Mining massive Datasets WS 2017/18

Problem Set 1

Rudolph, Marvin, Daniela Schacherer

October 26, 2017

Exercise 01

Given is a cluster of n machines, each having a probability p of failing.

a) The probability of one machine to not fail is $1 - p$.
The probability of ALL machines not failing is n times $1 - p$ which is $(1 - p)^n$.
The probability of at least one machine failing is the opposite event and thus $1 - (1 - p)^n$.

b) The probability p_k of exactly k machines failing can be described using the binomial distribution. The binomial distribution describes the discrete probabilities of the number of successes in a sequence of independent experiments. As we have independent machines in the cluster with the number of successes corresponding to a machine failing we can write:

$$p(k|p,n) = \binom{n}{k} p^k (1 - p)^{n-k}$$

p^k is the probablity that k machines fail which has to be multiplied to the probability that the other $n - k$ machines do not fail. The binomial coefficient is the combinatoric element and describes in which way k elements can be chosen from n elements.

c) Zz.: $p_1 + p_2 + ... + p_n = 1 - (1 - p)^n$
We have $p_1 = p_2 = ... = p_n = p = \binom{n}{k} p^k (1 - p)^{n-k}$

$$p_1 + p_2 + ... + p_n = \sum_{k=1}^{n} \binom{n}{k} p^k (1 - p)^{n-k}$$

We can use the binomial theorem: $\sum_{k=0}^{n} \binom{n}{k} y^k x^{n-k} = (x+y)^n$ but have to subtract p_0 again

$$
\begin{aligned}
&= \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} - \binom{n}{0} p^0 (1-p)^n \\
&= ((1-p)+p)^n - (1-p)^n \\
&= 1^n - (1-p)^n \\
&= 1 - (1-p)^n
\end{aligned}
$$