1. ***Generalization bounds: theory and experiment***

[created by Fernando Valladares Monteiro and Keith Jenkins]

You are given a dataset D_{full} with 10 features and 10^6 samples. We want to use a simple perceptron to test how the generalization bound, ε varies in practice as a function of the number of samples N_{in}, the tolerance level δ, and the VC dimension d_{vc}. Whenever you vary one of these parameters, the others should be kept constant.

Hint: Remember that the VC dimension of the perceptron equals the number of features plus one, i.e., $d_{vc} = d + 1$.

a) To study the dependence on the tolerance level use $N_{in} = 1000$, $d_{vc} = 4$, $\delta = 0.01, 0.02, \ldots 0.5$ and follows the steps:

   i) Use the parameters above to compute theoretical generalization bound values as a function of δ.

   ii) Extract a dataset D_{out} with $N_{out} = 10^5$ samples and the first d (note that $d \neq d_{vc}$) features from D_{full}. You will use this dataset to estimate E_{out}.

   iii) Extract a dataset D_{in} with N_{in} samples and the first d features from D_{full}, train a perceptron on D_{in}, and compute E_{in} and E_{out}

   iv) Repeat steps (iii) **1000** times, remembering to save all your results in arrays.

   v) For each $\delta \in \{0.01, 0.02, \ldots 0.5\}$ in the values above, determine the maximum value of $|E_{out} - E_{in}|$ such that, over the 1000 runs, a proportion δ of the samples are above that value. Hint: for $\delta = 0.5$, this value is the median. This provides an estimate of the generalization bound as a function of the tolerance level.

   vi) Plot the results from (i) and from (v) side by side.

   Would you say the theoretical bound is tight, moderate, or loose in this experiment? (Let's define tight as within a factor of 2; moderate as a factor of 2-10; and loose as a factor of more than 10.) Would you say the relationship between the bound and the tolerance level is the same in theory and in the experiment, i.e., do the plots have similar shapes? If not, conjecture why if you can.

b) To study the dependence on VC dimension, you will vary the perceptron's complexity by selecting only the **first** d features. Use: $N_{in} = 1000$, $d_{vc} = [2 \ldots 11]$, $\delta = 0.1$ and follow the steps:

   i) Use the parameters above to compute theoretical generalization bound values as a function of d_{vc}

   ii) Extract a dataset $D_{out}^{(10)}$ with $N_{out} = 10^5$ samples and all the 10 features from D_{full}. You will use this dataset to estimate E_{out}.

iii) Extract a dataset $D_{in}^{(10)}$ with $N_{in} = 1000$ and all the features from the D_{full}.

iv) For each $d_{vc} = [2 \dots 11]$: create datasets $D_{out}^{(d)}$ and $D_{in}^{(d)}$ by extracting the first d (note that $d \neq d_{vc}$) features of $D_{out}^{(10)}$ and $D_{in}^{(10)}$ respectively, train a perceptron on $D_{in}^{(d)}$, and compute E_{in} and E_{out}.

v) Repeat steps (iii)-(iv) 100 times, remembering to save all your results in arrays.

vi) Determine the maximum value of $|E_{out} - E_{in}|$ such that, over the 100 runs, a proportion δ of the samples are above that value. This provides an estimate of the generalization bound as a function of the VC dimension.

vii) Plot the results from (i) and from (vi) side by side

Would you say the theoretical bound is tight, moderate, or loose in this experiment? (Let's define tight as within a factor of 2; moderate as a factor of 2-10; and loose as a factor of more than 10.) Would you say the relationship between the bound and the VC dimension is the same in theory and in the experiment, i.e., do the plots have similar shapes? If not, conjecture why if you can. (Hint: the model that the data was generated from is very simple.)

c) To study the dependence on the number of samples use: $N_{train} = [10,30,100,300,1000,3000,10000]$, $N_{test} = N_{train}/5$, $d_{vc} = 4$, $\delta = 0.1$ and follow the steps:

i) Use the parameters above to compute theoretical generalization bound values as a function of N_{train}.

ii) Use the parameters above to compute theoretical generalization bound values as a function of N_{test}. Remember that the generalization bound for train and test sets have different formulas.

iii) Extract a dataset D_{out} with $N_{out} = 10^5$ samples and the first d (note that $d \neq d_{vc}$) features fromD_{full}. You will use this dataset to estimate E_{out}

iv) For each N_{train}: extract a dataset D_{train} with N_{train} samples and d features from D_{full}, extract a dataset D_{test} with $N_{train}/5$ samples and d features from D_{full}, train a perceptron on D_{train}, and compute E_{train}, E_{test} and E_{out}.

v) Repeat step (iv) 100 times, remembering to save all your results in arrays.

vi) Determine the maximum value of $|E_{out} - E_{train}|$ such that, over the 100 runs, a proportion δ of the samples are above that value. This provides an estimate of the generalization bound as a function of the number of training samples.

vii) Repeat item (vi) for $|E_{out} - E_{test}|$.

viii) Plot the results from (i) and from (vi) side by side.

ix) Plot the results from (ii) and from (vii) side by side.