Probabilistic Modelling: A Few Case Studies in Data Analysis and Performance Analysis

Jeremiah Deng

Information Science / Telecommunications Programme
University of Otago

April 8, 2015 @ Sun Yat-sen University, Guangzhou
Acknowledgement

- Dr. Yuwei Xu
- Dr. Suet-Peng Yong
- Other collaborators
- Other inspiring researchers
Outline

Introduction

Common Distributions

Clustering
   Case 1
   Case 2

Performance Analysis
   Queueing fundamentals
   Case 3
   Case 4
Outline

Introduction

Common Distributions

Clustering
  Case 1
  Case 2

Performance Analysis
  Queueing fundamentals
  Case 3
  Case 4
Background

- Probability concepts have been around for some time.
  - **Democritus**: *Everything existing in the universe is the fruit of chance.*
  - **Boethius**: *Chance, too, which seems to rush along with slack reins, is bridled and governed by law.*
  - **Caesar, Julius**: *Lacta alea est.* (The die is cast.)
  - **Einstein, Albert**: *I will never believe that God plays dice with the universe.*
Background

- Probability concepts have been around for some time.
  - **Democritus:** *Everything existing in the universe is the fruit of chance.*
  - **Boethius:** *Chance, too, which seems to rush along with slack reins, is bridled and governed by law.*
  - **Caesar, Julius:** *lacta alea est. (The die is cast.)*
  - **Einstein, Albert:** *I will never believe that God plays dice with the universe.*
  
- 宋人方岳：“不如意事常八九，可与人語無二三。”

- **Pioneers:** Leibniz, Pascal

- **Laplace, Pierre Simon:** *It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge.*
Probability in Information Sciences

- Artificial Intelligence
  - Probabilistic inference
  - Decisions under partial information
  - Processing signals (e.g., speech, images)

- Computer Networks
  - Channel scheduling
  - Packet collision
  - Queueing behaviour at routers

- Software Engineering
  - Model failure of safety-critical systems

- Data Compression
  - Shannon Theorem
  - Huffman coding
Outline

Introduction

Common Distributions

Clustering
  Case 1
  Case 2

Performance Analysis
  Queueing fundamentals
  Case 3
  Case 4
There are continuous and discrete probability distributions.

- Uniform distribution
- Normal distribution (aka Gaussian)
- Bernoulli distribution
- Binomial distribution
- Poisson distribution
- Exponential distribution
- Pareto distribution
Most prominent distribution in statistics.

Central limit theorem: under mild conditions the sum of a large number of random variables is distributed approximately normally.

pdf: \( p(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \)
Exponential Distribution

\[ p(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad F(x) = 1 - e^{-\lambda x} \]

\[ E(X) = \frac{1}{\lambda}, \quad \text{Var}(X) = \frac{1}{\lambda^2} \]

Models

- Life span of equipments, call duration, job processing time ...

⇒ Question: How likely does it last longer than average?
Poisson Distribution

- Models random occurrence of discrete events.
  - Service requests received per hour.
  - Number of packets arriving at a node per second.

\[
P(n) = \frac{e^{-\lambda} \lambda^n}{n!}, \quad n = 0, 1, 2, \ldots
\]

- \( E(n) = \lambda = \text{Var}(n) \)
Age Distribution of Populations

▶ Guess how would it look like?
Age Distribution of Populations

▶ Guess how would it look like?
Age Distribution of Populations

- Guess how would it look like?

- Discussion point: Will the age distribution of the populations affect the outcome of population-based optimization?
Outline

Introduction

Common Distributions

Clustering
  Case 1
  Case 2

Performance Analysis
  Queueing fundamentals
  Case 3
  Case 4
Clustering

- Clustering algorithms assume data distributions within the clusters.
- Gaussian E-M: assume Gaussian ellipsoids.
  - k-means is a special case of GEM.
- Mixture of Gaussian models can be explored for classification, and anomaly detection.
Case 1: Scene analysis and novelty detection

- S-P Yong et al., *Pattern Recog.*, 45(9), 2012.
- 14×14 label co-occurrence matrices as ‘feature code’
- Clusters reviewed by PCA
- Are they Gaussian?
1-D Gaussian modeling

- High-dimensional space tricky to model
- Resort to modelling point-to-centre distances 😊
- Q-Q plot seems okay
- $\chi^2$ goodness-of-fit test: $p$-value=0.085 (null hypothesis not rejected)
- A simple threshold used for outlier detection
The Quincunx

http://www.mathsisfun.com/data/quincunx.html
Case 2: Learning the $k$ in $k$-means

- Lacking prior knowledge, often we don’t know $k$.
- The idea: Start with a low number (e.g., 1), and examine the clusters.
- If a cluster passes Gaussian test, keep it; otherwise split.
- Questions:
  - How to test the normality?
  - How to split?
Case 2: Learning the $k$ in $k$-means

- Lacking prior knowledge, often we don’t know $k$.
- The idea: Start with a low number (e.g., 1), and examine the clusters.
- If a cluster passes Gaussian test, keep it; otherwise split.
- Questions:
  - *How* to test the normality?
  - *How* to split?
  - Can we kill two birds with one stone?
Case 2: Learning the $k$ in $k$-means

- Lacking prior knowledge, often we don’t know $k$.
- The idea: Start with a low number (e.g., 1), and examine the clusters.
- If a cluster passes Gaussian test, keep it; otherwise split.
- Questions:
  - *How* to test the normality?
  - *How* to split?
  - Can we kill two birds with one stone? 一箭双雕
Algorithm 1 G-means($X, \alpha$)

Input: $X$ - dataset, $\alpha$ - a confidence level
Output: $k$ cluster centres $C$
1: Initialize $C$ as a set of centres (usually $C \leftarrow \{\bar{x}\}$)
2: $C \leftarrow \text{kmeans}(C, X)$.
3: for all $c_j \in C$ do
4: Let $C^{(j)} = \{x_i|\text{class}(x_i) = j\}$ be the set of datapoints assigned to center $c_j$.
5: Use a statistical test to detect if each $C^{(j)}$ follows a Gaussian distribution (at confidence level $\alpha$).
6: If the data look Gaussian, keep $c_j$. Otherwise replace $c_j$ with two centres.
7: end for
8: Repeat from step 2 until no more centres are added.
How to test a cluster?

1. Initialize two centres $c_1$ and $c_2$; re-cluster.
2. Get $v = c_1 - c_2$, and project data vector $x_i$ onto $v$: $x'_i = \frac{\langle v, x_i \rangle}{\|v\|}$.
3. Normalize $X'$ to zero mean and variance 1.
4. Calculate empirical cumulative density function $z_i = F(x'_i)$, and the Anderson-Darling statistics $A_2(Z) = -\frac{1}{n} \sum_i (2i - 1)[\log(z_i) + \log(1 - z_{n+1-i})] - n$.
5. If statistics within range of the critical value, keep the original centre, and discard $\{c_1, c_2\}$; otherwise, replace the current centre with $\{c_1, c_2\}$. 

How to test a cluster?

- Initialize two centres $c_1$ and $c_2$; re-cluster.
- Get $v = c_1 - c_2$, and project data vector $x_i$ onto $v$: 
  $x'_i = \langle v, x_i \rangle / \| v \|$. Normalize $X'$ to zero mean and variance 1.
- Calculate empirical cumulative density function $z_i = F(x'_i)$, and the Anderson-Darling statistics 
  $A^2(Z) = -\frac{1}{n} \sum_i (2i - 1)[\log(z_i) + \log(1 - z_{n+1-i})] - n$
- If statistics within range of the critical value, keep the original centre, and discard $\{c_1, c_2\}$; otherwise, replace the current centre with $\{c_1, c_2\}$. 
Another Take on Initialization

- What if we want a one-shot clustering with $k$ clusters?
- Requires a better way to do initialization.

**Algorithm 2** \( k \text{-means+++}(X, k) \)

**Input:** \( X \) - dataset, \( k \): number of cluster centres  
**Output:** \( C = \{c_i\} \): a set of \( k \) initial centres

1. Take one centre \( c_1 \), chosen uniformly at random from \( X \).
2. Take a new centre \( c_i \), choosing \( x \in X \) with probability \( \frac{D(x)^2}{\sum_{x \in X} D(x)^2} \).
3. Repeat Step 2 until we have taken \( k \) centres altogether

\( D(x) \): the shortest distance from data point \( x \) to the closest centre.
Outline

Introduction

Common Distributions

Clustering
  Case 1
  Case 2

Performance Analysis
  Queueing fundamentals
  Case 3
  Case 4
Stochastic Processes

- Over a continuous time parameter, SP is defined as a collection of random variables.
  - Denoted as \( \{X_t\}, \ t \in R \).
- Over a discrete time parameter, is defined as a collection of random variables.
  - Denoted as \( \{X_n\}, \ b \in Z \).
- These random variables are related and defined in the same probability space.
- Stationary stochastic process: statistics of the process will not vary over time.
Point Process

- Point process (aka counting process), is a process with random occurrence of points on a line.
- Denoted as \( \{N(t), t \geq 0\} \).
  - Number of customers arriving in a shop during time of \([0, t)\).
- If \( s < t \), then \( N(s) \leq N(t) \).
- Increment \( N(s) - N(t) \): Number of event occurrence within \((s, t)\).
Poisson Process

- $N$ has stationary increments.
- $N$ has independent increments.
- Probability of 1 arrival in small interval $h$:
  - $P[N(h) = 1] = \lambda h + o(h)$.
- Probability of 2 or more arrivals in $h$:
  - $P[N(h) \geq 2] = o(h)$.
- Such a point process is a Poisson Process with a rate of $\mu > 0$. 
Graphical Representation

- Markov Chains are used to describe system state transition in a Poisson process.
- A point process counting arrivals only: always growing
- Birth-death process can be used for queueing modeling.
  - Right links represent birth or arrival;
  - Left links are for death or departure.
Kendall Notation \((A/B/c/K/m/Z)\)

- **A**: interarrival time distribution
- **B**: service time distribution
  - \(M\) for exponential
  - \(G\) for general
- **c**: number of servers
- **Optional:**
  - **K**: maximum number of allowed customers
  - **m**: size of the customer population
  - **Z**: queueing discipline, typically FIFO
M/M/1 Queuing

- Interarrival time: exponentially distributed, mean $= 1/\lambda$
- Job processing time: exponentially distributed, mean $= 1/\mu$
- Steady state requires $\lambda < \mu$
- FIFO
- One server
- Chances the server is busy: $P[N \geq 1] = 1 - p_0 = \rho$
- Expected number in system: $L = E\{N\} = \sum_n np_n = \frac{\rho}{1-\rho}$
A Little Variation to $M/M/1$

- In $M/M/1$, there is no limit on the queue length.
- In reality, services usually don’t support unlimited queueing (memory, ports etc.)
- If a customer finds no available position in a limited queue, it is supposed to disappear!
M/M/1/K Analysis

- Transition diagram

\[ \begin{array}{c}
0 \xrightarrow{\lambda} 1 \xrightarrow{\lambda} 2 \xrightarrow{\lambda} \cdots \xrightarrow{\lambda} k-1 \xrightarrow{\lambda} k \\
\mu & \mu & \mu & \mu & \mu
\end{array} \]

- Steady state solutions:

- \[ \sum_{n=0}^{K} p_n = 1 \]

- \[ p_n = \left( \frac{\lambda}{\mu} \right)^n p_0, \quad 0 \leq n \leq K \] (K was \( \infty \) in \( M/M/1 \)).
Expected Customer Numbers

▶ Solution to the probabilities:

\[ p_n = \rho^n (1 - \rho) / (1 - \rho^{K+1}). \]

▶ Expected customer number in system:

\[ L = E\{N\} = \sum_{n=0}^{K} np_n = \frac{\rho}{1 - \rho} - \frac{(K + 1)\rho^{K+1}}{1 - \rho^{K+1}} \]

▶ Probability that an arriving customer is rejected is (simply) \( p_K \).

▶ Rejection rate is therefore \( p_K \lambda \).

▶ **Actual** arrival rate into the system is

\[ \lambda' = (1 - p_K)\lambda. \]
Case 3: Video streaming on wireless multi-hop networks

- Each node subject to MAC contention, interference, load, limited buffer, and other overhead ...
- Treated as M/M/1/K to form a Jackson network
- What is the best routing metric: #hops, load, delay ...?
In reality, we know $\lambda$, but not $\mu$ ...

We do know $Q$ ($Q \leq K$), and $\lambda_d$

The question: how to estimate $\mu$, so that the processing time for each node can be derived.

Hang on - how do you know you are handling Poisson traffic?
Simulation Settings

- Network Simulator 2 (NS-2) + EvalVid in an IEEE 802.11b/g networks of grid topologies: 5 × 5, 7 × 7, 9 × 9, 15 × 15

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance between two neighbours</td>
<td>150m</td>
</tr>
<tr>
<td>Antenna Type</td>
<td>Omnidirectional</td>
</tr>
<tr>
<td>Standard</td>
<td>802.11b</td>
</tr>
<tr>
<td>Transmission Range</td>
<td>250m</td>
</tr>
<tr>
<td>Transmission Rate</td>
<td>11 Mbits/s</td>
</tr>
<tr>
<td>Packet Size</td>
<td>1024 bytes</td>
</tr>
<tr>
<td>Queue Size</td>
<td>50 packets per node</td>
</tr>
<tr>
<td>Video Format</td>
<td>H.264/MEPG4</td>
</tr>
<tr>
<td>Duration</td>
<td>29 ∼ 66s</td>
</tr>
<tr>
<td>Number of Streams</td>
<td>3 ∼ 7</td>
</tr>
<tr>
<td>Minimum number of hops</td>
<td>4</td>
</tr>
</tbody>
</table>
Poisson Inflow / Outflow

Histogram with Poisson PDF

number of incoming packets

Probability

Theoretical Quantiles
Sample Quantiles

number of outward packets

Probability

Theoretical Quantiles
Sample Quantiles
Chi-Square Test Results

For video ‘Highway’:

<table>
<thead>
<tr>
<th>Node</th>
<th>Inflow</th>
<th>Outflow</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DF</td>
<td>$\chi^2$</td>
</tr>
<tr>
<td>N30</td>
<td>6</td>
<td>12.1596</td>
</tr>
<tr>
<td>N36</td>
<td>5</td>
<td>7.3989</td>
</tr>
<tr>
<td>N31</td>
<td>6</td>
<td>11.1698</td>
</tr>
</tbody>
</table>

For video ‘Grandma’:

<table>
<thead>
<tr>
<th>Node</th>
<th>Inflow</th>
<th>Outflow</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DF</td>
<td>$\chi^2$</td>
</tr>
<tr>
<td>N34</td>
<td>5</td>
<td>5.2467</td>
</tr>
<tr>
<td>N41</td>
<td>5</td>
<td>4.3583</td>
</tr>
<tr>
<td>N49</td>
<td>5</td>
<td>1.2288</td>
</tr>
</tbody>
</table>

Our assumption on Poisson traffic holds on the significance level of $\alpha = 0.05$. 
Case 4: Mobility Modelling and Simulation in DTNs

- Delay Tolerant Networks: opportunistic forwarding, e.g., Vehicular ad hoc networks
- Simulators use simple motion models such as Random Walk and Random Waypoint etc.
- Real traces are available but few
- Question: how good are the motion models in simulations?
- Karagiannis et al. (MobiCom’07): Power law and exponential decay of inter contact times between mobile devices.
Dichotomy exists

- Good linear fitting up to day-level (power law)
- Then exponential delay follows
Simple motion patterns (e.g. Random Waypoint) generate similar effects
Elsewhere

- Dirichlet distributions widely used for topic modelling in text and multimedia content analysis
- Markov chain models for weather and renewable energy data modelling and simulation
- ... (almost) the entire Machine Learning field
- ... Your contributions 😊
La vita è bella

美麗人生 - 1997