Homework number 4.

Theory question:
Show that if the concept classes C_i, $1 \leq i \leq d$ have VC dimension at most d, then their union $\mathcal{C} = \cup_i C_i$ has VC dimension at most $(2 + o(1))d$. For each d, show two concept classes C_1 and C_2 whose VC dimension is d, and whose union has VC dimension $2d$.

Programming assignment:
Let H_i be a class of concepts over $[0, 1]$. Each concept in H_i divides $[0, 1]$ to $i + 1$ equal length intervals, i.e. $(j/(i+1), (j+1)/(i+1)]$. For each interval the concept assigns a value of either 0 or 1. (Therefore there are 2^{i+1} concepts.) (The class H_0 has two concepts, the zero function and the one function.) Write a program to test model selection on this set of concepts.

For the input chose 1000 points uniformly distributed in $[0, 1]$. Classify them using:

1. Using a random hypothesis from H_{i*100}, for $0 \leq i \leq 10$, with noise 1/3. (I.e. each point with probability 2/3 keep its classification and probability 1/3 is flip it.)

2. Using a random hypothesis from H_{1000} with NO noise.

Each time check what is the *true* error of the hypothesis you generated (using SRM and CV) and compare it to the best possible error achievable.

Answer the following questions:

1. Compare the two criteria.

2. What happens when the number of intervals increase.

The homework is due in two weeks