Building Models

Max Turgeon

DATA 2010–Tools and Techniques in Data Science

Lecture Objectives

- Determine which modeling philosophy is behind a model
- Classify models according to our taxonomy

Motivation

- For the last part of the course, we will finally start to build models.
    - Linear and logistic regression
    - Clustering
    - Other machine learning approaches
- But first, we need more vocabulary and tools.
- We will start by discussing general principles for *building* models.
- Then we will discuss general principles for *evaluating* models.

Philosophies of Modeling

- When building models, it is useful to have guiding principles in mind.
- There are three main philosophies:
    - Occam's Razor/Parsimony
    - Bias-Variance Trade-Off
    - Probabilistic Reasoning

Occam's Razor

- The simplest explanation is best
    - Also called the “parsimony principle”, especially in statistics.
- In practice, this means that for similar performance, we should prefer the simpler model.
- E.g. if we can predict the Best Picture winner using 10 variables, why do it with 200?

Bias-Variance Trade-Off

- “All models are wrong, but some are useful.” George Box
- Assuming the same accuracy, reducing the bias of a model will increase its variability (and vice-versa).
- Bias usually occurs when we make erroneous assumptions.
- Variance usually means our model is sensitive to small changes in the data
    - Also related to the concept of *overfitting*.

Probabilistic Reasoning

- Models should provide probability statements (instead of point predictions)
- Predictions should be updated in response to new information
- This is related to Bayesian thinking (see STAT 4150)

Exercise

Describe which principle is behind choosing each of the following model:

1. Predicting velocity using laws of physics instead of a neural network based on 1 million observations of objects falling.
2. Probability of rain is between 50% and 80% instead of only saying 70%.
3. Predicting the Best Picture winner using 10 variables instead of 200.

Solution

1. Variance-bias trade-off
2. Probabilitistic Reasoning
3. Occam's Razor

Taxonomy of Models

- We can classify models along different axes:
    - Linear vs Non-linear
    - Blackbox vs Descriptive
    - First-Principle vs Data-Driven
    - Stochastic vs Deterministic
    - Flat vs Hierarchical

Linear vs Non-linear

- Linear models typically give a weight (or importance) to each variable and add up their contributions to create a score.
  - Linear and logistic regression are examples of linear models.
- Non-linear models will either implicitly or explicitly model complex relationships.
  - Polynomial, exponential, logarithm, complex interactions.
  - E.g. Random forests, Support Vector Machine

Blackbox vs Descriptive

- Descriptive models will be able to provide an explanation of the effect of a variable on the prediction/output.
- Blackbox models will only produce predictions/outputs without any explanation.
    - Blackbox models can be more accurate, but more difficult to interpret.

First-Principle vs Data-Driven

- First-principle models are typically constructed based on prior knowledge of the system.
    - Laws of physics, biological processes, etc.
- Data-driven models will use the collected observations to *learn* a model of the system.
- First-principle models are often biased, but data-driven models are at risk of overfitting.

Stochastic vs Deterministic

- This is related to the previous point.
- Deterministic models will provide a single output given the same output.
    - Changing the training data will not change the prediction.
    - First-principle models are often deterministic.
- Stochastic models incorporate the randomness of the data into the output.
    - Also relates to probabilistic reasoning.

Flat vs Hierarchical

- Flat models will consider all predictors on an equal footing.
- Hierarchical models incorporate a hierarchy of predictors.
  - To predict future price of a stock, use information about economy, company's finances, etc.
  - For the economy or company's finances, use a different model that incorporates other predictors.
- Hierarchical models are more complex, but also more flexible.
  - You can combine first-principle and data-driven models for the submodels.

Exercise

For each of the prediction tasks below, discuss whether first-principle or data-driven models seem to be the more promising approach:

- Movie gross.
- Baby weight.
- Art auction price.
- Snow on Christmas.
- Super Bowl champion.
- Future oil price.