CptS 483:04 Introduction to Data Science

Statistical Inference

Assefaw Gebremedhin
Processes and data

• The world we live in is complex, random and uncertain.
• The world we live in is one big data-generating machine.

• We would like ways to describe, understand, and make sense of processes around us
  • Because we just want to understand the world better,
  • Because understanding these processes is part of the solution to problems we are trying to solve.

• Data represents the traces of the real-world processes, and exactly which traces we gather are decided by our data collection or sampling method.
  • The observer is turning the world into data, and this is a subjective, not objective, process.
Sources of randomness and uncertainty

PROCESS

RANDOMNESS, UNCERTAINTY

DATA COLLECTION
Mathematical models or functions, known as Statistical estimators
Statistical Inference

Is the discipline that concerns itself with the development of procedures, methods and theorems that allow us to extract meaning and information from data that has been generated by stochastic (random) processes.
Populations and Samples

• Observations
  • Total: N
• Sample
  • Subset: n
  • Method should avoid bias
• Doesn’t “Big Data” make the notions of population and samples irrelevant?
  • Short answer: no
Why population/sample issue is still relevant in “Big Data”

• **Sampling solves some engineering problems**
  • Analysis/inference purposes (sampling fine)
  • Serving/UI purposes (may need all data)

• **Bias**
  • Conclusions should not be generalized beyond the source of the data
    • (Kate Crowford of Microsoft Research’s example of Hurrican Sandy)
  • Importance of context

• **Rethinking sampling**
  • In statistics, we make simplifying assumptions about the underlying generative process that created the data. The sample is one particular realization of the generative process.
  • In real-world, we have uncertainty in the sampling process (sampling distribution). We can then think of a full set of observations as a subset of some “super-population”, and hence a sample.

• **New kinds of data**
  • Traditional.  Text.  Records.  Geo-based data.
  • Network.  Sensor data.  Images.
A few ways to think about Big Data
(a side bar…)

• “Big” is a relative term and a moving target.
• “Big” is when you can’t fit it in one machine.
• Big Data is a cultural phenomenon.
• The 4 Vs: Volume, variety, velocity, and value.
Big Data: wrong assumptions and new meaning

• Faulty assumptions (sometimes made):
  • N = ALL
    • But, it is never so.
    • Example: election night polls
  • Data is objective
    • But, it is not. Context needs to be seen. Ignoring causation is a flaw.
    • Example: algorithm/model-based hiring
    • The emerging area of “fairness” (a mathematically rigorous study)
      (Cynthia Dwork’s keynote address at KDD17)

• Sample size of 1 (a new meaning):
  • User-level modeling
Modeling

What is a model?
Modeling: what is a model?

• Our attempt to understand and represent the nature of reality through a particular lens.

• An artificial construction where all extraneous detail has been removed and abstracted.

• Attention must always be paid to these abstracted details after a model has been analyzed to see what might have been overlooked.
  • For example, a model of protein backbone with side-chains is removed from the laws of quantum mechanics that govern the behavior of electrons, which ultimately dictate the structure and actions of proteins.
  • In a statistical model, we may have mistakenly excluded key variables, included irrelevant ones, or assumed structure divorced from reality.
Statistical modeling

• Getting a sense for what the underlying process might be.
  • What comes first? What influences what? What causes what? What is a test of that?
  • Write out mathematical expressions. Draw pictures.
    • Parameters and data

• How do you build a model?
  • Part art, part science
  • EDA is a great place to start!
  • Start simple, then build on complexity
  • Probability distributions are some of the building blocks
Common probability distributions

Interpreted as assigning a probability to a subset of possible outcomes, and have corresponding functions.
Kinds of distributions

• (Univariate) distributions
  • Distributions of single random variables (x). (Functions of one variable)
  • Expressed as \( p(x) \), which maps \( x \) to a positive real number
  • To be a probability density function, integral of \( p(x) \) (area under the curve) should be 1

• Joint distributions
  • Distributions of more than one random variable
  • E.g. in the case of two random variables, a distribution can be represented by the function \( p(x,y) \).
    • Takes values in a plane, and give nonnegative values
    • Double integral over the whole plane would be 1

• Conditional distribution
  • \( p(x|y) \) means density function of \( x \) given a particular value of \( y \)
  • Corresponds to subsetting when working with data
  • Same property as regular distributions (integral sums to 1, values are nonnegative)
Fitting a model

• Means you estimate the parameters of the model using the observed data

• Often involves optimization algorithms, such as MLE, to help get the parameters

• Is when you start actually coding
  • Your code reads in the data, you specify the functional form you wrote down on paper, R uses built-in optimization methods to give you the most likely values of the parameters given the data

• Overfitting: you used a dataset to estimate the parameters of your model, but your model isn’t that good at capturing reality beyond your sampled data
Contrasting three distributions: A side bar…

- Normal (Gaussian) Distribution
- Power Law Distribution
- Exponential Distribution
Nodes: WWW documents  Links: URL links

Over 3 billion documents

ROBOT: collects all URL’s found in a document and follows them recursively


Normal (Gaussian) Distribution vs Power Law Distribution

If it were Random

Normal

Power Law

$P(k) \sim k^{-\gamma}$
The difference between a power law and an exponential distribution

Above a certain $x$ value, the power law is always higher than the exponential.

$f(x) = cx^{-0.5}$

$f(x) = cx^{-1}$
The difference between a power law and an exponential distribution

This difference is particularly obvious if we plot them on a log vertical scale: for large $x$ there are orders of magnitude differences between the two functions.
Network Science: Scale-Free Property

Bell Curve
- Most nodes have the same number of links
- No highly connected nodes

Power Law Distribution
- Very many nodes with only a few links
- A few hubs with large number of links

Number of nodes with $k$ links
Number of links ($k$)

Number of nodes with $k$ links
Number of links ($k$)
Properties of Power Laws

• Scale invariance
  • Scaling $x$ in $f(x) = ax^{-k}$ by a constant $c$ causes only a proportionate scaling of the function:
    
    $f(cx)^{-k} = a(cx)^{-k} = c^{-k}ax^{-k} = c^{-k}f(x)$

• Lack of well defined average value
  • A power law $x^{-a}$ has a well defines mean over $[1, \infty)$ only if $a > 2$ and have a finite variance only when $a > 3$
  • Most identified power laws in nature have exponents such that mean is well defined but variance is not
  • This makes it incorrect to apply traditional stat based on variance and standard deviation. On the other hand, it allows “cost-effective interventions”

• Universality
  • Deeper origin in dynamical processes that generate power-law relations