum_{k=1}^n \sum_{m=1}^n \sum_{p=1}^n B_{m,p} \sum_{q=1}^n A_{p,q}^T C_{q,k} \frac{\partial A_{k,m}}{\partial A_{s,t}} \\
&= \sum_{k=1}^n \sum_{m=1}^n A_{k,m} \sum_{p=1}^n B_{m,p} \frac{\partial \sum_{q=1}^n A_{p,q}^T C_{q,k}}{A_{s,t}} + \sum_{k=1}^n \sum_{m=1}^n \sum_{p=1}^n B_{m,p} \sum_{q=1}^n A_{p,q}^T C_{q,k} \frac{\partial A_{k,m}}{\partial A_{s,t}} \\
&= \sum_{k=1}^n \sum_{m=1}^n A_{k,m} \sum_{p=1}^n B_{m,p} \sum_{q=1}^n \frac{\partial A_{p,q}^T C_{q,k}}{A_{s,t}} + \sum_{k=1}^n \sum_{m=1}^n \sum_{p=1}^n B_{m,p} \sum_{q=1}^n A_{p,q}^T C_{q,k} \frac{\partial A_{k,m}}{\partial A_{s,t}} \\
&= \sum_{k=1}^n \sum_{m=1}^n A_{k,m} \sum_{p=1}^n B_{m,p} \sum_{q=1}^n \frac{\partial A_{q,p} C_{q,k}}{A_{s,t}} + \sum_{k=1}^n \sum_{m=1}^n \sum_{p=1}^n B_{m,p} \sum_{q=1}^n A_{q,p} C_{q,k} \frac{\partial A_{k,m}}{\partial A_{s,t}} \\
&= \sum_{k=1}^n \sum_{m=1}^n A_{k,m} B_{m,t} C_{s,k} + \sum_{k=1}^n \sum_{m=1}^n \sum_{p=1}^n B_{m,p} \sum_{q=1}^n A_{q,p} C_{q,k} \frac{\partial A_{k,m}}{\partial A_{s,t}} \\
&= \sum_{k=1}^n \sum_{m=1}^n A_{k,m} B_{m,t} C_{s,k} + \sum_{p=1}^n B_{t,p} \sum_{q=1}^n A_{q,p} C_{q,s} \\
&= \sum_{k=1}^n C_{s,k} \sum_{m=1}^n A_{k,m} B_{m,t} + \sum_{p=1}^n B_{t,p} \sum_{q=1}^n A_{q,p} C_{q,s} \\
&= \sum_{k=1}^n C_{s,k} (AB)_{k,t} + \sum_{p=1}^n B_{t,p} (A^T C)_{p,s} \\
&= (CAB)_{s,t} + (BA^T C)_{t,s}
\end{aligned}$$

Using the property of matrix transpose

$$= (CAB)_{s,t} + (C^T AB^T)_{s,t}$$

Since $Y_{s,t} = (CAB)_{s,t} + (C^T AB^T)_{s,t}$

$$\therefore \nabla_A \text{tr} ABA^T C = (CAB) + (C^T AB^T)$$

T4

Here are the point assignment and center update steps of the K mean process using point (3,3), (2,2) and (-3,-3) as the starting point.

Considering point [1 2] , it is closest to the cluster 0 which has center located at [2. 2.]

Considering point [3 3] , it is closest to the cluster 1 which has center located at [3. 3.]

Considering point [2 2] , it is closest to the cluster 0 which has center located at [2. 2.]

Considering point [8 8] , it is closest to the cluster 1 which has center located at [3. 3.]

Considering point [6 6] , it is closest to the cluster 1 which has center located at [3. 3.]

Considering point [7 7] , it is closest to the cluster 1 which has center located at [3. 3.]

Considering point [-3 -3] , it is closest to the cluster 2 which has center located at [-3. -3.]

Considering point [-2 -4] , it is closest to the cluster 2 which has center located at [-3. -3.]

Considering point [-7 -7] , it is closest to the cluster 2 which has center located at [-3. -3.]

New assign center for cluster 0 is [1.5 2.]

New assign center for cluster 1 is [6. 6.]

New assign center for cluster 2 is [-4. -4.6666665]

Recalculated error function is 4.693376

Considering point [1 2] , it is closest to the cluster 0 which has center located at [1.5 2.]

Considering point [3 3] , it is closest to the cluster 0 which has center located at [1.5 2.]

Considering point [2 2] , it is closest to the cluster 0 which has center located at [1.5 2.]

Considering point [8 8] , it is closest to the cluster 1 which has center located at [6. 6.]

Considering point [6 6] , it is closest to the cluster 1 which has center located at [6. 6.]

Considering point [7 7] , it is closest to the cluster 1 which has center located at [6. 6.]

Considering point [-3 -3] , it is closest to the cluster 2 which has center located at [-4. -4.6666665]

Considering point [-2 -4] , it is closest to the cluster 2 which has center located at [-4. -4.6666665]

Considering point [-7 -7] , it is closest to the cluster 2 which has center located at [-4. -4.6666665]

New assign center for cluster 0 is [2. 2.3333333]

New assign center for cluster 1 is [7. 7.]

New assign center for cluster 2 is [-4. -4.6666665]

Recalculated error function is 1.5365908

Considering point [1 2] , it is closest to the cluster 0 which has center located at [2. 2.3333333]

Considering point [3 3] , it is closest to the cluster 0 which has center located at [2. 2.3333333]

Considering point [2 2] , it is closest to the cluster 0 which has center located at [2. 2.3333333]

Considering point [8 8] , it is closest to the cluster 1 which has center located at [7. 7.]

Considering point [6 6] , it is closest to the cluster 1 which has center located at [7. 7.]

Considering point [7 7] , it is closest to the cluster 1 which has center located at [7. 7.]

Considering point [-3 -3] , it is closest to the cluster 2 which has center located at [-4. -4.6666665]

Considering point [-2 -4] , it is closest to the cluster 2 which has center located at [-4. -4.6666665]

Considering point [-7 -7] , it is closest to the cluster 2 which has center located at [-4. -4.6666665]

New assign center for cluster 0 is [2. 2.3333333]

New assign center for cluster 1 is [7. 7.]

New assign center for cluster 2 is [-4. -4.6666665]

Recalculated error function is 0.0

Since the center for every cluster doesn't change anymore, we finish the K-mean process.

T5

Here are the point assignment and center update steps of the K mean process using point (-3,-3), (2,2), (-7, -7) as the starting point.

Considering point [1 2] , it is closest to the cluster 1 which has center located at [2. 2.]

Considering point [3 3] , it is closest to the cluster 1 which has center located at [2. 2.]

Considering point [2 2] , it is closest to the cluster 1 which has center located at [2. 2.]

Considering point [8 8] , it is closest to the cluster 1 which has center located at [2. 2.]

Considering point [6 6] , it is closest to the cluster 1 which has center located at [2. 2.]

Considering point [7 7] , it is closest to the cluster 1 which has center located at [2. 2.]

Considering point [-3 -3] , it is closest to the cluster 0 which has center located at [-3. -3.]

Considering point [-2 -4] , it is closest to the cluster 0 which has center located at [-3. -3.]

Considering point [-7 -7] , it is closest to the cluster 2 which has center located at [-7. -7.]

New assign center for cluster 0 is [-2.5 -3.5]

New assign center for cluster 1 is [4.5 4.6666665]

New assign center for cluster 2 is [-7. -7.]

Recalculated error function is 3.7230513

Considering point [1 2] , it is closest to the cluster 1 which has center located at [4.5 4.6666665]

Considering point [3 3] , it is closest to the cluster 1 which has center located at [4.5 4.6666665]

Considering point [2 2] , it is closest to the cluster 1 which has center located at [4.5 4.6666665]

Considering point [8 8] , it is closest to the cluster 1 which has center located at [4.5 4.6666665]

Considering point [6 6] , it is closest to the cluster 1 which has center located at [4.5 4.6666665]

Considering point [7 7] , it is closest to the cluster 1 which has center located at [4.5 4.6666665]

Considering point [-3 -3] , it is closest to the cluster 0 which has center located at [-2.5 -3.5]

Considering point [-2 -4] , it is closest to the cluster 0 which has center located at [-2.5 -3.5]

Considering point [-7 -7] , it is closest to the cluster 2 which has center located at [-7. -7.]

New assign center for cluster 0 is [-2.5 -3.5]

New assign center for cluster 1 is [4.5 4.6666665]

New assign center for cluster 2 is [-7. -7.]

Recalculated error function is 0.0

Since the center for every cluster doesn't change anymore, we finish the K-mean process.

The cluster assignments are different from the prior run due to the starting point being different.

T6

The one from T4 is better since it displays more clustering behavior.

One of the possible ways to evaluate the performance of the K mean learning model numerically is to use the sum of the square of the distance of each point to its assigned centroid and divide it by the number of total points.

From the idea above, we calculate the performance of T4 and T5 model as follows:

MSE value for the T4 is 3.26

MSE value for the T5 is 8.65

Since the lower value means better performance, we can conclude that the starting points in T4 produce better model.

OT1

Using elbow method:

We find fraction of explain variance for each K with our data.

K=1, fraction of explain variance=0.0

K=2, fraction of explain variance=0.762

K=3, fraction of explain variance=0.930

K=4, fraction of explain variance=0.981

K=5, fraction of explain variance=0.9888

From the values above, we can see that between k=4 and k=5, there is no significant change in fraction of explain variance.

Therefore, k=4 would be the suitable k value for this set of points.

T7

Median age of the training set is 28.

T8

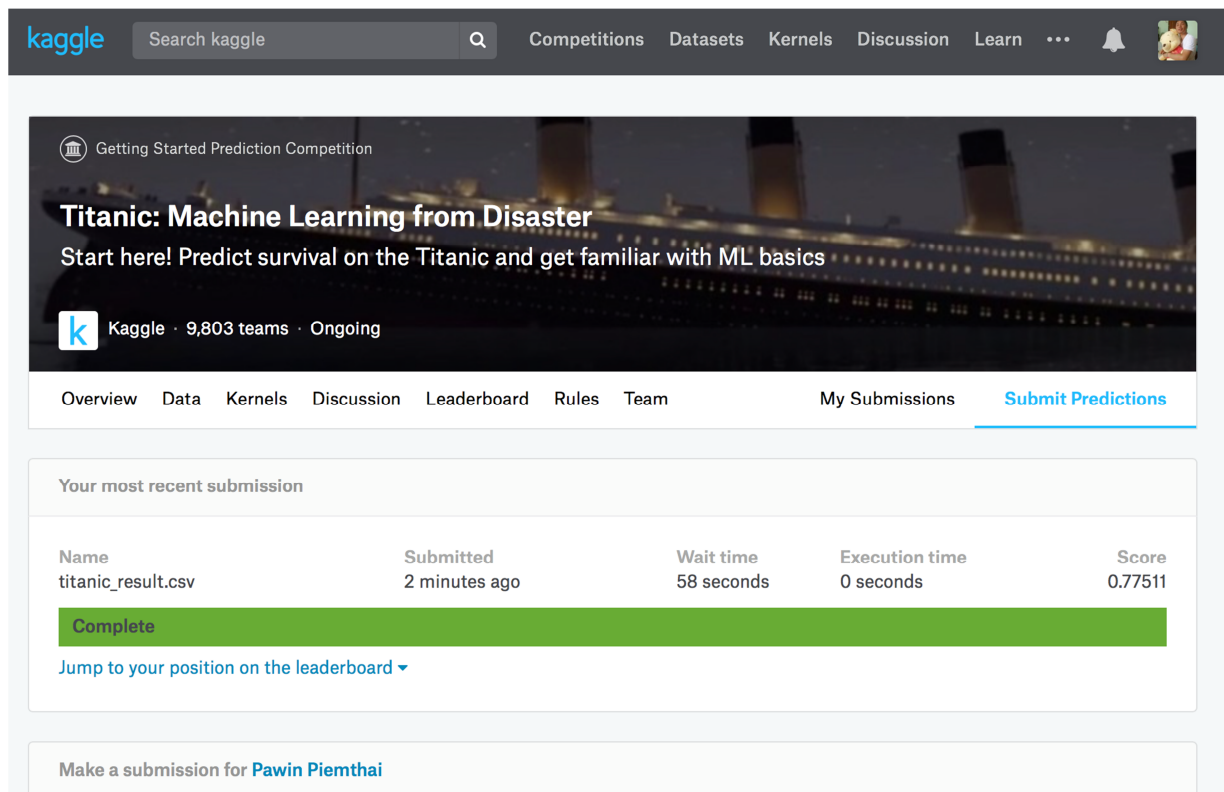
Mode of Embarked is 0 (S)

T9

The code is in the submitted file (section T9).

T10

Here is the submitted result:



The screenshot shows the Kaggle website interface for the 'Titanic: Machine Learning from Disaster' competition. The header includes the Kaggle logo, a search bar, and navigation links for Competitions, Datasets, Kernels, Discussion, Learn, and a user profile. The main banner features the competition title, a subtitle 'Start here! Predict survival on the Titanic and get familiar with ML basics', and the Kaggle logo with '9,803 teams · Ongoing'. Below the banner is a navigation bar with links: Overview, Data, Kernels, Discussion, Leaderboard, Rules, Team, My Submissions, and Submit Predictions. The 'My Submissions' section displays a table of recent submissions. The first submission, 'titanic_result.csv', is highlighted with a green bar and the status 'Complete'. Below the table is a link to 'Jump to your position on the leaderboard'. At the bottom, there is a prompt to 'Make a submission for Pawin Piemthai'.

Name	Submitted	Wait time	Execution time	Score
titanic_result.csv	2 minutes ago	58 seconds	0 seconds	0.77511

Complete

[Jump to your position on the leaderboard](#)

Make a submission for [Pawin Piemthai](#)

OT2

The code is in the submitted file (section OT2).

OT3

The weight matrix (θ) obtained from two methods are converging on the same value.

The weight matrix obtained from linear regression using gradient descend is:

```
[[ 0.77652116] [-0.18843415] [ 0.49087112] [-0.00505407] [ 0.04911372]]
```

The weight matrix obtained from linear regression using matrix inversion is:

```
[[ 0.77654442] [-0.18843944] [ 0.49086711] [-0.00505436] [ 0.04911346]]
```

The difference between two matrices is:

```
[[-2.32609857e-05] [ 5.29192529e-06] [ 4.01169591e-06] [ 2.92394898e-07] [ 2.62349718e-07]]
```

Mean square error of the difference is 1.17065191e-10

The convergence will be more pronounced if the amount of iterations used to train the model with gradient descend is increased.

The code is in the submitted file (section OT3).

OT4

For comparison,

The accuracy of the base model is 0.8058

The accuracy of the training set of OT4A (adding age*sex and age²) is 0.7957

The accuracy of the training set of OT4B (adding age*sex and sex²) is 0.7868

The accuracy of the training set of OT4C (adding age*sex and embarked²) is 0.3962

From the result above, it has been seen that adding high level variables to the model does not produce any significant gain in performance (as oppose to the computational cost it incurs).

Adding highly correlated features such as age will slightly decrease the performance of the model in some case.

However, when adding the low correlated features to the model (such as embarked²), those defects will be amplified making the performance of the model significantly decreased.

OT5

When the features are reduced to just age and sex, we found the accuracy of the model to be 0.7868.

Which is still pretty close to the accuracy of the base model (about 2% drop)

This is due to the high correlation between the features remaining and the predicted targets. In other words, those two features are enough to capture the characteristic pattern that can predict the survivability.