Homework 3. Machine Learning (MIIS)

Marc Juvillà Garcia

November 12, 2016

1. When is the bound of Theorem 3 better than the bound of Theorem 2 presented in the lecture? (What is the size of the smallest hypothesis class for which the former is better than the latter?)

   Theorem 2 (Halving):

   $$\sum_{t=1}^{n} l(x_{\mathrm{t}}, y_{\mathrm{t}}, \hat{y}_{\mathrm{t}}) \leq \log_2 |H| \tag{1}$$

   Theorem 3 (Randomized Hypothesis Elimination):

   $$E\left[\sum_{t=1}^{n} l(x_{\mathrm{t}}, y_{\mathrm{t}}, \hat{y}_{\mathrm{t}})\right] \leq 1 + \ln |H| \tag{2}$$

   Solve:

   $$1 + \ln |H| < \log_2 |H| \tag{3}$$

   Knowing that:

   $$\log_2 |H| = \frac{ln|H|}{ln(2)} \tag{4}$$

   If we plug (4) into (3):

   $$1 + \ln |H| < \frac{ln|H|}{ln(2)} \tag{5}$$

   And isolate $|H|$:

   $$|H| > e^{\frac{ln(2)}{1-ln(2)}} \approx 9,57 \tag{6}$$

   That is, the loss bound of RHE is better than the one from Halving when the hypothesis class has at least 10 items.

2. What is the computational complexity of Randomized Hypothesis Elimination? (Hints: How expensive is it to generate a random hypothesis and test if it is in the version space? How many random hypotheses do we need to generate until we find a consistent one?)

   As Randomized Hypothesis Elimination draws only one hypothesis from the model class on each iteration, it has the same complexity as Hypothesis Elimination, plus maybe the cost of computing this random hypothesis, but I think this cost is negligible. For this reason, it only needs to find a consistent hypothesis instead of checking them all (like Halving does).

3. Construct a hypothesis class and a sequence of examples for which Hypothesis Elimination (the non-randomized one) makes exactly $|H| - 1$ mistakes. (Hint: a possible good idea is choosing the $\mathrm{x_t}$'s from the natural numbers 1, 2, ..., $|H|$ and consider the hypothesis class H consisting of functions of the form $h_{\mathrm{k}}(x) = I\{x \leq k\}$ (i.e., $h_{\mathrm{k}}(x) = 1$ if $x \leq k$ and $h_{\mathrm{k}}(x) = 0$ otherwise) for all k = 1, 2, ..., $|H|$.)

Strategy: We have an hypothesis class of size $|H|$. Let $\mathrm{h}_{|H|}$ be the target function (that is, $h_{|H|}(x) = I\{x \leq |H|\}$), and $\mathrm{x_t}$ be a sequence of natural numbers in the traditional order.
For $x_1 = 1$ we will draw h_1, which will predict 1 correctly.
For $x_2 = 2$ we will draw h_1 again, which will predict incorrectly. h_1 will be removed from the hypothesis class.
At $x_{|H|}$ we will be incorrect again but we will delete the last wrong hypothesis, leaving the class hypothesis with just one item: the target hypothesis.

Example:

If we create an hypothesis class in which $h_{\mathrm{k}}(x) = 1$ if $x \leq k$ and 0 otherwise we will have these hypothesis:

$h_1(x) = 1\{x \leq 1\}$
$h_2(x) = 1\{x \leq 2\}$
$h_3(x) = 1\{x \leq 3\}$

For a $|H|$ of 10 and target function H_{10}:

$t_1 : x_1 = 1 \rightarrow$ draw $H_1 \rightarrow H_1(x_1) = 1 \rightarrow y_1 = 1 \rightarrow H_1(x_1) = y_1 \rightarrow$ keep H_1

$t_2 : x_2 = 2 \rightarrow$ draw $H_1 \rightarrow H_1(x_2) = 0 \rightarrow y_1 = 1 \rightarrow H_1(x_2) \neq y_2 \rightarrow$ delete H_1

$t_3 : x_3 = 3 \rightarrow$ draw $H_2 \rightarrow H_2(x_3) = 0 \rightarrow y_3 = 1 \rightarrow H_2(x_3) \neq y_3 \rightarrow$ delete H_2

(...)

$t_{10} : x_{10} = 10 \rightarrow$ draw $H_9 \rightarrow H_9(x_{10}) = 0 \rightarrow y_{10} = 1 \rightarrow H_9(x_{10}) \neq y_{10} \rightarrow$ delete H_9

$t_{11} : x_{11} = 11 \rightarrow$ draw $H_{10} \rightarrow H_{10}(x_{11}) = 1 \rightarrow y_{11} = 1 \rightarrow H_{10}(x_{11}) = y_{11} \rightarrow$ keep H_{10}

The classifier mistook every hypothesis from t_2 to t_{10} (that is, 9 mistakes, $|H| - 1$).

4. Construct a hypothesis class and a sequence of examples for which Halving makes exactly $log_2|H|$ mistakes. (Hint: The same hypothesis class as above and the same pool of $\mathrm{x_t}$'s may work here too.)

   Using the same hypothesis class, the same target function but not the same pool of $\mathrm{x_t}$'s as before, and breaking voting ties with $\hat{y}_t = 0$.

   The sequence of observations should be:

   $x_t = floor(mean(H_t) + 1)$

   That is, we choose the hypothesis for which the lower hypothesis will be wrong and also majority (at least until just one hypothesis -the correct one- is left) or will tie with the higher hypothesis, in which case we will break the tie with $\hat{y}_t = 0$

   Example:

   $t_1 : x_1 = 6 \rightarrow H_1 - H_5$ predict 0, $H_6 - H_{10}$ predict $1 \rightarrow \hat{y}_1 = 0 \neq y\_1 = 1 \rightarrow$ delete $H_1 - H_5$

   $t_2 : x_2 = 9 \rightarrow H_6 - H_8$ predict 0, $H_9 - H_{10}$ predict $1 \rightarrow \hat{y}_2 = 0 \neq y_2 = 1 \rightarrow$ delete $H_6 - H_8$

   $t_3 : x_3 = 10 \rightarrow H_9$ predicts 0, H_{10} predicts $1 \rightarrow \hat{y}_3 = 0 \neq y_3 = 1 \rightarrow$ delete H_9

   And in t_4 we only have H_{10} (the target function), having committed 3 mistakes ($\approx \log_2(10)$)

5. Prove Theorem 3.

$$E\left[\sum_{t=1}^{n} l(x_{\mathrm{t}}, y_{\mathrm{t}}, \hat{y}_{\mathrm{t}})\right] \leq 1 + \ln |H| \tag{7}$$

If we assume that at every time step one and only one hypothesis from the hypothesis class will be wrong, the expected value of the probability that we will choose this wrong hypothesis in every time step (and of course remove it from the hypothesis class once we observe the true label) is:

$$E\left[\sum_{t=1}^{n} l(x_{\mathrm{t}}, y_{\mathrm{t}}, \hat{y}_{\mathrm{t}})\right] = \sum_{k=2}^{|H|} \frac{1}{k} \tag{8}$$

As the probability is uniform, the probability of drawing a certain hypothesis from a set of size N is $\frac{1}{N}$, and since we are making sure that at every time step we make it, this size is getting reduced at every step. Also, as only one hypothesis is wrong at every step, the cost is 1. Finally, the sum starts from 2 as when the size is 1 it is impossible to make a mistake, as the remaining hypothesis is the target function (assuming of course that the labels and the observations are consistent).