Homework 1

Professor Lydia Y.Chen

CS4215: - Quantitative Performance Evaluation for Computing systems

September 8, 2021

Exercise 1. (10 Points)
The following output was obtained from a learning algorithm that performed a 2-way ANOVA with the two factors being A and B.

1. Fill in the blanks in the ANOVA table. How many replicates of experiments were performed?
2. We want to know which factors (A, B, and AB) are significant. What statistics should we compare against the computed F statistics for factor A,B, and AB, respectively? Assuming the significant level is 0.05. You also need to find out values for those statistics.
3. Figure out the P-values
4. If you want to fit a regression model, which factors shall be included? And why?
5. What conclusion would you draw about this experiment?

2-way ANOVA: y versus, A, B					
Source	DF	SS	MS	F	P
A	-	80	40	-	-
B	-	100	33.33	-	-
Interaction (AB)	-	40	6.67	0.8	-
Error	-	100	-		
Total	23	320			

Exercise 2. (15 Points)
In this exercise we want to see the effect of two independent variables, one system parameter and one model parameter, on the training time of *Random Forest* (RF) on a UCI dataset *Bank Marketing*. The system parameter we want to examine is the number of cores and the model parameter is the number of estimators for RF. The code for this question will be provided and you need to modify the corresponding parameters to achieve the training time by running the experiments on the Google Cloud Platform. Report the training time based on the following levels in the table below, 1) number of cores: 1, 2, 8, and, 2) number of

Table 1: Training Time (seconds)

	Number of Cores		
Number of Estimators	1	2	8
100	-	-	-
	-	-	-
	-	-	-
200	-	-	-
	-	-	-
	-	-	-
400	-	-	-
	-	-	-
	-	-	-
500	-	-	-
	-	-	-
	-	-	-

estimators: 100, 200, 400, 500. Number of cores and estimators can be adjusted using the parameters `n_jobs` and `n_estimators` respectively. Repeat each experiment for three times.

Perform a two-way ANOVA analysis and indicate if the null hypothesis *"there's no difference between the training time for the different number of cores and different number of estimators"*, should be rejected or not. Report the details of your analysis.

Exercise 3. (10 points)
Analyze the 2^{4-1} design shown in the sign table below:

A	B	C	D	y
1	-1	-1	-1	100
1	1	-1	-1	120
-1	-1	1	-1	15
-1	1	1	-1	10
-1	-1	-1	1	40
-1	1	-1	1	20
1	-1	1	1	30
1	1	1	1	50

1. Quantify all main effects.
2. Quantify percentages of variation explained.
3. List all confoundings.
4. Can you propose a better design with the same number of experiments.
5. What is the resolution of the design?

Exercise 4. (12 Points)
For the interactive system in Figure 1, suppose that we are given the following information:

- mean user think time = 3 seconds
- expected service time at device i = 0.01 seconds
- utilization of device i = 0.4
- utilization of CPU = 0.7
- expected number of visits to device i per visit to CPU = 15
- expected number of jobs in the central subsystem (the cloud shape) = 20
- expected total time in system (including think time) per job = 50 seconds.

How many jobs are there in the queue portion of the CPU on average, $E[N_Q^{cpu}]$?

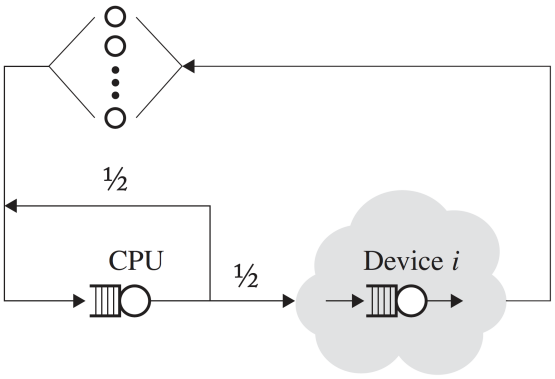

Figure 1: Figure for Exercise 4

Exercise 5. (15 Points)
Marty is running his database as a closed interactive system with N = 50 users. Each user submits a screenful of data to the database (her "job") to process, waits until she gets back an answer from the system, spends $E[Z] = 10$ seconds entering a new screenful of data (think time), and then submits that new job to the database. This process repeats ad infinitum. Marty realizes that his system's CPU utilization and his disk utilization are both high. He considers two modifications to his database to increase throughput. The first is to buy a second CPU (new CPUs on the market run at twice the speed of old ones) and divide the CPU load among the old CPU and the new one according to some optimal split. The second is to buy a second disk (new disks on the market run at three times the speed of old ones) and divide the disk load among the old disk and the new one according to some optimal split.
You obtain the following measurements of Marty's original system:

- C = 100 (number of jobs that completed during the observation period)
- $C_{CPU} = 200$ (number of completions at the CPU during observation)
- $C_{disk} = 500$ (number of completions at the disk during observation)
- $B_{CPU} = 800$ sec (time that the CPU was busy during observation)
- $B_{disk} = 1{,}500$ sec (time that the disk was busy during observation)

Your job is to answer two questions:

1. Assuming that the new disk and new CPU are equally priced, which should Marty buy to increase throughput?
2. Assuming that he chooses to buy the new disk (CPU), how should he optimally split requests between the old disk (CPU) and the new one? Work this out for whichever device you chose.

Exercise 6. (18 Points)
Consider the following Markov chains:

$$\boldsymbol{P}^{(1)} = \begin{pmatrix} 0 & 2/3 & 0 & 1/3 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 1/3 & 0 & 2/3 \\ 2/3 & 0 & 1/3 & 0 \end{pmatrix}$$

$$\boldsymbol{P}^{(2)} = \begin{pmatrix} 1/3 & 2/3 & 0 & 0 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 1/3 & 0 & 2/3 \\ 0 & 0 & 1/3 & 2/3 \end{pmatrix}$$

1. Draw the corresponding Markov chains for $\boldsymbol{P}^{(1)}$ and $\boldsymbol{P}^{(2)}$.
2. Solve for the time-average fraction of time spent in each state for both $\boldsymbol{P}^{(1)}$ and $\boldsymbol{P}^{(2)}$. First try to use the time-reversibility equations, and if they do not work, then use the balance equations.
3. For those chain(s) that were time-reversible, explain why it makes sense that for all states i, j in the chain, the rate of transitions from i to j should equal the rate of transitions from j to i.

Exercise 7. (20 points) This is a review exercise and meant to showcase you the potential project topics. To complete this exercise, you need to strictly follow the review template available on the brightspace Review exercise guideline. You need to first select ONE of the following four papers which talk about performance optimization issues of machine learning clusters, consisting of heterogeneous CPU and GPU. The details of this exercise and papers are provided here. Make sure you check them before you start the exercise.

1. Prediction-Based Power Oversubscription in Cloud Platforms. Paper
2. Habitat: A Runtime-Based Computational Performance Predictor for Deep Neural Network Training. Paper
3. INFaaS: Automated Model-less Inference Serving. Paper
4. Heterogeneity-Aware Cluster Scheduling Policies for Deep Learning Workloads. Paper
5. Elastic Parameter Server Load Distribution in Deep Learning Clusters. Paper

   InferLine: Latency-Aware Provisioning and Scaling for Prediction Serving Pipeline.Paper