Fitting distributions

CS 700

Objective

- Example workload characterization
- 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

Fitting Distributions to Data

- First step: 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

Heuristics for hypothesizing a distribution

- Summary statistics can provide some information
  - Coefficient of variation (CV) (standard deviation/mean)
    - 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 (two valued distribution),
  - Lexis ratio $\tau = s^2/\sigma^2$
  - (sample variance/mean) has the same role that CV does for continuous distributions
    - $\tau = 1$ for Poisson, $\tau < 1$ for binomial, $\tau > 1$ for negative binomial

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
  - Visually compare the shape of the histogram to that of known distributions

5

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

6

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

7

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 probability mass function for this distribution. If we have observed the data $X_1, X_2, \ldots, X_n$, 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 $L(\theta)$

8

Maximum Likelihood Estimation

- Construct likelihood function modeling probability of the data
- Take log of likelihood
- Take partial derivatives with respect to parameters
- Set to 0 and solve for parameters
- Example MLE of normal distribution
- Example MLE of binomial distribution

9

Bayes' Rule

- Bayes Rule for point probabilities

$$P(a \mid b) = \frac{P(b \mid a)P(a)}{P(b)}$$

- or in distribution form

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

- Useful for assessing diagnostic probability from causal probability:

P(Cause|Effect) = P(Effect|Cause) P(Cause)/P(Effect)
E.g., let *M* be meningitis, *S* be stiff neck:

$$P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)} = \frac{0.8 \times 0.001}{0.1} = 0.008$$

Bayes' Theorem applied to distributions

- Bayes Rule for distribution

$$P(\theta \mid data) = \frac{P(data \mid \theta)P(\theta)}{P(data)}$$

- We would like to use some preferences for certain values of parameters to estimate the posterior parameters
- Posterior ~ Likelihood x Prior
- Voting example black-board

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 $D = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$
  - 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