Homework 2

Professor Lydia Y.Chen

CS4215: - Quantitative Performance Evaluation for Computing systems

October 7, 2021

Exercise 1. (10 Points)
A packet-switched Jackson network routes packets among two routers according to the routing probabilities shown in Figure 1. Notice that there are two points at which packets enter the network and two points at which they can depart.

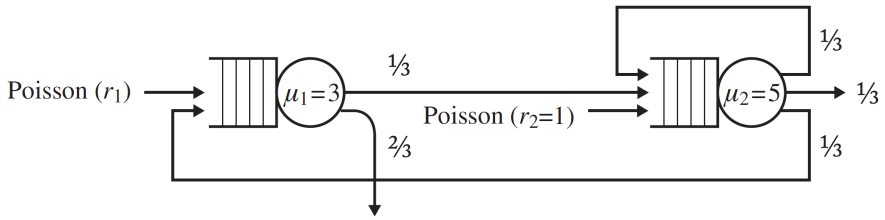

Figure 1: Figure for Exercise 1

1. What is the maximum allowable rate r_1 that the network can tolerate? Call this r_1^{max}.
2. Set $r_1 = 0.9r_1^{max}$. What is the mean response time for a packet entering at the router 1 queue?

Exercise 2. (15 Points)
In the *memory-usage.csv* file, the size of each job (MB) at the first column and the corresponding service speed at the second column are provided. The service speed denotes the amount of job which is processed in a second by the service node. In the third column, you can find the inter-arrival time for each job. The job size and inter-arrival time are derived from Pareto distribution and service speed is derived from exponential distribution. In the following, you can find the formulations of M/M/1 and G/G/1 systems, where the service rate is μ and arrival rate is λ. The mean time in the M/M/1 system and mean time in queue are:

$$E[T] = \frac{1}{\mu - \lambda}$$

$$E[T_Q] = \frac{\rho}{\mu - \lambda}$$

Also, an approximation for the mean waiting time in G/G/1 queue is:

$$E[W_Q] \approx (\frac{\rho}{1-\rho})(\frac{c_s^2 + c_a^2}{2})\tau$$

where τ is the mean service time (i.e. $\mu = \frac{1}{\tau}$ is the service rate), λ is the mean arrival rate, $\rho = \frac{\lambda}{\mu}$ is the utilization, c_a is the coefficient of variation for arrivals (that is the standard deviation of arrival times divided by the mean arrival time) and c_s is the coefficient of variation for service times.
Your job is to simulate an G/G/1 queue by considering the input file information. Also, base on input file and provided formulation, calculate the waiting time for M/M/1 and G/G/1 to compare the result with simulation.

Figure 2: Figure for Exercise 2

Exercise 3. (15 Points)
Your system consists of a single CPU with finite buffer capacity. Jobs arrive according to a Poisson process with rate λ jobs/sec. The job sizes are Exponentially distributed with mean $\frac{1}{\mu}$ seconds. Jobs are serviced in FCFS order. Let $N-1$ denote the maximum number of jobs that your system can hold in the queue. Thus, including the job serving, there are a maximum of N jobs in the system at any one time (this is called an M/M/1/N queue). If a job arrives when there are already N jobs in the system, then the arriving job is rejected.
Your DARPA proposal requires that you reduce the loss probability in your system. To do this you could either ask for money to double the buffer size, or, alternatively, you could ask for money to double the speed of the CPU so that jobs get processed faster, thereby lowering the probability that there are N jobs in the system. Assuming both proposals have the same cost, which do you choose? (Asking for both makes you seem greedy.)
These are the specific questions you should answer:

1. Draw the CTMC and derive the limiting probabilities, then derive a closed-form expression for $\boldsymbol{E}$[Number in system].

2. Determine a closed-form expression for $\boldsymbol{E}$[T] for only those jobs that enter the system.

3. Consider an M/M/1/2 with arrival rate λ, service rate μ, and finite capacity of 2. Derive $\boldsymbol{E}[T_{1,0}]$, the mean time to go from having one job in the system until the system is empty.

Exercise 4. (10 Points)
Consider an M/M/k system, where the service rate at each server is $\mu = 0.85$. Fix system utilization at $\rho = 0.50$. Now increase the number of servers, k, as follows - 1, 2, 4, 8, 16,

32, 64 - adjusting the arrival rate λ, accordingly. For each number of servers, derive (i) the fraction of customers that are delayed and (ii) the expected waiting time for those customers who are delayed. We are just looking for numerical answers here. Feel free to write a math program to evaluate the needed summations. Explain the trend that you see.

Exercise 5. (15 Points)
Bianca observes that her database throughput drops when she runs too many transactions concurrently (this is typically due to thrashing). She also observes that if she runs too few transactions concurrently, her database throughput drops as well (this is often due to insufficient parallelism). To capture these effects, Bianca models her time-sharing database system as an M/M/1/PS queue with load-dependent service rate, $\mu(n)$, where n denotes the number of concurrent transactions. The function $\mu(n)$ is shown in Figure 3.

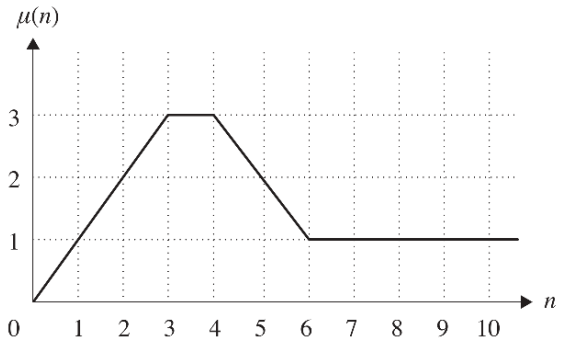

Figure 3: Figure for Exercise 5

1. Solve for the mean response time under Bianca's M/M/1/PS system. Assume arrival rate $\lambda = 0.95$. [Hint: Use a Markov chain.]

2. Bianca has a great idea: Rather than allow all transactions into the database as before, she decides to allow at most 4 transactions to run concurrently in the database, where all remaining transactions are held in a FCFS queue. Bianca's new queueing architecture is shown in Figure 4. Compute the mean response time for Bianca's new architecture, again assuming $\lambda = 0.95$, and Exponentially distributed service times with rates from Figure 4. What is the intuition behind Bianca's hybrid FCFS/PS architecture?

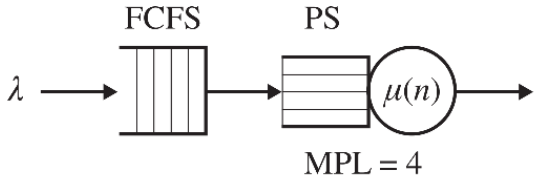

Figure 4: Processor-Sharing with limited multiprogramming level, MPL = 4.

Exercise 6. (35 Points)
Using the given dataset for Google Cluster Usage Traces, the purpose of this assignment is to build several classifiers to predict the event type that happens by the end of each task. This dataset consists of 663043 data instances with two event types of "successful", "unsuccessful". These event types are indicated as numerical values in the last column of each data row. Each data row consists of several features that are the values of resource usage records per task, e.g. memory usage and CPU usage. We use 33% of the data for test and the rest for training. We have provided you the code to read the data and balance it if necessary. The specific questions you need to answer are:

1. Find out the best hyper-parameter setting (list of hyper-parameters are specified for each classifier in the following) for each classifier resulting in the highest test accuracy. Explain your approach and observations in finding those hyper-parameters.

   - Logistic Regression → [solver, penalty, C]
   - Support Vector Machine (SVM) → [C, gamma, kernel]
   - Decision Tree → [criterion, splitter, max_depthint, min_samples_split]
   - Random Forest → [bootstrap, max_depth, max_features, min_samples_leaf, min_samples_split, n_estimators]
   - Multi-Layer Perceptron (MLP) → [hidden_layer_sizes, activation, solver, learning_rate]

2. What are the key attributes to predict the failed jobs? Investigate this by analysing `feature_importances_` in Random Forest.

3. (**Bonus**) For the classifier answered in the question one, choose maximum three hyper-parameters and perform an ANOVA analysis considering three levels per parameter. By this analysis you need to show if different levels of the parameters have significant effect on the test accuracy (consider $\alpha = 0.05$).

To answer these questions, the code is provided and you should fill in the blanks based on the following steps to build the classifiers and report the accuracy. In addition to answering the questions, attach your code to your submission. The following steps are required to pre-process the data and train the classifiers:

- Load the data for train and test and plot the class label population with pie plot. (For more analysis on parameter, if you want, you can concatenate the test and train data, shuffle them and split them randomly each time with the above mentioned sizes.)
- If the dataset is not balanced with regard to the class labels, try to resample the data to achieve a balanced dataset.
- Standardize the features.
- (**Bonus step**) Use Principal Component Analysis (PCA) to reduce the feature space dimension into k orthogonal components. Provide justification of your choice of k.

- Train the following classifiers and answer the three questions above.
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - Random Forest
  - Multi-Layer Perceptron (MLP)

You can find the data at the following link:
`https://drive.google.com/file/d/1e5mHEtJynC23txci8be8WLJKaMTgOUt4/view?usp=sharing`

Exercise 7. (7 Points - **Bonus**)
We define a threshold queue with parameter T as follows: When the number of jobs is $< T$, then jobs arrive according to a Poisson process with rate λ and their service time is Exponentially distributed with rate μ, where $\lambda > \mu$ (i.e., the queue is running in overload). However, when the number of jobs is $> T$, then jobs arrive with Exponential rate μ and are served with Exponential rate λ.

Figure 5 shows the CTMC for the case of $T = 2$. Compute $\boldsymbol{E}[N]$, the mean number of jobs in the system, as a function of T. As a check, evaluate your answer when $T = 0$. Note that when $T = 0$, we have $\rho = \mu/\lambda$.

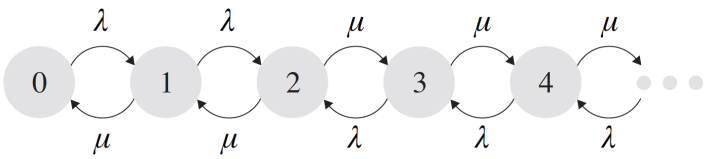

Figure 5: Threshold queue with $T = 2$.