5-6. (a) \( \mathbb{E}_\mathbf{z}(F) = \mathbb{E}_\mathbf{z} \left[ \sup_{f \in \mathbb{F}} \frac{1}{m} \sum_{i=1}^{m} f_i(Z_i) \right] = \mathbb{E}_\mathbf{z} \left[ \sup_{f \in \mathbb{F}} \frac{1}{m} \sum_{i=1}^{m} J_i(W_i, X_i) \right] \)

\[ = \mathbb{E}_\mathbf{z} \left[ \sup_{f \in \mathbb{F}} \frac{1}{m} \left( \sum_{i=1}^{m} f_i(W_i, X_i) \right) \right] \]

\[ = \mathbb{E}_\mathbf{z} \left[ \frac{1}{m} \left( \sum_{i=1}^{m} f_i(W_i, X_i) \right) \right] \]

\[ = \frac{1}{m} \mathbb{E}_\mathbf{z} \left[ \sum_{i=1}^{m} f_i(W_i, X_i) \right] \]

where 1 comes from definition
2 comes from Cauchy-Schwarz inequality and the equality can be achieved
3 comes from Jensen's inequality
4 comes from \( P(J_i = 1) = P(J_i = -1) = \frac{1}{2} \) and \( J_i, J_j \) are iid
5 comes from \( \|X_i\|_2 \leq 1 \).

(b) From Theorem 5.7, we know that with probability at least 1-\( \delta \), for all weight vectors \( \mathbf{w} \in \mathbb{R}^n \) with \( \|\mathbf{w}\|_2 = 1 \),

\[ P_F [y(w \cdot x) \leq \mathbb{E} + \mathcal{R}_S(\emptyset, F)] \leq P_F [y(w \cdot x) \leq \Theta] + 2\mathcal{R}_S(\emptyset, F) + \frac{2m}{n} \]

where \( \Theta \) comes from theorem with \( \emptyset \) as in Figure 5.3
3 comes from (5.25) and (a)

-7 Consider the segment \([0, 1]\) with the uniform distribution, i.e. \( x \sim \text{Unif}[0, 1] \).
Let \( Y = \begin{cases} 1 & \text{if } x \text{ is rational} \\ -1 & \text{if } x \text{ is irrational} \end{cases} \)

For any training example, we assign a delta function to it. Hence, every training example is correctly labeled. Larger than \( m \)

We can then assign a big positive constant if \( x \) is not in the training set.
The combined classifier would be a delta function plus a big positive constant

where the domain of every delta function is every training point and

domains of the big constant function is everything except the training pair.

Therefore, the training error is zero and the generalization error is

\( P(x \text{ rational}) P(\text{error} | x \text{ rational}) + P(x \text{ irrational}) P(\text{error} | x \text{ irrational}) \)

\[ = P(\text{error} | x \text{ irrational}) = 1 \]

. Generalization error is 100%.

The rational/irrationals are not needed.

The target function is constant, say 1, and the rule is the constant 0 with delta function on training variables.
8. (a) Let's consider the labeling: \[ +---+++---++ \]

We can group same and consecutive labeling together to get: \[ +-----+ \]

Then we can do the majority vote: \[ +-----+ \]

\[ +-----+ \] each one can be implemented by a decision stump.

\[ +-----+ \]

\[ +-----+ \] majority vote of each column

It means that if we see \(+\) (or \(-\)) in the grouped labeling, then we can use the decision stump such that everything before and including it should have the same sign. Everything following it should have different sign.

The above procedure holds if we have \(2k-1\) groups where \(k \geq 1\).

If we have \(2k\) groups where \(k \geq 1\), say \(+----\), it should be careful to deal with it.

If the groups are even, then we can get rid of the last one. Then redo all the things as in the odds groups case and finally append the last group labeling different from the one before. (or we can use decision stumps to isolate the last group.

(b) Let's consider \(n=2\).

Let's consider the labeling \[ +---+ \]

Without loss of generality, let's think the combined classifier:

\[ \text{sign}(w_1 I(x_1 < 0) + w_2 I(x_2 < 0) + w_3 I(x_1 < 0)) \]

\[ I(x_i < 0) = \begin{cases} 1 & x_i < 0 \\ -1 & \text{otherwise} \end{cases} \]

In figure 1, we need \( w_1 - w_2 + w_3 > 0 \) --- \( \Box \)

\( w_1 + w_2 + w_3 < 0 \) --- \( \therefore \)

\( -w_1 - w_2 + w_3 < 0 \) --- \( \therefore \)

\( -w_1 + w_2 + w_3 > 0 \) --- \( \therefore \)

From \( \Box \) and \( \therefore \), we need \( w_3 > 0 \) and from \( \therefore \) and \( \therefore \), we need \( w_3 < 0 \).

Thus, we get a contradiction, so it's impossible to use decision stumps to get the desired consistent result.

Same arguments hold for dimensions, \( n \geq 2 \). (Neighbors have different labeling)

8. (a) It's better to think that we can isolate each group labeling by decision stumps. For example, \[ +++-+---+++-++---+ \]

So, the answer is positive for \( n=1 \).