Hypothesis testing and Data Analysis
(Assignment 3)

Project Management and Research Methodologies

Author Ladislav Šulák

Email er1281@edu.teicrete.gr

Institute Technical Educational Institute of Crete

Heraklion 01/2018

Exercise 1

(a) The design of the experimental setup for this particular problem will be as follows.

- The information about how much data the system will have on the input and how often new data are coming (25 images per second) is precisely known. It means that we can estimate the maximum computation time for our algorithm. We are dealing with real-time system, so our algorithm cannot slow the final system down. The maximum execution times of other modules are known, so now we have to ensure, that the execution time of our module will fit so the final system can work on 25 frames per second without any delay. In numbers, it is in average 1 frame per 50ms, the other modules need maximum 20ms together so our algorithm could have, in theory, maximum execution time of 30ms.
- We will work only with our module and not with the other modules, because we want to be exact and measure only our part. If we would measure everything together, we wouldn't know what exactly is the maximum execution time of our algorithm, because this information could be inaccurate - sometimes there could be a very small execution time of other modules.
- The frames from an original video will be different, which means that we can have a big variety in the input and it is necessary to perform our measurement on a bigger dataset. For the input, we would prepare frames which are already acquisited from camera and are pre-processed. The original video data lasts 20 minutes. The video captures 4 different scenes and each has 5 minutes duration. There is a day in 3 scenes and night in one scene. Each scene is taken from a different place so that we have a bigger variety in out dataset as well.
- The goal is also to estimate how good our algorithm is regarding detection. We would also create labels in the input data frames with known information about face(s) in them. Then the measurement itself would be performed on such input.

(b) We would measure the execution time and the correctness of our algorithm by constructing the confusion matrix and by calculating accuracy, precision and recall.

(c) Confounding variables can be ones, which correlates our algorithm's execution time and its correctness (and which we didn't count account for).

For example, some frames can contain more information and it could cost our algorithm more time to finish. More complicated image with more faces could take longer to compute and also the time for writing to memory now requires more number of operations (so also CPU time), in comparison to image almost without any visible objects and no faces.

(d) For mitigating the negative effects of confounding variables by following: if we would find out, that the complexity of the input frame influences the algorithm performance in a noticeable way, we would try to optimize the algorithm or some other parts of the module. In case that memory operations are very expensive, we would try to reduce the number of writes (for example to write less often, not every frame). Or we would maybe even try to lower down the number of frames per second, which is however a big intervention to the specifications and we would need to discuss it with the project manager and probably also with the other members of the team.

(i) sample mean: $\bar{x} = \frac{751}{36} = 20,86$

(ii) standard deviation: $\sigma = \sqrt{\frac{105,25}{36}} = \sqrt{2,92} = 1,71$

(iii) Null hypothesis: the execution time of our algorithm is always less or equal to 20ms.

(iv) Alternative hypothesis: the execution time of our algorithm is always less or equal to 25 (we can afford it since we have to process 25 frames per second and other modules need 20ms to perform their actions).

(v) 99% confidence level means that we are testing the null hypothesis with significance level $a = 0,01$.

- The calculation of test statistics: $z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{20,86 - 20}{\frac{1,71}{\sqrt{36}}} = \frac{0,86}{0,285} = 3,02$
- We will use standard normal table from the assignment which is one-tailed. From the table, we got 2,33 as our normal table value (99% of confidence). Since we have one-tailed test, it means that we are searching if our z value is in critical region (if yes, our null hypothesis will be rejected): $z < 2,33$.
- Since $z = 3,02$ and $z > 2,33$, our null hypothesis is retained.

(vi) The task was to implement a face detection module which co-operates with other modules with following specifications:

- The selected camera captures images at 25 frames per second.
- The video acquisition and pre-processing module has a maximum execution time of 10ms
- The face tracking and the decision making and alerting subsystems together have a maximum execution time of 10ms.
- Your subsystem (i.e. algorithm) reads data (i.e. video frames) and writes (the coordinates of the detected face, if any) to volatile memory.

According to this specification, 36 tests were done with following execution times (in ms):

- $[t1 \ldots t12] = [20.8, 21.0, 18.0, 19.0, 18.3, 19, 24.0, 22.3, 21.5, 22.8, 21.6, 20.5]$
- $[t13 \ldots t24] = [22.0, 22.0, 19.5, 19.0, 24.0, 18.0, 20.4, 20.8, 25, 18.9, 20.5, 21.3]$
- $[t25 \ldots t36] = [19.1, 22.3, 21.4, 21.5, 22.8, 21.6, 20.5, 19.0, 18.9, 22.0, 21.3, 20.4]$

From these values we calculated sample mean, standard deviation value and we stated the null and alternative hypothesis. Sample mean was 20,86 and standard deviation was 1,71. We estimated null hypothesis as to be true if and only if the z value is bigger than 2,33 since we are working with 99% of confidence.

Summary: The final system has limited maximum CPU time for the whole processing part done on each frame. We were able to implement face detection module which reads pre-processed data from a camera and it writes its output, in the form of face(s) coordinates, to the volatile memory. All specifications were met and system can work in real time since our solution can process data in a reasonable amount of time, which had been also proved by hypothesis testing with 99% level of confidence calculated from 36 experiments.

Exercise 2

1.
   - Algorithm 1:

     $TP = 145$

     $TN = 4725$

     $FP = 35$

     $FN = 95$

     $$Accuracy = \frac{TN+TP}{TN+FP+FN+TP} = \frac{4870}{5000} = 0,974$$

     $$Precision = \frac{TP}{TP+FP} = \frac{145}{180} = 0,806$$

     $$Recall = \frac{TP}{TP+FN} = \frac{145}{240} = 0,604$$

     $$F-score = \frac{2}{\frac{1}{Precision}+\frac{1}{Recall}} = \frac{2}{1,241+1,656} = 0,690$$

   - Algorithm 2:

     $TP = 150$

     $TN = 4710$

     $FP = 50$

     $FN = 90$

     $$Accuracy = \frac{TN+TP}{TN+FP+FN+TP} = \frac{4860}{5000} = 0,972$$

     $$Precision = \frac{TP}{TP+FP} = \frac{150}{200} = 0,75$$

     $$Recall = \frac{TP}{TP+FN} = \frac{150}{240} = 0,625$$

     $$F-score = \frac{2}{\frac{1}{Precision}+\frac{1}{Recall}} = \frac{2}{1,33+1,6} = 0,682$$

   - Algorithm 1 has better accuracy and F-score, so it outperforms algorithm 2.

2. If we would need to find as many references as possible and manually filter out the wrong ones, which is very often the case in systematic review, we would prefer algorithm 2 since it is only a less precise and less accurate, but it is able to find more articles than algorithm 1 since it has better recall.

   However, if our requirement would be that we want to have more accurate and precise results we would use algorithm 1 even to the fact that we would have less number of found articles. That would be, for example, in the case that there is extremely big number of articles and we have to filter out the results more strictly since we don't want to go manually through each of one if it is not necessary.