BayesianStatistics-2014-02-21

Presentation slides printable summary: **preliminary version**

Title: **The potential impact of big data and Bayesian statistics on the business school curriculum**

Author/Presenter: **Robin Snyder, robin@robinsnyder.com**

Venue: **South East Decision Sciences meeting in Wilmington, NC, February 19-21, 2014**

1. Introduction

- Big data and computer science
- Probability and statistics
- Bayes Rule and frequentist methods
- Business school application

2. Big data

- Available data in increasing in size and granularity.
- Business decision making requires intelligently using this data.
- Computer science: Networking, databases, multi-processing, etc.
- Probability and statistics: decision models

3. Trends

Once data becomes too large to look at all the data, and one needs results based on many factors, query results will (and sometimes now have) error bars associated with them.

In computer science, a linear algorithm is needed to at least look at all of the data once.

As databases become bigger and bigger, the only way to get sub-linear algorithms is to not look at all of the data, which requires probabilistic models.

4. Future

Michael Jordan: Berkeley (machine learning, computer science, statistics):

- Data is getting really big.
- Sublinear algorithms are needed (cannot look at all the data).
- Computer science and statistics will be merged in 50 years.

See his YouTube video at http://www.youtube.com/watch?v=LFiwO5wSbi8.

5. Information

What is information?

How does it relate to statistics?

What is the probability that a fair coin, when flipped, lands heads?

6. Physics wave particle duality

Physics has two correct ways of looking at reality. Both are correct. One may work better in a given situation.

- Wave theory
- Particle theory

Physics has pretty much accepted wave-particle duality as both being correct ways of looking at reality.

7. Statistics

Statistics has two correct ways of looking at reality. Both are correct. One may work better in a given situation.

- Frequentist statistics (null hypothesis, confidence intervals, etc.)
- Bayesian statistics (inverse probability, probability of causes, etc.)

Many statisticians disagree over both frequentist and Bayesian statistics being correct ways of looking at reality.

8. Decision making

Business decision making (with or without computer assistance) requires making timely decisions often without complete information under ever changing circumstances.

Bayesian statistics works well for these types of decision-making problems.

On-line auctions and recommendation engines are other examples where hierarchical nonparametric Bayesian statistical models can be useful. Many SPAM filters use, in part, Bayesian statistics/logic.

9. Thomas Bayes

Rev. Thomas Bayes (mid 1700's) has the original idea as a specific instance of a problem, uses Newtons cumbersome notation to describe it.

10. Laplace

Laplace in late 1700's and early 1800's independently develops idea, credits Bayes for one insightful idea, refines and formalizes the ideas in an elegant way.

11. Others

Independently developed and used by many during the next 150 years (e.g., Alan Turing to help break Enigma during World War II, etc.).

12. Computational methods

In the 1980's, algorithmic advances using MCMC (Marchov Chain Monte Carlo) techniques make Bayesian system computationally tractable.

- Technique: Gibbs Sampling
- Software: BUGS, WinBuGS, OpenBUGS

13. Difference

According to Jordan, in general:

14. Frequentist approach

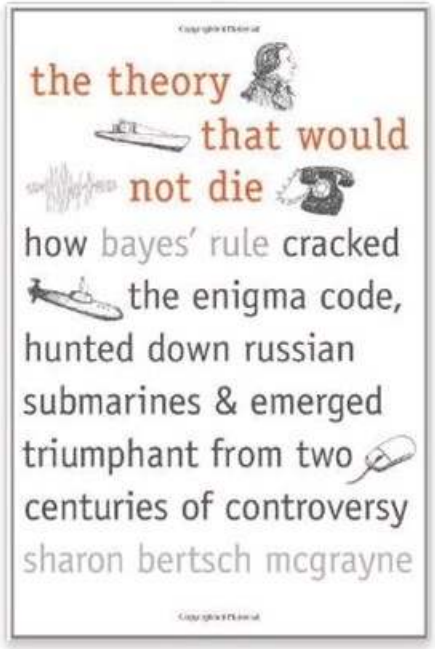
A frequentist approach will average over the possible data to see if the sample result is within a certain limit (i.e., confidence interval).

15. Bayesian approach

A Bayesian approach will look at just the available data, and what is known about the past data, in making a decision.

16. Reference

A good book on the history of Bayes Rule, and also of frequentist statistics, is "**The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy**" by Sharon Bertsch McGrayne.



17. Conditional probability

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

$$P(B|A) = \frac{P(A,B)}{P(A)}$$

These two (symmetric) conditional probability equations can be related by the common joint probability to get Bayes Rule.

18. Bayes Rule

Bayes Rule:

$$P(A|B)P(B) = P(B|A)P(A)$$

In usual form Bayes Rule appears as follows.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

19. Cromwell's Rule

"*As long as you are set that the probability is going to be zero, then nothing's going to change your mind.*" Mandansky.

"*I beseech you, in the bowels of Christ, think it possible you may be mistaken.*" Oliver Cromwell.

20. Arthur Bailey

Favorite introduction to Bayes by Author Bailey, accountant and Bayes Rule popularizer in the early 1950's.

"*If thou canst believe, all things are possible to him that believeth.*" Mark 9:23.

21. Bayes Rule

Let **A** be the proposed **Model** and **B** be the observed **Data**. Then Bayes rule becomes the following.

$$P(Model|Data) = \frac{P(Data|Model)P(Model)}{P(Data)}$$

- The ***posterior*** is **P(Model | Data)**.
- The ***likelihood*** is **P(Data | Model)**.
- The ***prior*** is ***P(Model)***.
- The ***evidence*** is **P(Data)**.

In real world calculations, the ***posterior*** is proportional to the ***likelihood*** times the **prior** so that the ***evidence*** can be ignored (in the denominator of the right side).

22. Bayesian statisticians

How do you recognize a Bayesian statistician?

"*Ye shall know them by their posteriors.*"

23. Stumbling block

Stumbling block: In the absence of information about the domain of application, what value should be used for the ***prior*** probability?

24. One approach

In the absence of information, the error in the ***prior*** probability is minimized by assuming equal odds (`50.0%`, or `0.5` probability for a sample space of two events) or a **uniform prior**.

Does this make you uncomfortable?

25. Graphical models

Consider A, B, and C as three distinct subsets of nodes an a DAG (Directed Acyclic Graph).

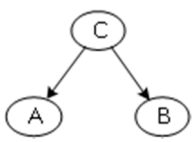
For nodes, A, B, and C, the joint probability is as follows.

$$P(A, B, C) = P(A \wedge B \wedge C)$$

The joint probability of nodes A, B, and C can be determined from the graphical model, which implicitly contains independence relations.

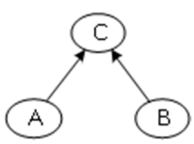
There are three primary ways in which distinct A, B, and C can be related. Other ways can be reduced to these three by symmetry considerations. For each, the joint probability can be expressed.

26. C to A and B



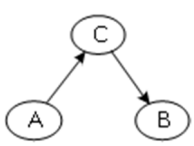
$$P(A, B, C) = P(A|C)P(B|C)P(C)$$

27. A and B to C



$$P(A, B, C) = P(C|A, B)P(A)P(B)$$

28. A to B to C



$$P(A, B, C) = P(B|C)P(C|A)P(A)$$

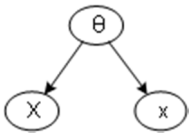
29. Compositional models

Grahapical models can be composed to allow arbitrarily complex models to be used.

30. Generative graphical models

A graphical model provides a convenient way in which to represent problems involving probabilities so that the **Model** can be inferred from the **Data**.

Consider the following simple graphical model.



In this diagram, θ represents a probability distribution of X that is to be estimated.

Assume that θ is known.

31. Machine learning

Machine learning is a collection of statistical techniques to infer models and make predictions, often in the processing of big data.

- supervised
- unsupervised

Everything being equal, an unsupervised method is preferred to a supervised method.

32. Topic modeling

- Infer topics from documents containing words
- Infer objects from images containing pixels
- Infer similarity from DNA containing nucleotides
- Recommend products from customers and purchases

33. Idea

A credible business school statistics course should cover both Bayesian and frequentist statistics, as each has areas of application.

Problems, if course time is not increased:

- What to include from Bayesian statistics.
- What to omit from traditional statistics course.

Any ideas?