Big Data
Exercises: sampling for frequent itemset mining

Exercise 1

(a) If the true relative support of an itemset Z is $\pi_Z = 0.1$, then what is the probability that a random sample of size $n = 100$ transactions will have $\hat{\pi}_Z > 0.2$? Here $\hat{\pi}_Z = \frac{m}{n}$, where m is the number of transactions in the sample that contain Z. (Hint: use the binomial distribution. In `R`, use `pbinom` or `dbinom` to compute binomial probabilities).

(b) Use the Hoeffding inequality to bound the probability that $\hat{\pi}_Z > 0.2$ in the situation described under (a). Hint: since we only want to bound the error in one direction, use only half of the two-sided bound. Compare the bound to the probability you computed under (a).

(c) You may have found that the bound provided by the Hoeffding inequality is not very "tight" when compared to the exact probability provided by the binomial distribution. But when we want to bound the probability of error in estimating the true support of an itemset in the database, using the Hoeffding inequality has a distinct advantage over using the binomial distribution. What is that advantage?

Exercise 2

On slide 6 of lecture 4 it is stated that if we want

$$\mathbb{P}(|\pi - \hat{\pi}| > \epsilon) < \delta,$$

then, using the Hoeffding inequality, we should choose n such that

$$\delta \geq 2e^{-2\epsilon^2 n}.$$

Verify that it follows we should choose

$$n \geq \frac{1}{2\epsilon^2} \ln \frac{2}{\delta}.$$

Hint: recall some basic properties of logarithms (not all of them necessarily required!)

1. $\ln(x \times y) = \ln x + \ln y$.
2. $\ln \frac{x}{y} = \ln x - \ln y$.
3. $\ln e^x = x$.
4. $\ln x^y = y \ln x$.
5. The logarithmic function is monotone increasing.

Exercise 3

Consider a frequent itemset mining problem with k items (the number of different products in the collection of the supermarket is k). Suppose we want to guarantee with high probability ($\geq 1 - \delta$) that for *all* itemsets Z

$$|\pi_Z - \hat{\pi}_Z| \leq \epsilon.$$

That is, we want the estimated relative support to be close to the true relative support for all itemsets. Formally stated, the requirement is:

$$\mathbb{P}\left(\bigcup_{Z \subseteq \{1,\ldots,k\}} |\pi_Z - \hat{\pi}_Z| > \epsilon \right) \leq \delta$$

(a) Use the Hoeffding bound in combination with the union bound (slide 35 of lecture 1) to determine the required sample size n, in terms of ϵ, δ and k.

(b) Referring to the table on slize 7 of lecture 4, what would be the required sample size for $\epsilon = 0.01$, $\delta = 0.01$, and $k = 50$?

(c) The union bound assumes that the events $|\pi_Z - \hat{\pi}_Z| > \epsilon$ for different itemsets Z are mutually exclusive. This is good if you want to give an absolute guarantee that the bound holds. But is this assumption realistic?

(d) Suppose that the events $|\pi_Z - \hat{\pi}_Z| > \epsilon$ for different itemsets Z are *independent* of each other. Derive a bound on

$$\mathbb{P}\left(\bigcup_{Z \subseteq \{1,\ldots,k\}} |\pi_Z - \hat{\pi}_Z| > \epsilon \right)$$

for this case. Is the bound much better than the one based on the union bound? (plot the two bounds for $\epsilon = 0.01$, $k = 5$, and n in the range from $15,000$ to $45,000$.)

Exercise 4: Lowering the threshold

Suppose that X is a frequent itemset at minimum (relative) support threshold $t = 0.01$, that is, $\pi_X \geq 0.01$. If the sample size is $n = 100,000$, and we want the probability that we miss X on the sample to be at most one in a thousand, what should be the minimum support threshold t' that we use to mine on the sample? (round to 3 decimal places). Hint: see theorem 3 in the paper by Toivonen, and the slides of lecture 4.

Instead, suppose we use the normal approximation of the binomial distribution, and a pessimistic estimate of the variance. What value of t' do we obtain in that case?