max planck institut
informatik

Machine Learning SS 2007

Bast, Celikik, Chang, Funke, Giesen

http://www.mpi-inf.mpg.de/~bast/ml-ss07

Exercise Sheet 9

complete until Thursday, July 12th

In class, we studied the parameter δ in the PAC learning model and sketched a solution to the following problem:

Exercise 1 Suppose Algorithm A learns a concept class $\mathcal{C}$, such that for all distributions $\mathcal{D}$ (from which the training set is drawn), for all $\epsilon > 0$ and for some **fixed** $\eta > 0$, with probability at least η the hypothesis output by A has error at most ϵ. Define $N(\epsilon)$ to be the size of the training set needed by A.

Show that A can be used as a black box to find an algorithm A' such that, for any $\delta > 0$, with probability at least $1 - \delta$, the error of the resulting hypothesis is at most ϵ.

Rigorously prove that your scheme works and carefully calculate the number of samples needed for your new algorithm.

The following theorem will be helpful for proving that the "test sets" work:

Theorem (Chernoff Bound) Let $X_1, \ldots, X_n$ be independent, identically distributed random variables such that $X_i = 1$ with probability p and $X_i = 0$ with probability $1 - p$ for some $0 \leq p \leq 1$. For any choice of error $\alpha \in (0, 1)$:

$$\mathbf{Pr}\left[\left|\sum \frac{X_i}{n} - p\right| > \alpha\right] \leq 2e^{-2\alpha^2 n}.$$

Exercise 2 If you didn't finish the previous exercise sheet, try to complete it.