Big Data
Solutions for Exercise Class 7

Exercise 1: The VC generalization bound

Starting from $\delta = 4\tau_{\mathcal{H}}(2m)e^{-\frac{1}{8}\epsilon^2 m}$, we solve for ϵ

$$\begin{aligned}
\delta &= 4\tau_{\mathcal{H}}(2m)e^{-\frac{1}{8}\epsilon^2 m} && \text{(divide by } 4\tau_{\mathcal{H}}(2m)\text{)} \\
\frac{\delta}{4\tau_{\mathcal{H}}(2m)} &= e^{-\frac{1}{8}\epsilon^2 m} && \text{(take the natural logarithm)} \\
\ln\left(\frac{\delta}{4\tau_{\mathcal{H}}(2m)}\right) &= -\frac{1}{8}\epsilon^2 m && \text{(multiply by } -1\text{)} \\
\ln\left(\frac{4\tau_{\mathcal{H}}(2m)}{\delta}\right) &= \frac{1}{8}\epsilon^2 m && \text{(multiply by 8)} \\
8\ln\left(\frac{4\tau_{\mathcal{H}}(2m)}{\delta}\right) &= \epsilon^2 m && \text{(divide by } m\text{)} \\
\frac{8}{m}\ln\left(\frac{4\tau_{\mathcal{H}}(2m)}{\delta}\right) &= \epsilon^2 && \text{(take square root)} \\
\sqrt{\frac{8}{m}\ln\left(\frac{4\tau_{\mathcal{H}}(2m)}{\delta}\right)} &= \epsilon
\end{aligned}$$

Hence, we have:

$$L_{\mathcal{D}}(h_D) \leq L_D(h_D) + \sqrt{\frac{8}{m}\ln\left(\frac{4\tau_{\mathcal{H}}(2m)}{\delta}\right)}$$

For hypothesis classes with finite VC-dimension d, $\tau_{\mathcal{H}}(2m)$ is bounded by $(2m)^d$ (see exercise 1). Substituting this bound in the inequality above, we obtain

$$L_{\mathcal{D}}(h_D) \leq L_D(h_D) + \sqrt{\frac{8}{m}\ln\left(\frac{4(2m)^d}{\delta}\right)}$$

Consider the term

$$\sqrt{\frac{8}{m}\ln\left(\frac{4(2m)^d}{\delta}\right)}$$

How does it grow with m? The term

$$\ln\left(\frac{4(2m)^d}{\delta}\right) = \ln\left(\frac{4}{\delta}\right) + \ln\left((2m)^d\right) = \ln\left(\frac{4}{\delta}\right) + d\ln(2m)$$

grows only logarithmically with m. Hence, division by m will drive the whole expression to zero. This means $L_D(h_D)$ gets arbitrarily close to $L_{\mathcal{D}}(h_D)$, as $m \to \infty$, for hypothesis classes with finite VC-dimension.

Exercise 2: d-index and d-bound

(a) The d-index is 3: for example, {ABE,ABD,ABC} are 3 itemsets of size 3 that form an anti-chain (no itemset is a subset of any other itemset). The d-bound is also 3, since there is only 1 itemset of size bigger than 3.

(b) Now there still are 3 itemsets of size at least 3, namely {ABE,ABD,ABCE}, but they no longer form an anti-chain with respect to set inclusion, since ABE $\subset$ ABCE. The d-index drops to 2, and the d-bound remains 3.

(c) Both the d-index and the d-bound stay the same, because their definitions require the transactions to be *different*.

(d) According to the slides of lecture 10, and LEMMA 5.1 of the journal article of Riondato and Upfal:

$$|S| = \min\left\{|D|, \frac{4c}{\epsilon^2}\left(d + \ln\frac{1}{\delta}\right)\right\}$$

Apparently, experience shows $c \leq \frac{1}{2}$, so let's take $c = \frac{1}{2}$:

$$\frac{2}{0.05^2}\left(80 + \ln\frac{1}{0.05}\right) = 66,397$$

So a sample of size 66,397 suffices.

(e) For $\epsilon = 0.01$, we get:

$$\frac{2}{0.01^2}\left(80 + \ln\frac{1}{0.05}\right) = 1,659,915$$

We have $|D| = 1000,000$, so we'll just have to use the whole database!

Exercise 3: An even looser upper bound

(a) Define the e-bound as the largest integer e such that D contains at least e transactions of size (length) at least e. This number is at least as big as the d-bound, since we only dropped a condition: the transactions no longer have to be different. This bound

may seem silly, but it evaluated to the same number as the d-bound on the data sets "accidents","BMS-POS","kosarak" and "retail" as used by Riondato and Upfal (see table II in their journal paper). I didn't check on the other data bases. Of course, the e-bound is not immune to copying of transactions.

(b) This is actually the "retail" data set as used by Riondato and Upfal. Take the cumulative sum of "# transactions", starting at the last entry of the table working backwards, and return the size corresponding to the first entry for which this cumulative sum is bigger than the corresponding size. This is 58.

Exercise 4: Some intuition about the VC-dimension

I would expect the data base with independent items to have a higher VC-dimension than a data base with positively correlated items. Loosely speaking, if the items are independent there will be relatively little overlap between the "bad events". On the other hand, if there are (strong) positive correlations between the items, the overlap between "bad events" tends to be relatively large. To take an extreme case of positive correlation, suppose that always if someone buys item A, this person also buys item B. In this case the closure of A contains B, and the item sets $\{A\}$ and $\{A, B\}$ are supported by the same set of transactions.