**Q1** []. Given $n$ data points with three features. The variance of the first feature with respect to all data points is near zero. The variance of the second feature with respect to all data points is a small positive number. And the variance of the third feature with respect to all data points is a large positive number. If we want to select two features out of three, which features should be selected? Give an explanation.

**Q2** []. If the original set of features is $\left\{f\_{i}\right\}, i=1,…,N$ and the set of features after some preprocessing is $\left\{f\_{j}\right\}, j\in i$. Then, what does the preprocessing do to our set of features?

**Q3** []. Why Min-hashing can be considered as a feature extraction technique?

**Q4** []. Given the following feature-by-data matrix $X$:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | data1 | data2 | data3 | data4 |
| f1  | 2 | 8 | 1 | 7 |
| f2 | 3 | 9 | 3 | 8 |
| f3 | 4 | 8 | 4 | 9 |
| f4 | 1 | 7 | 2 | 8 |
| f5 | 2 | 9 | 3 | 7 |
| f6 | 4 | 8 | 1 | 9 |

And the following linear projection matrix $U$

|  |  |  |
| --- | --- | --- |
| 1 | 8 | 3 |
| 2 | 9 | 4 |
| 1 | 8 | 3 |
| 2 | 9 | 4 |
| 1 | 8 | 3 |
| 2 | 9 | 4 |

1. [] Compute the projection of the data using the projection matrix $U$
2. [] Does the projection preserve the data similarity? Explain your answer
3. [] Show how to recover row vector f1 in matrix $X$using the projected data $Y$

**Q5** []. Given the following data

|  |  |  |  |
| --- | --- | --- | --- |
| Day | Month | Year | Activities |
| 4 | January | 2001 | Stay home |
| 8 | July | 2002 | Go outside |
| 12 | December | 2003 | Stay home |
| 5 | February | 2004 | Stay home |
| 22 | August | 2005 | Go outside |
| 15 | June | 2006 | Go outside |
| 7 | July | 2007 | Go outside |

Apply feature engineering to simplify the presentation of the data

**Q6** []. Given the following data:

|  |  |  |  |
| --- | --- | --- | --- |
| data1 | data2 | data3 | data4 |
| 2 | 8 | 1 | 7 |
| 3 | 9 | 3 | 8 |
| 4 | 8 | 4 | 9 |
| 1 | 7 | 2 | 8 |
| 2 | 9 | 3 | 7 |
| 4 | 8 | 1 | 9 |

Apply standard scaler to normalize every data

**Q7** []. Given the following data:

|  |  |  |  |
| --- | --- | --- | --- |
| data1 | data2 | data3 | data4 |
| 2 | 8 | 1 | 7 |
| 3 | 9 | 3 | 8 |
| 4 | 8 | 4 | 9 |
| 1 | 7 | 2 | 8 |
| 2 | 9 | 3 | 7 |
| 4 | 8 | 1 | 9 |

1. [] Compute the output of the first PCA step (sample mean)
2. [] Compute the output of the second PCA step (normalization)
3. [] Compute the covariance matrix
4. [] Compute eigenvectors and eigenvalues
5. [] What is the principal eigenvector (the eigenvector corresponds to the largest eigenvalue)?
6. [] What will be the best approximation of the covariance matrix by using the first $K$ eigenvectors and eigenvalues where $K=2$
7. [] What will be the approximation error of the matrix above (your answer in f)?
8. [] If we want to preserve at least 90% of the information in the data, how many principal eigenvectors / eigenvalues need to be used?

**Q8** []. If the trend/pattern found in whole data is opposite than the trend/pattern found in the several groups in the data, which trend/pattern that has a higher chance to be the correct one?

**Q9** []. Given the following data about the success rate of treatment using drug A and drug B.

1. [] What is the pattern in case 1, case 2, and case 3?
2. [] What is the pattern in the overall case?
3. [] Why there can be discrepancy between the patterns?
4. [] In the overall case, drug B is better than drug A even though drug B could not outperform drug A in all other cases. By looking at the data, in which case (1, 2, or 3) that drug B has much more samples than drug A?