COMP 6321 Machine Learning Assignment

Qingbo Kang, Concordia University 40058122

1 VC dimension

1.1 a

$VC(H) = 1$.
Analysis: Consider 1 point p_1, if $p_1 > a$, then we can label p_1 as $+$, otherwise $p_1 < a$, then labelled as $-$. But if we have 2 points p_1, p_2 and $p_1 < p_2$, if we label p_1 as $+$ and p_2 as $-$. Then the class of hypotheses of the form $[a, +\infty)$ cannot shatter these 2 points. Therefore, the VC dimension for this class is 1.

1.2 b

$VC(H) = 2$.
Analysis: Consider 2 points p_1, p_2 and $p_1 < p_2$, we have 4 cases:

(1) $label(p_1) = -$ and $label(p_2) = -$, then we have hypothesis $(-\infty, a]$ and $p_1 > a$.

(2) $label(p_1) = -$ and $label(p_2) = +$, then we have hypothesis $[a, \infty]$ and $p_1 < a < p_2$.

(3) $label(p_1) = +$ and $label(p_2) = -$, then we have hypothesis $(-\infty, a]$ and $p_1 < a < p_2$.

(4) $label(p_1) = +$ and $label(p_2) = +$, then we have hypothesis $[a, \infty)$ and $a < p_1$.

But if we have 3 points p_1, p_2, p_3 and $p_1 < p_2 < p_3$. If we label p_1 as $+$, p_2 as $-$, p_3 as $+$, then the class of hypothesis of one-sided intervals cannot shatter these 3 points. Therefore, the VC dimension for this class is 2.

1.3 c

$VC(H) = 2$.
Analysis: For dichotomy which consists of 2 points, this class of hypothesis can shatter them, the analysis is above. However, if there are 3 points for example p_1, p_2, p_3 and $p_1 < p_2 < p_3$, if we label p_1 as $+$, p_2 as $-$, p_3 as $+$, the class of hypothesis cannot shatter this. Therefore, the VC dimension of this class is 2.

1.4 d

$VC(H) = 4$.
Analysis: Consider there are 4 points p_1, p_2, p_3, p_4 and $p_1 < p_2 < p_3 < p_4$, and $label(p_1) = +, label(p_2) = -, label(p_3) = +, label(p_4) = -$. The class of hypothesis can shatter this situation when $a < p_1 < b < p_2 < c < p_3 < d$. But if we add one extra point $p_5, p_4 < p_5$ and $label(p_5) = +$, for this situation, the class of hypothesis of the form $[a, b] \cup [c, d]$ cannot shatter it. Therefore, the VC dimension of this class is 4.

1.5 e

$VC(H) = 2k$.
Analysis: Consider there are $2k$ points $p_1, p_2, p_3, ..., p_{2k}$ and $label(p_1) = +, label(p_2) = -, label(p_3) = +, ..., label(p_{2k}) = -$, the positive and negative label are appear alternately. Then the unions of k intervals can shatter them, as a simple example are illustrated in the above question. But if we add 1 extra point p_{2k+1} and label it as $+$, then we need unions of $k + 1$ intervals to shatter them, in other words, unions of k intervals cannot shatter $2k + 1$ points. Therefore, the VC dimension of this class is $2k$.

2 KL-divergence

2.1 a

Prove:

$$\begin{aligned} -KL(P||Q) &= -\sum_x P(x) \log \frac{P(x)}{Q(x)} \\ &= \sum_x P(x) \log \frac{Q(x)}{P(x)} \end{aligned}$$

And by Jensen's inequality, let $f(x) = \frac{Q(x)}{P(x)}$, then:

$$\begin{aligned} -KL(P||Q) &= \sum_x P(x) \log \frac{Q(x)}{P(x)} \\ &\leq \log(\sum_x P(x) \frac{Q(x)}{P(x)}) \\ &\leq \log(\sum_x Q(x)) \\ &\leq \log(1) \\ &\leq 0 \end{aligned}$$

Here because $Q(x)$ is a probability distribution, so $\sum_x Q(x) = 1$.
Therefore, we can obtain: $KL(P||Q) \geq 0, \forall P, Q$.

2.2 b

In order to have

$$KL(P||Q) = 0$$

we need

$$\log \frac{P(x)}{Q(x)} = 0$$

that is

$$\begin{aligned} \frac{P(x)}{Q(x)} &= 1 \\ P(x) &= Q(x) \end{aligned}$$

Therefore, when $P(x) = Q(x)$ which means these two probability distributions are equal, then we have $KL(P||Q) = 0$.

2.3 c

When $Q(x) = 0, P(x) \neq 0$, then we have

$$\lim_{Q(x) \to 0} KL(P||Q) = \infty$$

Therefore, the maximum value possible of $KL(P||Q)$ is ∞, when $Q(x) = 0$ and $P(x) \neq 0$.

2.4 d

No, $KL(P||Q) \neq KL(Q||P)$.
Explanation: Here is a counterexample: suppose $x_i = \{0, 1\}$ and $P(0) = \frac{1}{3}, P(1) = \frac{1}{3}, Q(0) = \frac{1}{2}, Q(1) = \frac{1}{2}$.

$$KL(P||Q) = \frac{1}{3}\log\frac{2}{3} + \frac{2}{3}\log\frac{4}{3} = 0.0566$$
$$KL(Q||P) = \frac{1}{2}\log\frac{3}{2} + \frac{1}{2}\log\frac{3}{4} = 0.0589$$

Therefore we can see that $KL(P||Q) \neq KL(Q||P)$.

2.5 e

$$\begin{aligned}
KL(P(X,Y)||Q(X,Y)) &= \sum_x \sum_y P(x,y)\log\frac{P(x,y)}{Q(x,y)} \\
&= \sum_x \sum_y P(x,y)\log\frac{P(y|x)P(x)}{Q(y|x)Q(x)} \\
&= \sum_x \sum_y P(x,y)(\log\frac{P(x)}{Q(x)} + \log\frac{P(y|x)}{Q(y|x)}) \\
&= \sum_x \sum_y P(x,y)\log\frac{P(x)}{Q(x)} + P(x,y)\log\frac{P(y|x)}{Q(y|x)} \\
&= \sum_x \sum_y P(x|y)P(y)\log\frac{P(x)}{Q(x)} + P(y|x)P(x)\log\frac{P(y|x)}{Q(y|x)} \\
&= \sum_x \sum_y [P(x|y)P(y)\log\frac{P(x)}{Q(x)}] + \sum_y [P(y|x)P(x)\log\frac{P(y|x)}{Q(y|x)}] \\
&= \sum_x \log\frac{P(x)}{Q(x)} \sum_y [P(x|y)P(y)] + P(x)\sum_y [P(y|x)\log\frac{P(y|x)}{Q(y|x)}] \\
&= \sum_x P(x)\log\frac{P(x)}{Q(x)} + P(x)\sum_y P(y|x)\log\frac{P(y|x)}{Q(y|x)} \\
&= KL(P(X)||Q(X)) + KL(P(Y|X)||Q(Y|X))
\end{aligned}$$

2.6 f

Let $\hat{P}$ be the empirical distribution, then

$$\begin{aligned}
KL(\hat{P}||P_\theta) &= \sum_x \hat{P}\log\frac{\hat{P}}{P_\theta} \\
&= \sum_x \hat{P}\log\hat{P} - \sum_x \hat{P}\log P_\theta \\
&= -H(\hat{X}) - \sum_x \hat{P}\log P_\theta
\end{aligned}$$

Since $H(\hat{X})$ and $\hat{P}$ are irrelevant with θ and they are both greater or equal to 0. Therefore

$$\arg\min_\theta KL(\hat{P}||P_\theta) = \arg\max_\theta \sum_{i=1}^{m} \log P_\theta(x_i)$$

3 K-means

The code implemented in Q3.m and Kmeans.m.

3.1 How many clusters there are in the end

In the end, there are 6 cluster. Figure 1 shows the running result.

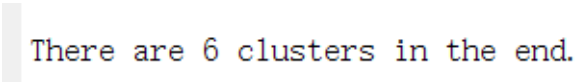

Figure 1: Running result of how many

3.2 The final centroids of each cluster

Figure 2 shows the running result of the final centroids of each cluster.

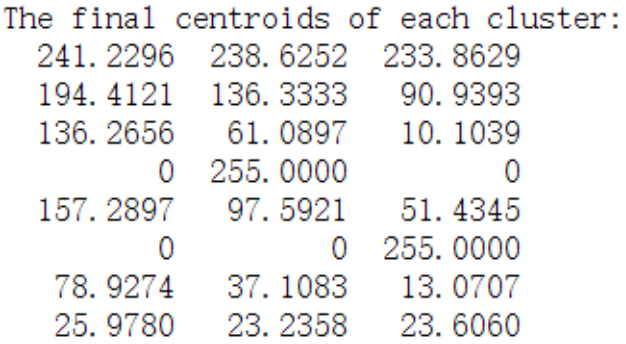

Figure 2: Running result of final centroids

3.3 The number of pixels associated to each cluster

Figure 3 shows the running result of the number of pixels associated to each cluster.

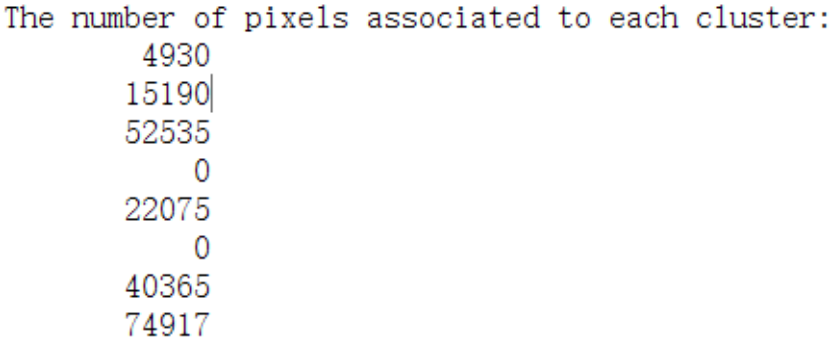

Figure 3: Running result of number of pixels

3.4 The sum of squared distance from each pixel to the nearest centroid after every iteration of the algorithm

Below shows the printed result of the sum of squared distance from each pixel to the nearest centroid after every iteration of the algorithm.
Iteration: 1 Sum of squared distance: 266210702.563910 Iteration: 2 Sum of squared distance: 138915406.016151
Iteration: 3 Sum of squared distance: 132690991.261671 Iteration: 4 Sum of squared distance: 119105157.581780
Iteration: 5 Sum of squared distance: 93196172.771484 Iteration: 6 Sum of squared distance: 92828148.615213
Iteration: 7 Sum of squared distance: 93207347.981797 Iteration: 8 Sum of squared distance: 93054811.480479
Iteration: 9 Sum of squared distance: 92732360.180222 Iteration: 10 Sum of squared distance: 92512834.769297
Iteration: 11 Sum of squared distance: 92394458.088659 Iteration: 12 Sum of squared distance: 92100159.340050
Iteration: 13 Sum of squared distance: 91946523.120799 Iteration: 14 Sum of squared distance: 91763713.267196
Iteration: 15 Sum of squared distance: 91697382.883979 Iteration: 16 Sum of squared distance: 91548780.441668

Iteration: 17 Sum of squared distance: 91414807.916007 Iteration: 18 Sum of squared distance: 91263223.079678
Iteration: 19 Sum of squared distance: 91110186.148409 Iteration: 20 Sum of squared distance: 90902302.054784
Iteration: 21 Sum of squared distance: 90529329.051193 Iteration: 22 Sum of squared distance: 89987875.893149
Iteration: 23 Sum of squared distance: 89262298.917681 Iteration: 24 Sum of squared distance: 87909145.666518
Iteration: 25 Sum of squared distance: 85504512.695055 Iteration: 26 Sum of squared distance: 83033746.956877
Iteration: 27 Sum of squared distance: 81154394.350952 Iteration: 28 Sum of squared distance: 79532091.176771
Iteration: 29 Sum of squared distance: 78348386.675937 Iteration: 30 Sum of squared distance: 77630585.049322
Iteration: 31 Sum of squared distance: 77244002.119292 Iteration: 32 Sum of squared distance: 77122065.439906
Iteration: 33 Sum of squared distance: 77025274.249422 Iteration: 34 Sum of squared distance: 77139516.818181
Iteration: 35 Sum of squared distance: 77165553.695484 Iteration: 36 Sum of squared distance: 77233116.927434
Iteration: 37 Sum of squared distance: 77250952.046324 Iteration: 38 Sum of squared distance: 77221331.651950
Iteration: 39 Sum of squared distance: 77201897.360321 Iteration: 40 Sum of squared distance: 77191054.277191
Iteration: 41 Sum of squared distance: 77183331.045046 Iteration: 42 Sum of squared distance: 77176410.091074
Iteration: 43 Sum of squared distance: 77172984.279567 Iteration: 44 Sum of squared distance: 77169922.743257
Iteration: 45 Sum of squared distance: 77168190.799665 Iteration: 46 Sum of squared distance: 77167499.624465
Iteration: 47 Sum of squared distance: 77166974.633727

3.5 Visualize your result

Figure 4 shows the result image.

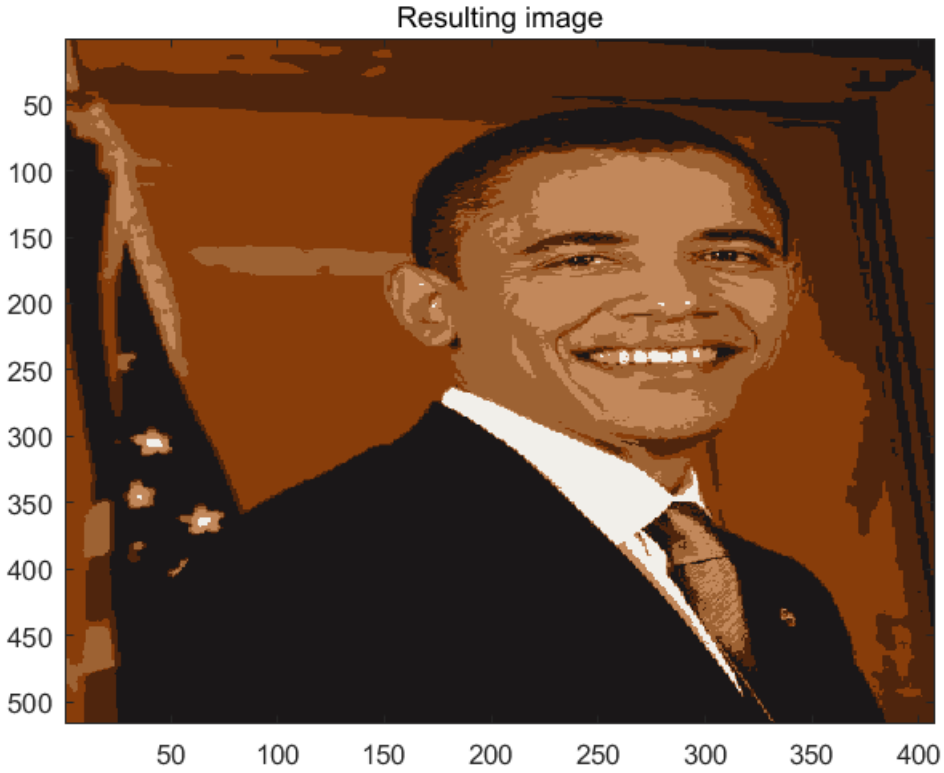

Figure 4: Result image

4 K-medoids

4.1 advantages

K-medoids algorithm is more robust to noise data or outliers. In addition, K-medoids using the pairwise distance measure instead of the sum of squared Euclidean distance type metric in K-means to evaluate

variance between data point which have more flexibility.

4.2 disadvantages

K-medoids have higher computing complexity than K-means, K-medoids is much more complex than K-means, thus it need more running time to obtain the result compare to K-means. Moreover, K-medoids converges much more slower than K-means.