1 Question 1

This question is answered with the following assumptions.

- $-1 \equiv$ *false* and $+1 \equiv$ *true*.
- In the second subproblem the original intended formula was $\sum_{i=1}^{n} w_i x_i \geq w_0$

a. In conjunctions like $x_1 \wedge \bar{x_2} \wedge x_3$, as soon as we determine a literal is *false*, the whole expression evaluates to *false*. For the disjunctions like $x_1 \vee \bar{x_2} \vee x_3$, as soon as we determine a literal is *true*, the whole expression evaluates to *true*. Based on this observation, we can construct a decision list such that each ℓ_i corresponds to a literal in our expression.

   - For the conjunctions case, we negate each literal i and set $b_i =$ *false* and $b_{n+1} =$ *true*. So when a literal evaluates to *false* we give *false* right away, and only when all the literals are *true* we return *true*.
   - Similarly for disjunctions, we use the exact literal i and set $b_i =$ *true* and $b_{n+1} =$ *false*. So when a literal is *true* we give *true* right away, and only when all the literals are *false* we return *false*.
   - For the conjunctions example we'll have "if $\bar{x_1}$ then -1, else if x_2 then -1, else if $\bar{x_3}$ then -1, else 1". For the disjunctions example we'll have "if x_1 then $+1$, else if $\bar{x_2}$ then $+1$, else if x_3 then $+1$, else -1.

b. To prove that it suffices to show a conversion from decision list to an affine threshold function. I provide a constructive proof and demonstrate correctness through an induction on n, number of the literals in our decision list (also referred to as list length). Note that for easier proof, x_i's are replaced with $\frac{x_i+1}{2}$ so that $x_i = 0$ corresponds to *false* and $x_i = 1$ as before, corresponds to *true*. After we're done with constructing a solution for this new domain, we can simply replace x_i's in the formula with $2.x_i - 1$ to go to the previous domain. Obviously, the resulting function will still be an affine threshold function.

   Proposition. Let P_n be the proposition that for all the decision lists of length n, we can set values w_i such that the affine threshold function $f(x) = 1$ iff $\sum_{i=1}^{n} w_i x_i \geq w_0$ expresses the same information as our decision list.

   Basis. For conciseness, I consider the P_0 as the basis. In this case, we simply have either *true* or *false*. I could also consider P_1 as the basis, but the possibilities are significantly greater and I will have to repeat the same consideration in the **Proof** part anyways.

   (a) if *true*, then $0 \geq w_0$, so we set $w_0 = 0$.

(b) if *false*, then $0 < w_0$, so we set $w_0 = 1$.

Hypothesis. Let's assume P_n is *true*.

Proof. I will prove P_{n+1} is also true. For any decision list of length $n+1$, we omit the first 'if ℓ_1 then b_1, else' to get a decision list of length n which looks like 'if ℓ_2 then b_2, else if …else b_m'. Based on the hypothesis, we know there exists $w_0, w_2, w_3 \dots$ such that gives an affine threshold function expressing the same information. Now we construct the solution for our primary decision list based on the answer to the subproblem as follows. For each possible ℓ_1 and b_1 we will adjust w_1 and w_0 in a way to enforce the final outcome b_1 for when the condition is satisfied. When the condition is not satisfied, the resulting formula will be the same as when we had the last n decisions.

- $\ell_1 = x_1$ and $b_1 = 1$. Then we know $w_1x_1 + \sum_{i=2}^{n+1} w_ix_i \geq w_0$ is true when $x_1 = 1$ and is left to be determined by the rest of formula when $x_1 = 0$. Hence,

$$w_1 + \sum_{i=2}^{n+1} w_ix_i \geq w_0 \Rightarrow w_1 \geq w_0 - \sum_{i=2}^{n+1} w_ix_i \Rightarrow w1 \geq w_0 - \sum_{i=2}^{n+1} |w_i|$$

$$\Rightarrow w_1 = w_0 - \sum_{i=2}^{n+1} |w_i| + 1.$$

As a result, when $x_1 = 1$ we are sure the left hand side is $\geq w_0$. When $x_1 = 0$, the resulting formula will be the same as when we only had n decisions, because the coefficient w_1 will have no effect.

- $\ell_1 = \bar{x_1}$ and $b_1 = 1$. Then $w_1(1 - x_1) + \sum_{i=2}^{n+1} w_ix_i \geq w_0$. $(1 - x_1)$ results from the fact that we need w_1 to be active when $x_1 = 0$, and inactive when $x_1 = 1$. The added expression will contribute to the value of w_0 at the end. When $x_1 = 0$,

$$w_1 + \sum_{i=2}^{n+1} w_ix_i \geq w_0 \Rightarrow w_1 \geq w_0 - \sum_{i=2}^{n+1} w_ix_i \Rightarrow w1 \geq w_0 - \sum_{i=2}^{n+1} |w_i|$$

$$\Rightarrow w_1 = w_0 - \sum_{i=2}^{n+1} |w_i| + 1.$$

$$w_1(1 - x_1) + \sum_{i=2}^{n+1} w_ix_i \geq w_0 \Rightarrow -w_1x_1 + \sum_{i=2}^{n+1} w_ix_i \geq (w_0 - w_1)$$

$$\Rightarrow w_1' = -(w_0 - \sum_{i=2}^{n+1} |w_i| + 1), w_0' = w_0 + w_1'.$$

So we simply set the w_1 to be $-(w_0 - \sum_{i=2}^{n+1} |w_i| + 1)$ and the new value for w_0 to be $w_0 + w_1$.

- $\ell_1 = x_1$ and $b_1 = -1$. Then we know $w_1 x_1 + \sum_{i=2}^{n+1} w_i x_i \geq w_0$ is false when $x_1 = 1$. So

$$w_1 + \sum_{i=2}^{n+1} w_i x_i < w_0 \Rightarrow w_1 < w_0 - \sum_{i=2}^{n+1} |w_i|$$
$$\Rightarrow w_1 = w_0 - \sum_{i=2}^{n+1} |w_i| - 1.$$

- And finally $\ell_1 = \bar{x_1}$ and $b_1 = -1$ will be similarly solved.

Conclusion. Assuming P_n holds, for any decision list of length $n + 1$, I first remove the first condition to get a decision list of length n. Given the assumption, there exists $w_0, w_2, w_3, \ldots$ such that forms an affine threshold function subject to the constraints which expresses the same information as our n length decision list. Then I extended that answer to take into consideration the first condition that I removed using the previously shown methods. In the new formula, when the first condition of the decision list is not met, we will get the same answer for the n length decision list. For the case where the first condition is met, we are assured the result will be the same as the result of the decision list. Therefore P_{n+1} also holds true.

After we have set the parameters $w_0, w_1, \ldots$, we will replace x_i's with $2.x_i - 1$ to get the answer in the original domain where $x_i \in \{-1, 1\}$. The resulting function is still an affine threshold function and w_i's can be updated accordingly.

2 Question 2

We first show there exists a setting for 6 points so that it can be shattered, and then we show any 7 points in our space can't be possibly shattered which results in VC-dimensionality of 6.

For the first part, we put a point in center of each 6 surfaces of a hypothetical cube. It's not difficult to see all 2^6 possible positive colorings can be easily put into a single cube, which leads to knowing that there exists a setting for 6 points which can be shattered.

For the second part, we show any set containing 7 or more points can't be shattered. If we have more than 7 points, for any subset S with $|S| = 7$, the following argument applies. We select a minimal axis aligned box of these points. Since it's a minimal box, all surfaces of this box must include at least one point. If we have at least one interior point in this cube, we can set this point negative, and all others positive. It can be easily seen that no consistent h exists for this case. If there are no interior points in this cube, than at least one of the surfaces includes more than one point. If P_1 and P_2 are on the same surface, the minimal box selected doesn't depend on exactly one of them (geometrical argument), meaning that if we remove that point, we will still end up with the same minimal box. We simply color that point negative and all others positive. It's easy to see no consistent h can be found again. Therefore, the VC-dimensionality of axis aligned boxes is 6.

3 Question 3

Following the same steps of the 3rd lecture we first show, $Pr[A] < 2.Pr[B]$, then we prove $Pr[B] \leq \delta/2$.

$$Pr[B] \geq Pr[A \cap B] = Pr[A].Pr[B|A]$$

If we show $Pr[B|A] > 1/2$, we have proved the first claim. Instead, we prove the equivalent claim $Pr[\bar{B}|A] \leq 1/2$.

$Pr[\bar{B}|A]$ is the case where $|R(h) - \hat{R}_S(h)| > \epsilon$ and $|\hat{R}_S(h) - \hat{R}_{S'}(h)| < \epsilon/2$. To prove probability of this event is less than 1/2 we will be using the Chernoff bound.

Define $X_i = 1$ iff the i_{th} data in S' disagrees in label with that of hypothesis h which was selected in event A for S. Given that X_i's are independent samples we have:

$$\mathbb{E}[X_i] = Pr[X_i] = R(h) \Rightarrow \sum_{i=1}^{m} \mathbb{E}[X_i] = mR(h)$$

Now we use the Chernoff bound to determine the probability of deviation of sum of disagreements from $mR(h)$ by $\epsilon/4$ above and below. If show this probability is $\leq 1/2$ we will be done with the first part.

Selecting the average $\mu = mR(h)$ and $\delta = \epsilon/4$, we will have:

$$\begin{aligned}
Pr[\sum X_i \geq (1+\delta)\mu_{max}] &\leq e^{-\delta^2\mu_{max}/3} \quad \mu = mR(h), \delta = \epsilon/4 \\
Pr[\sum X_i \geq \mu_{max} + \epsilon/4] &\leq e^{-\epsilon^2 mR(h)/48} \quad \text{For going above.(*)} \\
Pr[\sum X_i \leq (1-\delta)\mu_{min}] &\leq e^{-\delta^2\mu_{min}/2} \quad \mu = mR(h), \delta = \epsilon/4 \\
Pr[\sum X_i \leq \mu_{min} - \epsilon/4] &\leq e^{-\epsilon^2 mR(h)/32} \quad \text{For going below.(**)}
\end{aligned}$$

Plugging the minimum value for m we will get:

$$(*) \Rightarrow \leq e^{-R(h)(log(\Pi_H(m))+log(2/\delta))} \leq (\frac{\delta}{2\Pi_H(m)})^{R(h)} \leq \frac{\delta}{2\Pi_H(m)}$$
$$(**) \Rightarrow \leq e^{-R(h)(log(\Pi_H(m))+log(2/\delta))} \leq (\frac{\delta}{2\Pi_H(m)})^{R(h)} \leq \frac{\delta}{2\Pi_H(m)}$$
$$\bigcup_{h \in H} \frac{\delta}{\Pi_H(m)} = \delta$$

Assuming $\delta \leq 1/2$, our proof for the first part is complete.

For the second part:

$$\text{let event } B =$$
$$\{\text{ labeling y } \in \Pi_H(S \cup S')| \text{ such that causes the number of difference of disagreements to be } \geq m\epsilon/2\}$$

If there are d points incorrectly classified, we expect $d/2$ of them to be in S. Now we need to bound probability of going above and below that by $\epsilon/4$ to provide a guarantee of deviating by at least $\epsilon/2$. We also know that $d \geq m\epsilon/2$ (if less, B doesn't happen).

Using the Chernoff bound again we'll get:

$$Pr[\sum X_i \geq (1+\delta)\mu_{max}] \leq e^{-\delta^2\mu_{max}/3} \quad \mu = m\epsilon/4, \delta = \epsilon/4$$
$$Pr[\sum X_i \geq \mu_{max} + \epsilon/4] \leq e^{-\epsilon^2 m\epsilon/192} \quad \text{For going above.(*)}$$
$$Pr[\sum X_i \leq (1-\delta)\mu_{min}] \leq e^{-\delta^2\mu_{min}/2} \quad \mu = m\epsilon/4, \delta = \epsilon/4$$
$$Pr[\sum X_i \leq \mu_{min} - \epsilon/4] \leq e^{-\epsilon^2 m\epsilon/128} \quad \text{For going below.(**)}$$

Plugging the minimum value for m we will get:

$$(*) \Rightarrow \leq e^{-R(h)(log(\Pi_H(m))+log(2/\delta))} \leq (\frac{\delta}{2\Pi_H(2m)})^{\epsilon/4} \leq \frac{\delta}{4\Pi_H(2m)}$$
$$(**) \Rightarrow \leq e^{-R(h)(log(\Pi_H(m))+log(2/\delta))} \leq (\frac{\delta}{2\Pi_H(2m)})^{\epsilon/4} \leq \frac{\delta}{4\Pi_H(2m)}$$
$$\bigcup_{h \in H} \frac{\delta}{2\Pi_H(2m)} = \delta/2$$