Workload Characterization

1

Objective

- To observe the key characteristics of a workload, and develop a workload model that can be used to test multiple alternatives
  - Both analytical models and simulations require a workload model
- Example: modeling a web server
  - Inter-arrival process, service demands
    - Need information about distributions, not just summary statistics
  - Classes of requests

2

Wokload characterization techniques

- Select components and parameters
- Techniques: Averaging, Single Parameters Histogram, Fitting distribution to data, Multiple-parameters histogram, Principal Component analysis, Markov Models
- Clustering, Minimum Spanning tree
- Example of workload parameters: transaction types, instructions, packet sizes, page-reference pattern, source-destination of a packet

3

Workload Characterization

- Choose parameters that depend on the workload (e.g. type of requests) not system (elapsed time, CPU time)
- Example characteristics of service request
  - arrival time
  - type or request
  - duration of request
  - quantity of the resource demanded

4

Characterizing Data

- Previously: Averaging, histograms, multi-parameter histograms, fitting distributions to the data
- Fitting distribution to data: hypothesizing what family of distributions, e.g. Poisson, normal, is appropriate without worrying yet about the specific parameters for the distribution
  - Have to consider the shape of the distribution

5

Heuristics for hypothesizing a distribution

- Summary statistics can provide some information
  - Coefficient of variation (CV)
    - CV = 1 for exponential distribution, CV > 1 for hyper-exponential, CV < 1 for hypo-exponential, erlang
    - But CV not useful for all distributions, e.g., $N(0,\sigma^2)$
  - For discrete distributions, Lexis ratio $\tau = \sigma^2/\mu$ has the same role that CV does for continuous distributions
    - $\tau = 1$ for Poisson, $\tau < 1$ for binomial, $\tau > 1$ for negative binomial

6

Heuristics cont'd

- Histograms
  - Break up the data into k disjoint adjacent intervals of the same width and compute the proportion of data points that lie in each interval
    - Sturge's rule of thumb $k = \lfloor 1 + \log_2 n \rfloor$
      Given n data points
  - Visually compare the shape of the histogram to that of known distributions

7

Estimation of Parameters

- After hypothesizing a distribution, next step is to specify their parameters so that we can have a completely specified distribution
- Several techniques have been developed
  - Method of moments, Maximum likelihood estimators, Least-squares estimators

8

Method of moments

- Compute the first k moments of the sample data
- Equate the first few population moments with the corresponding sample moments to obtain as many equations as there are unknown parameters
  - Solve these equations simultaneously to obtain the required estimates
  - Example

9

Maximum Likelihood Estimation

- Suppose we have hypothesized a discrete distribution for our data that has one unknown parameter θ. Let $p_\theta(x)$ denote the pmf for this distribution. If we have observed the data X_1, X_2, …,X_n, we define the likelihood function $L(\theta)$ as follows: $L(\theta) = p_\theta(X_1) p_\theta(X_2) \ldots p_\theta(X_n)$

The MLE of θ is defined to be the value of θ that maximizes $L(\theta)$

For continuous distributions, $L(\theta)$ is defined analogously

10

MLE for exponential distribution

$$p(\beta) = 1/\beta e^{-x/\beta}$$

$$L(\beta) = (1/\beta e^{-X_1/\beta})(1/\beta e^{-X_2/\beta})....(1/\beta e^{-X_n/\beta})$$

$$= \beta^{-n} \exp(-\frac{1}{\beta}\sum_{i=1}^{n} X_i)$$

Taking logs on both sides, we have

$$\ln L(\beta) = -n \ln \beta - \frac{1}{\beta}\sum_{i=1}^{n} X_i$$

It can be shown through standard differential calculus by setting the derivative to 0 and solving for β that the value of β that maximizes L(β) is given by

$$\beta = (\sum_{i=1}^{n} X_i)/n = \overline{X}(n)$$

11

Determining how representative the fitted distributions are

- Both heuristic procedures and statistical techniques can be used for this
- Heuristics (Graphical/Visual techniques)
  - Density/Histogram Overplots and Frequency Comparisons
  - Q-Q plots
  - Probability plots (P-P plots)
  - Distribution Function Difference Plots

12

Statistical techniques

- Goodness-of-fit tests
  - Chi-square tests
  - Kolmogorov-Smirnov (KS) tests
  - Anderson-Darling (AD) tests
  - Poisson-process test

13

Chi-square tests

- First divide the entire range of the fitted distribution into k adjacent intervals
- Tally the number of data points in each interval o_i
- Compute the expected proportion of data points in each interval e_i
- Compute
  - D has a chi-square distribution with k-1 degrees of freedom
  - If the computed D less than $\chi^2(1-\alpha,k-1)$ then the observations come from the specified distribution
- Example

14

Chi-square tests cont'd

- Cell sizes should be chosen so that the expected probabilities e_i are all equal
- If the parameters of the hypothesized distribution are estimated from the sample then the degrees of freedom for the chi-square statistic should be reduced to k-r-1, where r is the number of estimated parameters
- For continuous distributions and for small sample sizes, the chi-square test is an approximation

15

Other tests

- Kolmogorov-Smirnov
  - Based on the observation that the difference between the observed CDF $F_o(x)$ and the expected CDF $F_e(x)$ should be small
- Anderson-Darling
  - More powerful in detecting differences in the tails of distributions

16

Fitting distributions to data in practice

- Use distribution-fitting software!
  - ExpertFit software from Averill Law
  - BestFit software
  - Download software and try it out on random data that you generate or data in exercises
    - See links on class web site

17

Principal Component Analysis

- Given set of workload parameters, determine set of factors
- Use weighted sum of parameters to classify the components

$$y = \sum_{i=1}^{n} a_i x_i$$

- PCA assigned weights to components such that they are maximally discriminative, weights are determined via PCA
- Principal components are ordered, first principal component explains highest variance

18

Example

- Number of packets send and received x_s, x_r by stations in local area network

Obs. No.	Variables x_s	Variables x_r	Normalized Variables x'_s	Normalized Variables x'_r	Principal Factors y_1	Principal Factors y_2
1	7718	7258	1.359	1.717	2.175	-0.253
2	6958	7232	0.922	1.698	1.853	-0.549
3	8551	7062	1.837	1.575	2.413	0.186
4	6924	6526	0.903	1.186	1.477	-0.200
5	6298	5251	0.543	0.262	0.570	0.199
6	6120	5158	0.441	0.195	0.450	0.174
7	6184	5051	0.478	0.117	0.421	0.255
8	6527	4850	0.675	-0.029	0.457	0.497
9	5081	4825	-0.156	-0.047	-0.143	-0.077
10	4216	4762	-0.652	-0.092	-0.527	-0.396
17	3644	3120	-0.981	-1.283	-1.601	0.213
18	2020	2946	-1.914	-1.409	-2.349	-0.357
$\sum x$	96336	88009	0.000	0.000	0.000	0.000
$\sum x^2$	567119488	462661024	17.000	17.000	32.565	1.435
Mean	5352.0	4889.4	0.000	0.000	0.000	0.000
Std. Dev.	1741.0	1379.5	1.000	1.000	1.384	0.290

19

Example

- Find correlation matrix
- Find eigenvalues of the correlation matrix
- Sort them in decreasing order
- Find corresponding eigenvectors

20

PCA example

- Compute correlation

$$R_{x_s,x_r} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_{si}-\bar{x}_s)(x_{ri}-\bar{x}_r)}{s_{x_s}s_{x_r}} = 0.916$$

- Create correlation matrix

$$\mathbf{C} = \begin{bmatrix} 1.000 & 0.916 \\ 0.916 & 1.000 \end{bmatrix}$$

- Covariance vs. correlations

$$\rho_{ij} = \frac{\text{cov}(X_i, X_j)}{\delta_i \delta_j}$$

21

$$\rho_{ij} = \frac{\text{cov}(X_i, X_j)}{\delta_i \delta_j}$$

PCA example

- Eigenvalues of the correlation matrix C
- Are 1.916 and 0.084
- Eigenvectors

$$\mathbf{q}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \quad \mathbf{q}_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$$

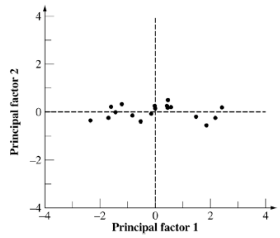
- Values of principal factors

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{x_s-5352}{1741} \\ \frac{x_r-4889}{1380} \end{bmatrix}$$

22

PCA Example

- First factor explains 95.7% of variations (sum of squares of principal factor 1 gives the percentage of variation explained by that factor)
- Second factor explains only 4.3% of the variations and can be ignored



- Can only use the first principal components to rank systems/servers
- Value of y_1 can rank systems into low, medium or high-load stations

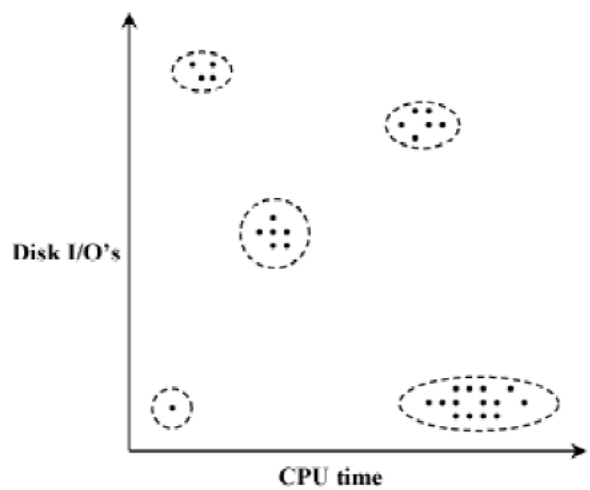
23

Clustering

- Many workloads consist of multiple classes of customers/requests
- Clustering is a technique used for classifying requests into multiple groups where members of one group are as "similar" as possible
  - Intragroup variance should be as small as possible whereas intergroup variance should be as large as possible
  - Non-hierarchical clustering: start with k clusters, move members around until intragroup variance is minimized
  - Hierarchical clustering: agglomerative and divisive

24

Example



25

Clustering

- Take a sample of workload data
- Select workload parameters
- Select distance measure
- Remove outliers
- Scale observations
- Perform Clustering
- Interpret results

26

Minimum spanning tree method

- Agglomerative hierarchical clustering technique
- Algorithm
  1. Start with k = n clusters
  2. Find the centroid of the ith cluster. The centroid has parameter values equal to the average of all points in the cluster
  3. Compute the intercluster distance matrix (distance between centroids)
  4. Find the smallest nonzero element of the distance matrix. Merge the two clusters with the smallest distance and any other clusters with the same distance
  5. Repeat steps 2 to 4 until all components are in the same cluster

27

Minimum spanning tree method cont'd

- Results of the clustering process can be represented as a spanning tree (a dendrogram) where each branch of the tree represents a cluster and is drawn at a height where the cluster merges with the neighboring cluster
- Given any maximum allowable intracluster distance, by drawing a horizontal line at the specified height we can obtain the desired clusters

28

Example

Consider a workload with five components and two parameters

Program	CPU time	Disk I/O
A	2	4
B	3	5
C	1	6
D	4	3
E	5	2

29

Example cont'd

First iteration:

	A	B	C	D	E
A	0	$2^{0.5}$	$5^{0.5}$	$5^{0.5}$	$13^{0.5}$
B		0	$5^{0.5}$	$5^{0.5}$	$13^{0.5}$
C			0	$18^{0.5}$	$32^{0.5}$
D				0	$2^{0.5}$
E					0

Minimum inter-cluster distance is between A and B, and D and E. The two pairs are merged

30

Example cont'd

- Second iteration:
  - Centroid of AB is {(2+3)/2),(4+5/2)}, I.e. {2.5,4.5}. Similarly for DE it is {4.5,2.5}

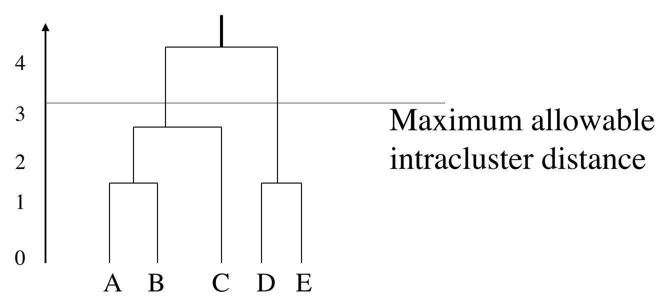
	AB	C	DE
AB	0	$4.5^{0.5}$	$8^{0.5}$
C		0	$24.5^{0.5}$
DE			0

  - Merge AB and C

31

Example cont'd

- Third iteration: merge ABC and DE to get a single cluster ABCDE



32

Additional Reading

- Articles on workload characterization by Calzorossa and Feitelson
  - On class web site
- More detailed discussion on clustering algorithms next week

33