Measurement analysis basics - I

Lecture slides for S-38.3183

14.3.2007

Mika Ilvesmäki

Contents

- Basic concepts (events, traces)
- Data preprocessing, sampling
- Basic statistics
  - ranges, avearages, variations etc.
- Distributions
  - concepts, characteristics, parameterization
- Histograms

Mandatory reading

- Please, download from the course webpages "Chapter 2" of the 'hopefully someday to be published' -book
  - Chapter 2 contains a lot of information on statistics.
    - However, it is a draft and, therefore, full of typos, inconsistencies etc. Beware!
      - And if you find errors etc. please let the course personnel know of them! Thank you!
  - The material in Chapter 2 has to be mastered in the exercises (and in the final exam).
  - http://www.netlab.tkk.fi/opetus/s383183/k07/draft/chapter2-2007.pdf

Goals of this lecture

- After this lecture you should know
  - Basic concepts related to traffic measurements
    - Trace, events, sampling, mask, aggregation, disaggregation
  - What can be measured in a network
  - What is done in data preprocessing
  - Basic statistics and their meaning
  - How distributions/histograms are formed from measurements and how they can be interpreted/characterized

What is there to measure?

- The event itself
  - Count of packets
- The size or some other quantitative property of the event itself
  - Packet size, flow duration
- Inter-event relation
  - Frequency of events, the time between two events

Sampling

- With sampling one tries to form a picture of the whole by looking at a small(er) part
- Sample of sampling methods:
  - Sample packets with a fixed probability p and trace headers of sampled packets. This is the approach used by Cisco Netflow.
  - Independent Sampling: Sample every packet independently with a probability 1/p. Difficult to implement. Easy to analyze.
  - Periodic Sampling: Sample every 1/pth packet with probability 1. Easy to implement. Difficult to analyze.

When to sample

- If you have too much data
  - …to fit into memory/spreadsheet/given amount of processor cycles etc.
- If you are short of time…
- Or exploring rarely happening events
- Remember: Trace is a sample of the network
  - Rather than sampling the trace file, sample the network (obtain more traces)
- Big question: Is the sample (and the sampling method behind it) representative of the phenomenon you're trying to observe?

Measurement file: TRACE

- A file that has a set of measured properties (or events) from the network is called a **trace (or a network trace)**
- Trace has the following (relevant) property:
  - Length, inidicating the number of events (packets, flows or sessions etc.) or the time it covers
- Event entry consists of relevant event data
  - Packet nr, flow id., timestamp, addresses, ports, duration (flow), volume (flow) etc.
    - If some event data is not available, you might be able to create it (e.g., packet timestamps and 5-tuple info result in 5-tuple flows)
      - Though this might not be very straightforward

Data preprocessing I

- Data normalization
  - – Normalization is done to achieve comparability of data across two or more sets of measurements
  - – Normalization is also a way to reduce variation in measurements (normalizing to a range)
  - – Examples:
    - Min-Max method
    - Z-score method
    - Decimal scaling

HELSINKI UNIVERSITY OF TECHNOLOGY
Mika Ilvesmäki, D.Sc. (Tech.)

Data preprocessing II

- Data cleaning
  - – Caveat! Are you cleaning away the noise or a previously undetected phenomena
- Data integration
  - – Different sources, same concept but values expressed differently
  - – Careful measurement planning, coherent use of measurement equipment
- Data reduction
  - – Methods to reduce the dataset to smaller representations of the original data.

Masking & (Dis)Aggregating

- Regrouping packets by selected (parts of) header information -> obtaining sets of packets with common header value(s)
  - This new set is a network event that can be measured (size, nr of contained elements, etc.)
  - 5-tuple flow is one of the most common ones
- Iterative masking may be used to (dis)aggregate traffic further and provide reference points
  - Group by 5-tuple -> set of flows
  - Group flows by TCP/UDP Sport value -> set of flows originating from different Sports
  - OR group initially by TCP/UDP Sport value
    - What statistics remain the same?

Basic statistics – ranges and quantiles

- Statistical range indicates the range in which data lies
- Quantiles perform the division of data
  - Quartile -> four groups
  - Percentile -> 100 groups
  - Interquantile ranges (for instance, the range between 2nd and 3rd quartile).

Basic statistics – indications of average

- Arithmetic mean is one of the most common statistic
  - Uses all data available
    - Is affected by extreme values
  - instant, exponential
- Median
  - Is the middlemost value in an ordered set
  - Not affected by extreme values

Basic statistics – indications of variation

- Variance is a measure of absolute variation
  - Depends on values and scale of measurements
  - Standard deviation is the squre root of variance
- Coefficient of Variation (CofV) is used to compare variation between several sets of data
- Mean deviation
  - Descriptive statistic, mean deviation from the mean
    - Uses absolute values, analytical calculations are harder to perform
    - Good, ”intuitive” measure of variation

Higher-moment statistics - Skewness

- Used to describe histograms
- Skewness
  - 3rd moment statistic
  - Measure of asymmtery of a frequency distribution
  - Towards the tail, larger skewness, longer tail
    - Large positive, right sided tail
    - Large negative, left sided tail

Higher-moment statistics - Kurtosis

- Kurtosis
  - 4th moment statistic
  - Measure of combined weight of tails in relation to the rest of the distribution
  - Heavy tails, larger kurtosis value
    - Peaked distribution, Kurtosis >0, Leptokurtic
    - Flat distribution, Kurtosis <0, Platykurtic

Moving averages

- MAs are lagging indicators of trends in the dataset
  - If there are no trends, MAs are pretty useless
  - Otherwise MAs smooth the behavior and make it easy to follow trends
- Several types of MAs to choose from
  - Simple moving average (SMA)
  - Exponential moving average (EMA)
  - Smoothed moving average
  - Linear Weighted moving average

Arithmetic moving average

- Average value over a set number of observations
  - Determine
    - Window size (how many samples)
      - Longer windows produce more reliable results of trends but are not that sensitive to sudden changes
    - Move the start point as you get new samples
- Better to identify long-term trend changes

Exponential moving average

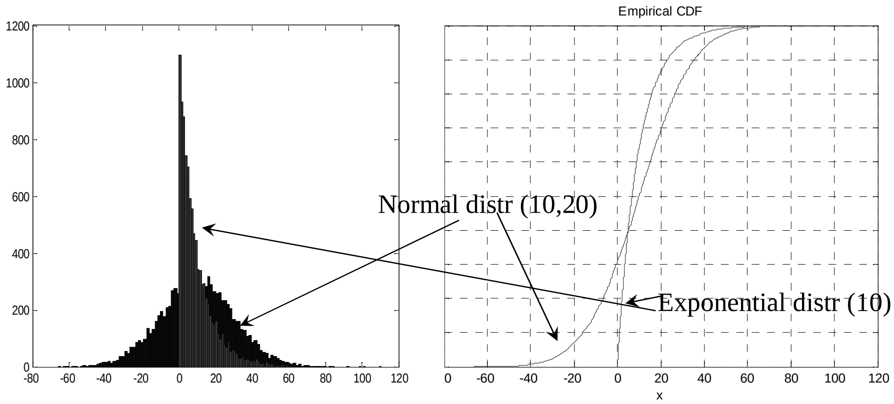
- Reduces the lag of AMA by applying more weight to recent values
- The weight is determined by the window size
  – The shorter the window, the more weight on recent values
- Very sensitive to quick changes

Probability distributions - concepts

- A probability distribution gives the frequency (probability) of possible events
  – Sample space: individual IATs
  – Events: intervals of IAT (0.01s-0.02s)
- In probability the distributions are completely described by distribution type and parameters
  – Inifinite number of independent random events -> normal distribution
  – Rare events -> Poisson distribution
  – Reference point to statistical distributions
  – Verification of assumptions

Distribution plots

- Probability Density/Distribution Function
  - Indicates how the density of the phenomena events(values) is distributed
- Cumulative Distribution Function
  - How much of data is below/over a certain threshold value



HELSINKI UNIVERSITY OF TECHNOLOGY
Mika Ilvesmäki, D.Sc. (Tech.)

Stable distribution parameters

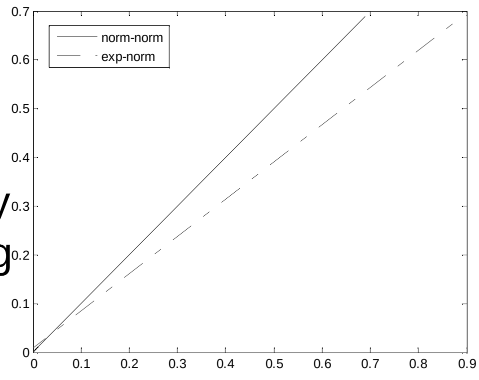
- Location
  - Either the midpoint (mean or median) or lower endpoint
- Spread
  - Variance, coefficient of variation. Determines the scale of the distribution
- Shape, general description or indicated by
  - Skewness that indicates the asymmetry around the mean
  - Kurtosis that indicates the peakedness/flatness or the weight in the tails of the distribution

Distribution analysis

- To find a closed form equation for an experimental distribution is very hard
- Procedure includes
  - Comparison of appearance
    - Overlaid histograms (use relative values)
      - Sensitive to bin choices
  - Comparison of parameters
  - P-P plots or Q-Q plots

Probability-Probability plots

- Plot the reference (known) distribution
  - Preferably solid line
- Plot the unknown distribution
  - Usually dots
- If you find similarity focus on explaining the differences



Histograms

- To summarize data, make a histogram of it.
  - Graphical display of tabulated frequencies
  - Categories are nonoverlapping intervals of a variable
  - Frequencies can be displayed in the histogram
    - As is
    - Divided by the total number of cases
      - Area under curve is 1
      - This is preferred for easier comparison with other distributions

Bins?

- Bin is the category or class of the variable
- When plotting a histogram one must decide the bin size / number of bins
  - Too many bins -> flat histogram
  - Too few bins -> high towers –histogram
  - Appropriate bin size improves the possibility of getting the actual distribution

Determining bin width

- Bin width as function of range and number of samples

$$W = \frac{R}{1+\log_2 n}$$

- Bin width as function of standard deviation and number of samples

$$W = 3.49sN^{-\frac{1}{3}}$$

- Bin width as function of inter-quartile range (IQR) and number of samples

$$W = 2IQR \cdot N^{-\frac{1}{3}}$$

Bin size and common sense

- Keep the bin width the same over traces if you intend to compare them
- Always experiment with other bin width values to get the best ”look & feel”.
- Bin size should be a multiple of data precision and the limits should be between possible readings
  - Bin 10<bin≤12 appeares to have center at 11
    - If data resolution is 1 then possible values within the bin are 11 and 12 (-> center at 11.5)
    - Choosing bin 10.5<bin≤12.5 has the center 11.5 and you do not have to worry about <, >,≤,≥ because no datapoint value will ever be 10.5 or 12.5.

Measurement analysis I summary

- Aim is to understand the nature of the measured phenomena with
  - Descriptive statistics
    - Means, measures of variation, range
    - Higher-order statistics
  - Distribution statistics
    - Histograms, bin sizes
    - Distribution comparison