Overview

- Me: Doug Leith
- Topics we'll cover
- Motivation
  - In computer science
  - Elsewhere
- Counting
  - Why lean to count ?
  - Sum rule of counting
  - Product rule of counting

Introduction

Module is roughly split into four parts:

1. <u>Random events</u>: counting, events, axioms of probability, Bayes, independence

2. <u>Random variables</u>: discrete RVs, mean and variance, correlation, conditional expectation
   Mid-term

3. <u>Inequalities and laws of large numbers</u>: Markov, Chebyshev bounds, sample mean, weak law of large numbers, central limit theorem, confidence intervals, bootstrapping

4. <u>Statistical models</u>: continuous random variables, logistic regression, least squares

Probability in Computer Science

- Machine learning, data analytics, "big data"

"Unsupervised learning" and the future of analytics
An interview with analyst Seth Grimes on the growing use of machine learning in business intelligence
The Economist

Up to 21,000 big data jobs in next 4 years

Monday, April 20, 2015

Irish Examiner

Demand for Computer Systems Analysts with big data expertise increased 89.9% in the last twelve months

Forbes / Tech

Probability in Computer Science

- Cloud applications, data centres, network design
  - Data centres e.g. load balancing, queues, job arrivals, capacity and reliability

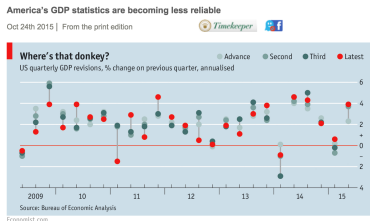Google is harnessing machine learning to cut data center energy

BIGQUERY
A fast, economical and fully managed data warehouse for large-scale data analytics.
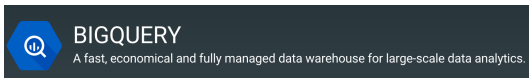
  - Networks e.g. WiFi random access, queues, traffic arrivals, information theory

WiFi

- Spam, passwords, hash tables etc

Probability in Real Life

- Insurance, financial markets, the economy

Balancing reliability and timeliness

- Health

More to meat and cancer link than what's in headlines

Analysis: Association of disease with processed and red meat must not be overstated

Mon, Oct 26, 2015, 21:03 | Updated: Tue, Oct 27, 2015, 10:00

THE IRISH TIMES

Logistics

- Web page: www.scss.tcd.ie/doug.leith/ST3009/
- Book: "A First Course In Probability", Sheldon Ross, Prentice-Hall (in library, google to find pdf).
- Lots of online resources e.g.
  - Khan Academy (www.khanacademy.org/math/statistics-probability)
  - Stanford (lagunita.stanford.edu/courses/OLI/ProbStat/Open/about)
  - MIT (ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/)
  - edX (www.edx.org/course/introduction-probability-science-mitx-6-041x-2)

  although most concentrate just on the first two parts of the module.
- Matlab (or octave) for this course. R or Python scikit.
- Will use Matlab in labs/tutorials and lectures – learning how to programme Matlab is <u>homework</u>.
- We'll be using some maths ...
- 20% of marks mid-term test (held right after study week).
- 10% of marks from weekly assignments (one for each of first 5 weeks, two weeks to complete each)

Why Learn to Count ?

Probability is "the likelihood of something happening" – OED

- Perform an experiment several times
- We need to be able to count the number of times the outcome we're interested in happens
- Divide this by number of experiments to get the frequency ("probability") with which it occurs

Aside: What about events that can only happen once ? We'll come back to this later.

Why Learn to Count[1] ?

<u>Example</u>:
In a family of two children, where the oldest child is a girl, what is the probability they are both girls?

---
[1]https://en.wikipedia.org/wiki/Boy_or_Girl_paradox

Why Learn to Count ?

- Toss a coin twice. What is the probability of two heads ?
  - Possible outcomes are TT, TH, HT, HH. So expect two heads seen 1 time in 4.
  - What about probability of one head and one tail ?
- Now toss coin 10 times ... what is probability of two heads now ?
- Instead of tossing coins, think of errors in packets transmitted over a wireless link
- Or hash table collisions
- Or chance of guessing a well chosen password

Sum Rule of Counting

- Suppose we have a set A of n outcomes and a set B of m outcomes
- Suppose no outcome in A is in B, and vice versa
- Suppose an experiment is performed by drawing one outcome from <u>either</u> A or B.
- For the experiments together there are $m + n$ possible outputs
- Just enumerate the possible outcomes of the experiment:

$$1, 2, \cdots, n, n+1, n+2, \cdots, n+m$$

- In set notation, number of outcomes is $|A| + |B|$

Sum Rule of Counting

Example:

- Suppose we have two sets of servers, one in Dublin and one in London, with 100 servers on Dublin and 50 servers in London.
- A web request arrives and must be routed to one of the servers. How large is the set of servers that it may get routed to ?
- Request can be sent to either of two locations, none of the machines in either location is the same. So sum rule applies and the request could potentially be routed to any of the 150 $(= 100 + 50)$ servers.

Product Rule of Counting

- Suppose two experiments are performed.
- Experiment 1 can result in any one of m possible outcomes.
- Experiment 2 can result in any one of n possible outcomes.
- For the two experiments together there are mn possible outputs
- Just enumerate the possible outcomes for the pair of experiments:

$$
\begin{array}{cccc}
(1,1) & (1,2) & \cdots & (1,n) \\
(2,1) & (2,2) & \cdots & (2,n) \\
\vdots & & & \\
(m,1) & (m,2) & \cdots & (m,n)
\end{array}
$$

Product Rule of Counting

Example:

- A byte consists of 8 bits. Each bit takes the value 0 or 1. How many different bytes are there ?

- The first bit takes one of 2 possible values. The second bit takes one of 2 possible values. And so on.

- So the total number of potential byte values is
  $2 \times 2 \times 2 \cdot 2 = 2^8 = 256$.

Product Rule of Counting

Example:

- Two 6-sided dice, with faces numbers 1 through 6, are rolled. How many possible outcomes of the roll are there ?
- The first die can come up with 6 possible values, and the second with 6 possible values (regardless of what appeared on the first die). So the total number of potential outcomes is 36 ($= 6 \times 6$).

Product Rule of Counting

Example:

- Consider a hash table with 100 buckets. Two arbitrary strings are independently hashed and added to the table. How many possible ways are there for the strings to be stored in the table ?

- Each string can be hashed to one of 100 buckets. Since the results of hashing the first string do no affect the second, there are $100 \times 100 = 10,000$ ways that the two strings may be stored in the hash table.

Generalised Product Rule of Counting

- Suppose r experiments are performed.
- Experiment 1 can result in any one of n_1 possible outcomes.
- Experiment 2 can result in any one of n_2 possible outcomes.
- and so on ...
- For the r experiments together there are $n_1 \times n_2 \times \cdots \times n_r$ possible outputs

Generalised Product Rule of Counting[2]

- Suppose I choose a password of length 4 using letters a-z.
- Number of possible passwords is $26 \times 26 \times 26 \times 26 = 456,976$.
- Now suppose I use letters a-z and A-Z
- Number of possible passwords is now
  $52 \times 52 \times 52 \times 52 = 7,311,616$.
- Letters a-z, A-Z and numbers 0-9
- Number of possible passwords is now
  $62 \times 62 \times 62 \times 62 = 14,776,336$.
- Suppose its of length 6. Number of passwords is $62^6 \approx 5 x 10^{10}$
- Of length 8. Number of passwords is $62^8 \approx 2 x 10^{14}$

---

[2]Pushing on string: the 'don't care' region of password strength,
https://dl.acm.org/citation.cfm?id=2934663

Generalised Sum Rule of Counting for 2 Sets

Some examples:

- Suppose we have a set A of n outcomes and a set B of m outcomes
- Some outcomes in A may also be in B, and vice versa
- Suppose an experiment is performed by drawing one outcome from either A or B.
- For the experiments together there are $|A| + |B| - |A \cap B|$ possible outputs
- When $|A \cap B|$ is the empty set we have the basic sum rule of counting.
- This is also called the **inclusion-exclusion principle**.

Generalised Sum Rule of Counting for 2 Sets

Example:

- An 8 bit string is sent over a network. The valid set of strings recognised by the receiver must start 01 or end with 10. How many valid strings are there ?

- Set A is the set of possible strings that begin with 01. $|A| = 64$ since of the 8 bits 2 are fixed while the remaining 6 bits can take on binary values, so there are $2^6 = 64$ such strings.

- Set B is the set of possible strings that end with 10. $|B| = 64$ since of the 8 bits 2 are fixed.

- $A \cap B$ is the set of possible strings that begin with 01 and end with 10. Since there are 4 bits fixed only 4 out of the 8 can vary so there are $2^4 = 16$ such strings i.e. $A \cap B = 16$

- By the generalised sum rule there are
$112 = |A| + |B| - |A \cap B| = 64 + 64 - 16$ such strings