Heuristics & Assessing results

TDT4187 – Exercise 2 September 26, 2017

Heuristics & expected running times

*1) Consider the Exact pattern matching problem introduced in the lecture. Consider the brute force algorithm using branch-and-cut. That is, over all positions of the text, the algorithm tries to match the pattern until the first mismatch is encountered, at which point it starts from scratch at a new position.

What is the *expected* running time of this brute force algorithm for:

(a) Text and Pattern symbols distributed uniformly at random.

(b) Text distributed uniformly, but each symbol of the Pattern is independently distributed according to the discrete distribution specified by: $p_A = 0.2$, $p_C = 0.4$, $p_T = 0.3$, $p_G = 0.1$.

(c) Both Text and Pattern distributed unevenly.

2) Solve the k-difference global alignment problem, as well as standard global alignment for Edit Distance scoring function δ, sequences $\omega_1 = GGCTCTA$ and $\omega_2 = CTCTAGC$, and k=2.

- Do you get the correct solution?
- When are you guaranteed that the algorithm's solution is the actual solution for the ED problem?

antonink@idi.ntnu.no

*3) Suppose that you know that GA-score distribution follows the following probability distributions. Can you design an algorithm that computes Edit Distance GA, and uses *on average* asymptotically less than $\mathcal{O}(n \cdot m)$ time?

(a) Binomial distribution $\mathrm{Bi}(\lfloor log(n) \rfloor, 1/2)$.

(b) Probability distribution specified by probability mass function $p(k) = \frac{1}{2} p_{\mathrm{Bi}(\lfloor log(n) \rfloor, \frac{1}{2})}(k) + \frac{1}{2} p_{\mathrm{Geo}(\lambda)}(k - \lfloor log(n) \rfloor)$ [1].

where $p_{Bi(\lfloor log(n) \rfloor, \frac{1}{2})}$ ($p_{Geo(\lambda)}$) denotes the probability mass function of the respective Binomial (Geometric) distribution.

(c) Will your solution work for the following distributions?

- Uniform probability distribution over all possible GA values.
- Probability distribution specified by probability mass function $p(k) = \frac{1}{2} p_{Bi(C \cdot n, \frac{1}{2})}(k) + \frac{1}{2} p_{Geo(\lambda)}(k - \lfloor log(n) \rfloor)$, where $C \in (0, \frac{1}{2})$ is a constant.

[1] There is something fishy with p_{Geo} in this context, can you tell what it is? How would you redeem it?

p-values

4) You have found the LA-score for sequences ω_1, ω_2 to be 8. Based on your parameters for the LA problem, and composition of the sequences, you model the LA-score distribution for unrelated ω_1, ω_2 by the following distributions.

   What are the p-values for the LA alignment you have found?
   Would you consider the sequences to be homologous?

   a) Poisson distribution with probability mass function $f_S(s) = \frac{\lambda^s e^{-\lambda}}{s!}$, with parameter $\lambda = 4$.

   b) Probability distribution with cumulative distribution function $P(S \leq s) = e^{-100 \cdot e^{-s}}$.

   *c) Consider your LA-score was 30, and you model the LA values with Poisson distribution with $\lambda = 25$. Approximate the Poisson distribution with normal distribution[2], and determine the p-value.

   That is, use $X \sim \mathcal{N}(\lambda, \lambda)$ instead of $X \sim Po(\lambda)$. Use the table[3] on the next page with values of CDF N(0, 1). You thus need to transform the LA-value to standard score, also known as Z-score.

   Compare your results with and without the continuity correction[4], with the real value of Poisson CDF[5]. Try the same with LA-score of 20.

[2] This is appropriate given λ is sufficiently large
[3] In the table, read the first decimal place along the rows, second along the columns.
[4] http://www.statisticshowto.com/what-is-the-continuity-correction-factor/
[5] There are many online calculators, such as `http://stattrek.com/m/online-calculator/poisson.aspx`.

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9958	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

Classifiers & ROC-curves

5) You have developed a test to determine homologous sequences. Homologous sequences get score X with distribution $X \sim F_X$, whereas non-homologous get $X \sim G_X$. Here, F_X and G_X are the respective cumulative distribution functions.

   a) You created a small control set to see how your clasifier performs. Your positive cases got scores $\{4, 6, 8, 9, 10, 12\}$ and your negative cases got scores $\{1, 2, 3, 3, 5, 7\}$.
   Draw the ROC curve for these cases, and determine the confusion matrix for threshold of 5,5.

   b) Next, you run your algorithm on all the sequences in the database. Each sequence got a unique score, and the sequences with the highest scores were labelled, in decreasing-score order, as follows:
   (P, P, P, N, P, N, P, P, P, N, N, P, P, P, P, N, ...)
   Compute ROC_5 for these results. What is its interpretation?

   c) You want to share the sequences with your colleagues. There is a good distinction between the distributions, so you decide your priority is that the sequences you share include at least 90% of all homologous sequences.

   Express in terms of F_X and G_X how to choose the cut-off value.
   Can you interpret the task graphically on the ROC curve?

   d) Assume that, from experience, you know that only 1 in 1000 sequences are homologous. You have tested all the sequences that were available in the database, order of 10^6, so you can believe that your results will hardly deviate from the pattern. You want to make sure your colleagues use their time wisely looking at sequences that indeed are homologous, but at same time you want to have them look at as many as possible. Thereby, you want to provide them with a set of sequences, where you can expect that at most 5% of them not to be homologous.

   Express in terms of F_X and G_X how you would choose the cut-off value. Can you interpret the task graphically on the ROC curve?

*e) Your method seems to work nicely, and you decided to share it with your colleagues. You don't think they would know themselves what threshold to choose, so you decided to choose the threshold that correctly labels the highest portion of sequences, whether that correct label is positive or negative.

- How do you find the threshold, and what does it represent on the ROC curve? Consider both discrete and continuous case[6].
- How does the answer change if negatives and positives are not distributed evenly?

[6]Example of the discrete case is the ROC curve in a). In continuous case, the ROC curve is represented as a continuous curve, based on F_X and G_X.