Caio Brighenti
COSC 480 - Learning From Data
Fall 2019
Problem Set 8

1. Given hypothesis set with VC dimension 4 and 40,000 observations, how likely are you to have a generalization error of at most 0.1?

   We have that $N = 40,000$, $d_{vc} = 4$, and are concerned with an error of at most 0.1, so $\epsilon = 0.1$. We can use the VC inequality to solve this as follows:

   $$Pr(|E_{in}(g) - E_{out}(g)| \geq \epsilon) \leq 4m_h(2N)e^{-\frac{1}{8}\epsilon^2 N} \quad \text{VC inequality} \quad (1)$$
   $$\leq 4(2N+1)e^{-\frac{1}{8}\epsilon^2 N} \quad \text{substitute growth function} \quad (2)$$
   $$\leq .0316 \quad \text{plug in variables and calculate} \quad (3)$$

   We thus have the probability that the error is greater than 0.1 is 0.0316. The probability that the error is at most 0.1 is thus $1 - 0.0316 \approx .97$.

2. Suppose you have a hypothesis set with VC dimension 2, 1000 training examples, and $E_in = 0.06$. Give an upper bound with 95% confidence for E_out.

   We are given that $d_{vc} = 2$ and $N = 1000$. Let σ be the probability given by the VC inequality such that $\delta = 4m_h(2N)e^{-\frac{1}{8}\epsilon^2 N}$. We thus have an expression in terms of δ, N, d_vc and ϵ. Given that we are considering a upper bound for 95% confidence, we have that $\sigma = 0.05$. We thus have values for all these variables, and thus can solve for ϵ as follows.

   $$4m_h(2N)e^{-\frac{1}{8}\epsilon^2 N} = \delta \quad \text{VC inequality} \quad (4)$$
   $$ln(4m_h(2N)) + ln(e^{-\frac{1}{8}\epsilon^2 N}) = ln(\delta) \quad \text{ln both cards} \quad (5)$$
   $$\epsilon = \sqrt{\frac{-8}{N} ln(\frac{\sigma}{4m_h(2N)})} \quad \text{algebra} \quad (6)$$
   $$\approx 0.389 \quad \text{plug in variables and calculate} \quad (7)$$

   We can thus bound the error as 0.389, meaning the misclassification rate could be as high as 0.439.

3. Suppose you have 10,000 training examples, a hypothesis set with VC dimension 2, and that $E_in = 0.06$. Give a low bound with 90% confidence for E_in. We approach this problem in the same way as above, given that we have $N = 1000$, $d_{vc} = 2$, and $\sigma = 0.1$. We can thus plug these values into the previous equation as follows:

   $$\epsilon = \sqrt{\frac{-8}{N} ln(\frac{\sigma}{4m_h(2N)})} \quad \text{algebra} \quad (8)$$
   $$\approx 0.301 \quad (9)$$

   Thus the error can be as low as $0.06 - 0.301 = -0.241$, which suggests that our bounds are too large to be meaningful.

4. Suppose you have a hypothesis set with VC dimension 5 and you want to be 95% certain the generalization error is less than 0.1. How many observations do you need?

   We have that $d_{vc} = 5$ and want to be 95% sure that the generalization error ($|E_{in} - E_{out}|$) is less than 0.1. We use the VC inequality to solve for N when $\epsilon = 0.1$.

   $$\sigma \leq 4m_h(2N)e^{-\frac{1}{8}\epsilon^2 N} \quad \text{VC inequality} \quad (10)$$
   $$\sigma \leq (8N^{d^{vc}} + 4)e^{-\frac{1}{8}\epsilon^2 N} \quad \text{substitute} \quad (11)$$

As we have multiple N terms, we cannot simply solve for it. Instead we plug in several values for N until we find the correct value for N that approaches $\sigma = 0.5$. This approach finds that the right value for N is approximately 445-450.

5. .

(a) Which bound on E_{out} provides a more accurate estimate of the true performance?

We solve for the error bound on the train and test set separately. Given that we have a single hypothesis with respect to the test set, we can use Hoeffding's inequality as follows.

$$\sigma = 2e^{-2\epsilon^2 N} \quad \text{Hoeffding's inequality} \tag{12}$$

$$ln(\sigma) = ln(2) + ln(e^{-2\epsilon^2 N}) \quad \text{ln both sides} \tag{13}$$

$$-2\epsilon^2 N = ln(\frac{\sigma}{2}) \quad \text{algebra} \tag{14}$$

$$\epsilon = \sqrt{\frac{ln(\frac{\sigma}{2})}{2N}} \quad \text{algebra} \tag{15}$$

$$\epsilon \approx 0.122 \tag{16}$$

We now find the bound for the training set. For this, we must consider all hypotheses as they were chosen based on the train set. We thus use Hoeffding's inequality for multiple hypotheses as follows.

$$\sigma = M2e^{-2\epsilon^2 N} \quad \text{Hoeffding's inequality} \tag{17}$$

$$ln(\sigma) = ln(2M) + ln(e^{-2\epsilon^2 N}) \quad \text{ln both sides} \tag{18}$$

$$-2\epsilon^2 N = ln(\frac{\sigma}{2M}) \quad \text{algebra} \tag{19}$$

$$\epsilon = \sqrt{\frac{ln(\frac{\sigma}{2M})}{-2N}} \tag{20}$$

$$\epsilon \approx 0.128 \tag{21}$$

We can see that the test set produces a slightly more accurate estimation of the out of sample error.

(b) Repeat the previous, splitting the size of the train and test sets.

For the test set, we use the previous expression $\epsilon = \sqrt{\frac{ln(\frac{\sigma}{2})}{2N}}$ to approximate $e \approx 0.071$. For the train set, we use the expression $\epsilon = \sqrt{\frac{ln(\frac{\sigma}{2M})}{-2N}}$ to approximate $\epsilon \approx 0.223$. In this case, the test set provides the best estimate of the error.

(c) Explain why the first scenario is preferred.

When learning from data, we have two objectives with respect to E_{in} and E_{out}. Firstly, we want E_out to be small, meaning that we have a model that accurately estimates the population of interest. As we cannot estimate this directly, we are concerned with the generalization error, meaning how good of an approximation is E_{in} is for E_{out}.
We have seen that the second scenario provides the lowest generalization error, meaning we have the most accurate estimate of E_{out}. This is certainly desirable, but says nothing of how low the error actually is. We might have a better estimate of E_{out}, but it is unlikely that we have a good model. It is likely not worth trading the better generalizability for the decreased model accuracy.