1. PCA:

1. w is [ -0.6 -0.8 ].

2.

Plot:

A graph with red and green arrows

Description automatically generated

3. Since we want to maximize , we need to find the maximum eigenvalue in the covariance matrix. Since the largest eigenvalue 20 appears at the third line and w needs to satisfy , and the optimal value of the program is 20.

2. Basics in information theory:

1. .

Since ,

.

2. . As we expand , and substitute and ,

.

According to the – divergence,

.

Thus we can expand this formula according to .

.

And this is same as . Thus, .

3. k-means with soft assignment:

1. Since , and the goal is to minimize the program within the set of in the soft assignment, which means to find values that minimize the program in the range of numbers between 0,1 that belongs to real number, and the hard assignment is to minimize the program in the range of only two values which is 0 or 1 which is , there must exist values that makes the program much smaller in than or at least they are equal, which corresponds to less than or equal to. Thus, we can say that:

.

2. First, for the hard assignments, since each point is assigned to its nearest cluster center , it’s goal is to minimize the distance under the constraint of . Since , any achievements accomplished by hard assignment can be completed by the soft one. As we are progressing, if there were a solution in the soft assignment that strictly minimized the objective function more than the optimal hard assignment, it would imply that there exists a set of cluster centers and a distribution of assignment in the solution that results in a lower objective value than assigning each data point to its nearest cluster center. However, the contradiction appears now that the initial premise of the hard assignment optimally assigns each data point to its nearest cluster center. Thus, further minimization is prohibited and impossible mathematically since the hard assignment configuration is already considered within the soft assignment solutions. Therefore, according to the contradiction proving above:

.

3. When we use optimization steps iteratively like gradient descent, every time when we update , it will render the increment of the strength for closer clusters and decreasing it for farther clusters. As the process going on, one will become nearly one since it’s near one of the cluster centers and the other will become 0 accordingly, which is nearly the output of a hard assignment. Thus, we can say that the soft assignment problem can converge to a hard assignment problem globally.

4. Bernoulli mixture model:

1. Since , and

,

.

Then we take the log for both sides:

.

2. Since , for each element in the formula:

is the likelihood of given it was generated by the k-th Bernoulli distribution, which is .

is prior probability of which is .

And is the total probability observing across all K distributions which is .

Thus, we can synthesis them into the final formula as:

.

3. Since the expected log-likelihood function involving is:

.

Then we can take the derivative respective to and set to zero to find the maximum:

.

.

.

.

.

Since the expected log-likelihood function involving is:

.

Now we can use a Lagrange multiplier to make the constraint: :

.

.

.

.

.

.

5. VAE

3.

A graph with numbers and a line

Description automatically generated with medium confidence

A graph with numbers and symbols

Description automatically generated with medium confidence

A graph showing a number of points

Description automatically generated