Statistics & Machine Learning Models

Samir Rachid Zaim

11/12/2018

Modeling: Table of Contents

- What is modeling and what problems are modelable?
- What models and tools do we have available?
- What are ‘black-box’ algorithms?

Modeling

"There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. . . If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools." - Breiman [1]

Take home message:

- Tools are a means to end to solve a problem.
- CDSS must use a variety of tools to offer best solution

What is modelable?

In short, a well-defined clinical question or problem. That is, if a question has a well-defined, measurable outcome, then it can be modeled. Some examples

- How many people show up to the emergency room on a daily basis? Is there seasonal/holiday variation?
- Should we increase staffing for cardio-vascular units after thanksgiving? Are there an increased number of cardio-vascular related discharge after thanksgiving?

Modeling may take the form of a formal hypothesis test with a p-value

$$\text{e.g., } H_0 : \mu = \mu_0, \quad H_A : \mu = \mu_A$$

or modeling simply may be a tool we use we want to better understand.

From open-ended to modelable

Some of the more interesting and operational questions may be too broad to model or or may not be modelable per se. However, they can be reformulated to allow for partial answers to the question. For example, the question "Do gaps in insurance coverage have an effect on a patient's health?" is quite vast, but can be reformulated into smaller, concrete aspects of 'quality of care'.

- Are generic drugs prescribed more during periods of insurance gaps? Are they less-effective but cheaper options?
- Is there an increase in emergency room visits attributable to medication/prescription non-compliance?
- Do surveys indicate lower patient and/or provider satisfaction during months of insurance gaps?

Formalizing questions that modelable

In order to better illustrate what problems can be modeled using CDSS, we propose the following definitions:

Definition (1.1)

A ***well-defined clinical question*** is a question to which a measurable outcome exists or can be approximated.

Definition (1.2)

A ***modelable clinical question*** is a ***well-defined clinical question*** to which there are resources available to study.

Understanding the types of questions that are modelable via CDSS and the tools available to model them will allow for more effective use and understanding of CDSS.

Models

CDSS will - for the most part - deploy and be built from the following types of models:

- Rule-based models
- Statistical models*
- Machine Learning models*
- Network/Ontology models
- Natural Language Processing models

Some models* fall into both ‘machine learning’ and ‘statistics’, and these differences of opinion are resolved in calling them ‘statistical learning’. In short, another name for the same family of tools.

Models (2)

The overview of the following models will be explained with how they can fit into the CDSS framework rather than the math. The two main differences we focus on are in the types of problems being modeled. The two main types of problems differ in the outcomes they measure.

- Continuous numerical outcomes: a real, continuous number (e.g., patient's temperature as a response to medication)
- Discrete numerical outcomes: a real, discrete number (e.g., number of bed availabilities during summer v. during winter months)
- Categorical outcomes: mutually-exclusive outcomes (e.g., fever vs. no fever or heart-attack vs. no heart-attack)

"While it is important to know what job is performed by each cog, it is not necessary to have the skills to construct the machine inside the box!" - Gareth, Witten, Hastie, and Tibshirani [2]

Rule-based models

These models incorporate rules of conditional statements to determine the most appropriate treatment. May incorporate some statistical modeling or data,but once they are determined, they are decision trees with a determined structure.

- Decision Trees
- Logic Circuits and Logic Gates (AND,NOT, OR, Not-AND, Not-oR, exclusive OR)

Statistical Models

These models are less complex and more interpretable than machine learning models, offering traditional parameter estimation, with coefficients, confidence intervals, and p-values. The trade-off is smaller prediction power for greater interpretability. Some tools widely used in CDSS and medicine are:

- propensity scores: identify similar cohorts for treatment and control
- logistic regression: model binary outcomes and establish log-odds of outcomes based on covariates
- anova: studying effects of treatments on different groups
- Bayesian Modeling: a statistical framework for incorporating prior knowledge. Extremely helpful in small-sample studies.

Machine Learning Models

Machine learning (ML) generates a lot of hype and mystique - ignore that, and focus on its actual benefits and drawbacks in CDSS. ML offers the ability to process big data in real-time to develop recommendations and risk profiling among patients. They sacrifice interpretability for greater prediction ability and can handle a wider range of problems than traditional statistical modeling.

- Neural Networks: a machine-learning tool
- Support Vector Machines: a machine-learning tool to predict and classify different groups of patients
- Linear Discriminant Analysis: a statistical-learning tool to predict and classify different groups of patients
- Random Forests: create many decision trees and estimate an optimal average response

Natural Language Processing (NLP)

Most machine learning and statistical models only handle categorical or numeric data (that is, 'structured' in some sense), and therefore are not suited to handle free text - one of the major data sources in clinical informatics. NLP ranges from simple word matching (i.e., find me the word "heparin") to more complex regular expression and grammars tools, to mine text. These are greatly powerful tools for summarizing and synthesizing:

- clinical notes
- radiology reports
- surveys and product reviews

NLP can highlight or extract relevant information which can then be used in ML or statistical modeling to identify risk factors or important data elements in text.

Avoiding 'black-box' algorithms and making risk-scores intepretable

Black-box algorithms are defined as 'uninterpretable' or 'unclear' algorithms which provides the end-user with little intepretive power when they receive a flag. Variable/Feature selection is a major area of statistics and machine learning that identify the most 'important' covariates in a model. Key concepts:

- not all models produce coefficients or confidence intervals
- not all models can identify whether a covariate is protective/harmful
- different models measure **interactions** very differently
- some models sacrifice interpretability for higher predictive power (and vice-versa)

Effective CDSS will usually provide actionable items/suggestions with their high-risk flags. These action-items are usually a product of variable/feature selection.

Programming Languages in CDSS

CDSS will likely be built using a variety of computational and programming tools. The below are a few [open-source] tools likely used in CDSS

- R: Statistical Programming w/ Machine Learning libraries and limited NLP support
- Python: Object-oriented programming w/ a plethora of machine learning libraries and limited statistical support. Supports NLP (e.g., NLTK)
- Java: Object-orient programming w/ a plethora of machine learning libraries and limited statistical support. Supports NLP (e.g., OpenNLP)
- MySQL (My Structured Query Language): A relational database language and tool that supports building dashboards, monitoring clinical outcomes, querying/defining cohorts of interest, and limited descriptive statistical/NLP functionalities.

References

[1] Breiman, L., 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science, 16(3), pp.199-231.

[2] James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An introduction to statistical learning. Vol. 112. New York: springer, 2013.