Zhangsheng Lai
1002554

Machine Learning: Homework 2

Task 1: K-means

1. To show that the algorithm terminates in finite number of steps, it suffices to show that the number of possible indicator matrix γ is finite and the algorithm used makes L a non-increasing function of the iteration. There are finite number of indicator matrix γ as each data point can belong to exactly one cluster, thus for n data points and k clusters, there are k^n possible configurations of the clusters. To show that L is non-increasing with each iteration, it can be seen that γ is determined by choosing the cluster j such that for each data point x_i for $i = 1, \ldots, n$ we have $\|x_i - \mu_j\| \leq \|x_i - \mu_{j'}\|, \ \forall j'$. Now we are left to show that the recomputation of the μ_j's using the formula

$$\frac{\sum_{i=1}^{n} \gamma_{ij} x_i}{\sum_{i=1}^{n} \gamma_{ij}} \tag{1}$$

which is equal to the mean of the data points belonging to the cluster, makes $\sum_{i=1}^{n} \gamma_{ij} \|x_i - \mu_j\|^2$ the minimum. This is true since for a given j,

$$\frac{\partial}{\partial \mu_j} \sum_{i=1}^{n} \gamma_{ij} \|x_i - \mu_j\|^2 = -2 \sum_{i=1}^{n} \gamma_{ij} (x_i - \mu_j)$$

and solving for the value of μ_j when we equate the equation above to 0 gives (1), the average of the data points in cluster j.

2.

$$\begin{aligned}
\sum_{j=1}^{k} \left(\sum_{i=1}^{n} \gamma_{ij} \right) W_j(x) + nB(x) &= \sum_{j=1}^{k} \sum_{i=1}^{n} \gamma_{ij} \left[\|x_i - \mu_j\|^2 + \|\mu_j - \widehat{x}\|^2 \right] \\
&= \sum_{j=1}^{k} \sum_{i=1}^{n} \gamma_{ij} \left[\|x_i\|^2 + \|\mu_j\|^2 + \|\mu_j\|^2 + \|\hat{x}\|^2 - 2x_i\mu_j - 2\mu_j\hat{x} \right] \\
&= \sum_{j=1}^{k} \sum_{i=1}^{n} \gamma_{ij} \|x_i - \hat{x}\|^2 + C \\
&= \sum_{i=1}^{n} \|x_i - \hat{x}\|^2 + C = nT(x) + C
\end{aligned}$$

where $C = 2x_i\hat{x} - 2\|\mu_j\|^2 + 2x_i\mu_j + 2\mu_j\hat{x}$.

From the relationship we estabilised above, we see that $nT(x) + C$ is a constant for fixed $\{\mu_1, \ldots, \mu_k\}$. Thus, when we try to minimize intra-cluster deviation, $W_j(x)$, we obtain new μ_j's with $W_j(x) \geq W_j(x)'$ where $W_j(x)'$ is the intra-cluster deviation with new μ_j's.

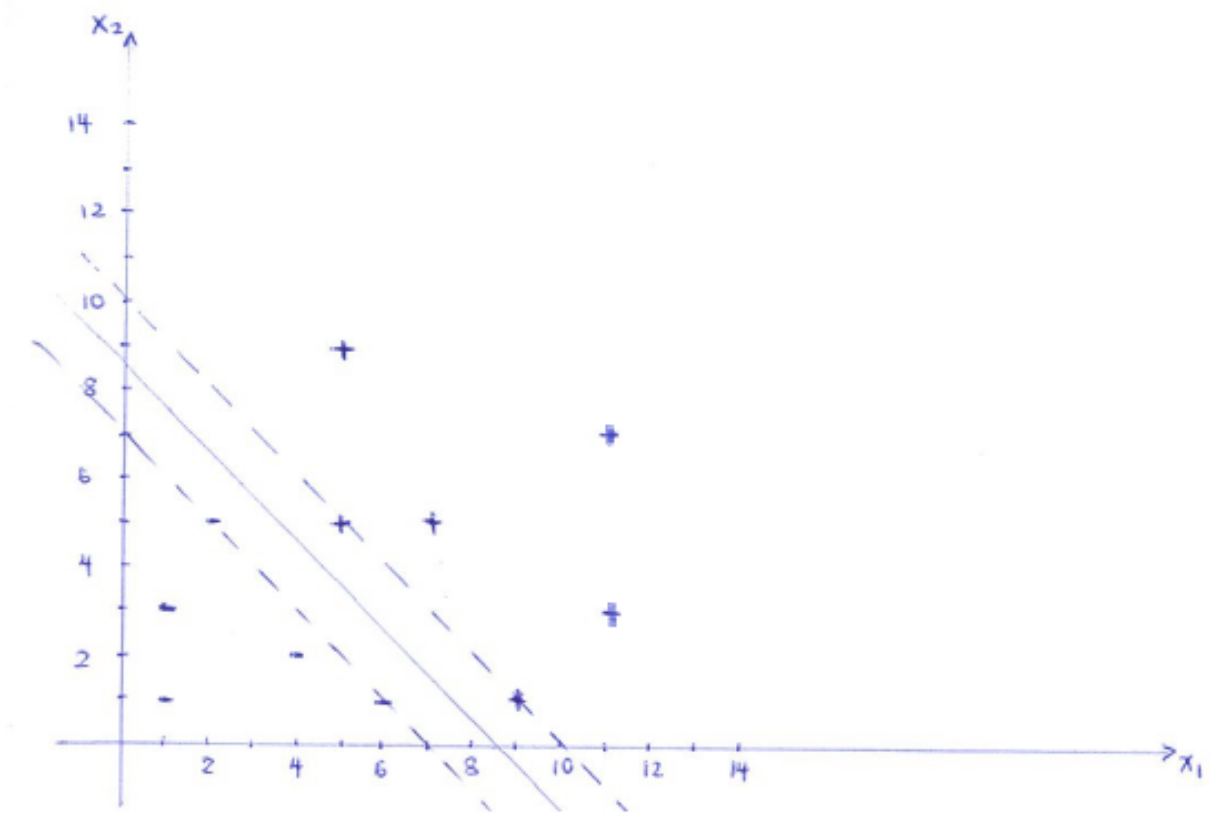
3. Let k be given and assume that we have the indicator matrix γ and the means of the clusters $\{\mu_1, \ldots, \mu_k\}$ such that $L(\gamma, \mu_1, \ldots, \mu_k)$ is a minimum. Now suppose we increase the number of clusters to $k+1$, and we initialize a new cluster mean μ_{k+1}. Here we can assume that this newly added cluster mean will have a new point in its cluster when an iteration of the algorithm is being performed, otherwise, this newly added cluster will be removed the moment it is added. This means there exists a data point x_a where $1 \leq a \leq n$ such that

$$\|x_a - \mu_j\| \geq \|x_a - \mu_{k+1}\|$$

for all $j = 1, \ldots, k$. Thus this implies that we get

$$\min_{\gamma', \mu_1, \ldots, \mu_{k+1}} L(\gamma', \mu_1, \ldots, \mu_{k+1}) \leq L(\gamma', \mu_1, \ldots, \mu_{k+1}) \leq \min_{\gamma, \mu_1, \ldots, \mu_k} L(\gamma, \mu_1, \ldots, \mu_k)$$

where γ' is the new clustering with at least one point in the newly added cluster. Thus ifwe try to choose the number of clusters by minimizing L, L will converge to zero with n clusters for the n data points, which is the trivial clustering of each data point into a single cluster.



Task 2: Support Vector Machine

1. The support vectors are $(2,5), (6,1), (5,5), (9,1)$.

2. The separating hyperplane has the formula $x_1 + x_2 - 8.5 = 0$. Thus $w_1 = 1$, $w_2 = 1$ and $b = 8.5$.

3. The scaled values are $w_1 = 2/3$, $w_2 = 2/3$ and $b = -17/3$. We see that the division of the formula is just doing a scalar multiplication of the whole equation by $2/3$. As linear equations are unique up to scalar multiplication, we can just choose the function to be represented by the one such that the values of the support vectors are -1 and $+1$.

4. Using $f(x_1, x_2) = \frac{2}{3}x_1 + \frac{2}{3}x_2 - \frac{17}{3}$,
   Functional margin for $((4,2), -1)$: $-1(17/3 - (2/3 \cdot 4) - (2/3 \cdot 2)) = 5/3$
   Geometric margin for $((4,2), -1)$: $-5/3 \div \sqrt{(-2/3)^2 + (-2/3)^2} = 5\sqrt{2}/4$
   The sign of the functional margin tells us if the classification of the point agree with the label, i.e. when the functional margin is positive, it means that the label and the classification of the point agree with each other. If the functional margin is negative, it means that the label and the classification of the point disagree with each other. The magnitude of the functional margin is the scaled value of the geometric margin.