1 VC dimension (20 Points) (Xun)

1. We first show \( H \) can shatter \( n + 1 \) points. Let \( S = \{x_i\}_{i=0}^n \) and \( y_i \in \{-1, 1\} \) be the label of \( x_i \). If we can place \( S \) such that \( y_i(a^\top x_i + b) \geq 0 \) holds for all \( y_i \), then \( S \) can be shattered by \( H \). Let \( x_0 = 0 \) and \( x_i \) be the unit vector on the \( i \)-th coordinate. Take \( b = y_0/2 \) and \( a_i = y_i \). Then

\[
y_0 \cdot (0 + b) = \frac{1}{2} y_0^2 \geq 0 \quad (1)
\]
\[
y_1 \cdot (y_1 + b) = y_1^2 + \frac{1}{2} y_0 y_1 \geq 0 \quad (2)
\]
\[
\vdots
\]
\[
y_n \cdot (y_n + b) = y_n^2 + \frac{1}{2} y_0 y_n \geq 0 \quad (3)
\]

always hold. Therefore \( \text{VCdim}(H) \geq n + 1 \).

Now let \( S \) contain \( n + 2 \) points, we show \( H \) cannot shatter \( S \). Let \( P = \{x : a^\top x + b \geq 0\} \) be the halfspace defined by \( h \in H \). Notice that \( S \subseteq P \implies \text{conv}(S) \subseteq P \), since

\[
a^\top \left( \sum_{i=1}^k \alpha_i x_i \right) + b = \sum_{i=1}^k \alpha_i (a^\top x_i + b) \geq 0. \quad (4)
\]

Similar for the opposite halfspace \( P^c \). Suppose \( H \) can shatter \( S \). Now \( H \) can separate any disjoint subsets \( S_1 \) and \( S_2 \) such that \( S_1 \subseteq P \) and \( S_2 \subseteq P^c \). By the claim above, this implies \( \text{conv}(S_1) \subseteq P \) and \( \text{conv}(S_2) \subseteq P^c \). However by Radon’s theorem there exist \( S_1 \) and \( S_2 \) whose convex hulls intersect. This is a contradiction. Hence \( \text{VCdim}(H) \leq n + 1 \).

2. We first show that \( H \) in \( \mathbb{R}^n \) can shatter \( 2n \) points. Pick points \( S = \{x_i, x'_i\}_{i=1}^n \), where \( x_i = e_i, x'_i = -e_i \) and \( e_i \) is the unit vector at the \( i \)-th coordinate. Let the corresponding labels be \( L = \{y_i, y'_i\}_{i=1}^n \). \( H \) can shatter \( S \) if the following can be satisfied for some small \( \epsilon > 0 \):

\[
a_i = \begin{cases} 
-1 - \epsilon & \text{if } y'_i = 1 \\
-1 + \epsilon & \text{if } y'_i = -1,
\end{cases} \quad b_i = \begin{cases} 
1 + \epsilon & \text{if } y_i = 1 \\
1 - \epsilon & \text{if } y_i = -1.
\end{cases} \quad (5)
\]

Clearly this is achievable, for instance by taking \( a_i = -1 - y'_i \epsilon \) and \( b_i = 1 + y_i \epsilon \).

Now show that \( H \) in \( \mathbb{R}^n \) cannot shatter \( 2n + 1 \) points. Given any placement of \( 2n + 1 \) points, let \( x_i^\text{min} \) and \( x_i^\text{max} \) be the points that have minimum and maximum value along the \( i \)-th coordinate. There are at most \( 2n \) such points in \( \mathbb{R}^n \), since some points might be the extremum along multiple coordinates. Then there are at least \( 1 \) point left inside the box created by the extremum points. If the internal points are labeled negative and all others are positive, then \( H \) cannot realize this labeling. Thus \( H \) in \( \mathbb{R}^n \) cannot shatter \( 2n + 1 \) points.

2 AdaBoost (30 Points) (Xun)

1. Define the correct set \( C = \{i : y_i h_t(x_i) \geq 0\} \) and the mistake set \( M = \{i : y_i h_t(x_i) < 0\} \).

\[
Z_t = \sum_{i=1}^m D_t(i) e^{-\alpha_i y_i h_i(x_i)} = \sum_{i \in C} D_t(i) e^{-\alpha_i} + \sum_{i \in M} D_t(i) e^{\alpha_i} = (1 - \epsilon_t) \cdot e^{-\alpha_t} + \epsilon_t \cdot e^{\alpha_t}. \quad (6)
\]
\[
\text{err}_{D_{t+1}}(h_t) = \sum_{i=1}^m D_{t+1}(i) 1_{y_i \neq h_t(x_i)} = \sum_{i \in M} \frac{D_t(i)}{Z_t} e^{\alpha_t} = \epsilon_t \cdot \frac{1}{2 \epsilon_t} = \frac{1}{2}. \quad (7)
\]
2. Expand $D_t(i)$ recursively.

$$D_{T+1}(i) = \frac{D_T(i)}{Z_T} e^{-\alpha_T y_i h_T(x_i)}$$

$$= \frac{D_{T-1}(i)}{Z_{T-1}} e^{-\alpha_{T-1} y_i h_{T-1}(x_i)} \cdot \frac{1}{Z_T} e^{-\alpha_T y_i h_T(x_i)}$$

$$= \frac{D_1(i)}{\prod_{t=1}^T Z_t} e^{-\sum_{t=1}^T \alpha_t y_i h_t(x_i)}$$

$$= \frac{1}{m \cdot \prod_{t=1}^T Z_t} e^{-y_i f(x_i)}.$$

3. Make use of the fact that exponential loss upper bounds the 0-1 loss: $1_{\{x<0\}} \leq e^{-x}$.

$$\text{err}_S(H) = \frac{1}{m} \sum_{i=1}^m 1_{y_i f(x_i)<0} \leq \frac{1}{m} \sum_{i=1}^m e^{-y_i f(x_i)} = \sum_{i=1}^m D_{T+1}(i) \prod_{t=1}^T Z_t = \prod_{t=1}^T Z_t.$$ (13)

4. Make use of the fact that $1-x \leq e^{-x}$.

$$\prod_{t=1}^T Z_t = \prod_{t=1}^T 2\sqrt{\epsilon_t (1-\epsilon_t)} = \prod_{t=1}^T \sqrt{1-4\gamma_t^2} \leq \prod_{t=1}^T e^{-2\gamma_t} = e^{-2\sum_{t=1}^T \gamma_t^2}.$$ (14)

5. From the result above, $\text{err}_S(H) \leq e^{-2\sum_{t=1}^T \gamma_t^2} \leq e^{-2T\gamma^2}$ $\xrightarrow{T \to \infty} 0$. Therefore

$$e^{-2T\gamma^2} \leq \varepsilon \implies T \geq \frac{1}{2\gamma^2} \log \frac{1}{\varepsilon},$$ (15)

hence we need $T = O(\frac{1}{\gamma^2} \log \frac{1}{\varepsilon})$.

6. See Table 1 and Figure 1. The red, green, and blue regions are the halfspaces defined by $h_1$, $h_2$, and $h_3$. The code is available on the course website.

<table>
<thead>
<tr>
<th>$t$</th>
<th>$\epsilon_t$</th>
<th>$\alpha_t$</th>
<th>$D_t(1)$</th>
<th>$D_t(2)$</th>
<th>$D_t(3)$</th>
<th>$D_t(4)$</th>
<th>$D_t(5)$</th>
<th>$D_t(6)$</th>
<th>$D_t(7)$</th>
<th>$D_t(8)$</th>
<th>$D_t(9)$</th>
<th>$\text{err}_S(H)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.222</td>
<td>0.626</td>
<td>0.111</td>
<td>0.111</td>
<td>0.111</td>
<td>0.111</td>
<td>0.111</td>
<td>0.111</td>
<td>0.111</td>
<td>0.111</td>
<td>0.111</td>
<td>0.222</td>
</tr>
<tr>
<td>2</td>
<td>0.143</td>
<td>0.896</td>
<td>0.071</td>
<td>0.071</td>
<td>0.071</td>
<td>0.071</td>
<td>0.071</td>
<td>0.071</td>
<td>0.071</td>
<td>0.250</td>
<td>0.250</td>
<td>0.222</td>
</tr>
<tr>
<td>3</td>
<td>0.125</td>
<td>0.973</td>
<td>0.042</td>
<td>0.042</td>
<td>0.042</td>
<td>0.250</td>
<td>0.250</td>
<td>0.042</td>
<td>0.042</td>
<td>0.146</td>
<td>0.146</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Table 1: AdaBoost results
3 Gaussian Mixture Model

1

\[ E[x] = \int xp(x)dx \]
\[ = \int x \sum_{k=1}^{K} \pi_k \mathcal{N}(x | \mu_k, \Sigma_k) dx \]
\[ = \sum_{k=1}^{K} \pi_k \int x \mathcal{N}(x | \mu_k, \Sigma_k) dx \]
\[ = \sum_{k=1}^{K} \pi_k \mu_k \tag{16} \]

2

\[ \text{Cov}[x] = \mathbb{E}[xx^T] - \mathbb{E}[x]\mathbb{E}[x]^T \]
\[ = \int xx^T \sum_{k=1}^{K} \pi_k \mathcal{N}(x | \mu_k, \Sigma_k) dx - \mathbb{E}[x]\mathbb{E}[x]^T \]
\[ = \sum_{k=1}^{K} \pi_k \int xx^T \mathcal{N}(x | \mu_k, \Sigma_k) dx - \mathbb{E}[x]\mathbb{E}[x]^T \]
\[ = \sum_{k=1}^{K} \pi_k \mathbb{E}_k[xx^T] - \mathbb{E}[x]\mathbb{E}[x]^T \]
\[ = \sum_{k=1}^{K} \pi_k (\Sigma_k + \mu_k \mu_k^T) - \mathbb{E}[x]\mathbb{E}[x]^T \tag{17} \]

where I denote \[ \mathbb{E}_k[x] = \int x \mathcal{N}(x | \mu_k, \Sigma_k) dx. \]
4  K-means

4.1  Proof:
\[ \sum_{x \in X} \|x - s\|^2 - \sum_{x \in X} \|x - \bar{x}\|^2 = \sum_{x \in X} (2x - s)(\bar{x} - s) \]
\[ = |X|(2\bar{x} - s)(\bar{x} - s) \]
\[ = |X| \cdot \|\bar{x} - s\|^2 \]  \hspace{1cm} (18)

2  Proof:
\[ \sum_{i=1}^{n_k} \sum_{j=1}^{n_k} \|x_{ki} - x_{kj}\|^2 \]
\[ = \sum_{i=1}^{n_k} \left( \sum_{j=1}^{n_k} \|x_{kj} - \mu_k\|^2 + n_k \|\mu_k - x_{ki}\|^2 \right) \]
\[ = n_k \sum_{j=1}^{n_k} \|x_{kj} - \mu_k\|^2 + \sum_{i=1}^{n_k} n_k \|\mu_k - x_{ki}\|^2 \]
\[ = 2n_k \sum_{i=1}^{n_k} \|\mu_k - x_{ki}\|^2 \]

Therefore,
\[ \sum_{i=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{j=1}^{n_k} \|x_{ki} - x_{kj}\|^2 \]
\[ = 2 \sum_{i=1}^{K} \sum_{i=1}^{n_k} \|\mu_k - x_{ki}\|^2 \]  \hspace{1cm} (20)

Proved.

3  In Step 1, as we fix the centroids, when we reassign the class memberships, every point \(x_i\) will find its new nearest centers, thus decreases the objective \(\omega\). In Step 2, we fix the class memberships and re-estimate the class centers. With Lemma 1 we know by replacing the old center with a new center we will decrease the objective.

4  if \(K > n\), we just set the centers as the points themselves which will give us a zero objective. if \(K < n\), we create a new cluster by picking any point \(x\) in the dataset which is not a center, and let \(x\) be the center. Denote the new memberships as \(f'\) and \(\mathcal{U}_{K+1} = \mathcal{U}_K + \{x\}\), then
\[ \Omega(K) \geq \omega(\mathcal{U}_{K+1}, f'; \mathcal{X}) \geq \Omega(K + 1) \]  \hspace{1cm} (21)
Proved.
Since there are at most $k^n$ assignments of points to cluster centres, the above objective can only achieve one of $k^n$ different values and one of $k^n$ different assignments. Therefore, it has to terminate in a finite number of steps as the objective is non-increasing.

4.2

1

See the code.

2

min objective: $2.0614e+09$. See the objective v.s. iterations in Fig.2. Some runs converged, but some not due to randomness. The mean faces are visualized as in Fig.3.

3

See the code. See the objective v.s. iterations in Fig.4. Most converged. The mean faces are visualized as in Fig.5. With Kmeans++, the objectives converged faster and better.
Figure 3: The mean faces of kmeans.

Figure 4: The objective v.s. iterations for kmeans++.
Figure 5: The mean faces of kmeans++. 