Additional Topics

Michael Thomas

AI in Marketing - Semester II, 2022

Preview of Topics

- Data cleaning
- Feature engineering
- Model Tuning
- Neural Networks
- Unsupervised learning
- Reinforcement learning
- Privacy Considerations
- AI Bias
- Synthetic data

Cleaning Data

- Lots of time is spent cleaning data.
  - Put variables in the right data types.
  - Fix missing or incorrect entries.
  - Combine data from different sources correctly.
- These activities are essential to the success of causal estimation with observational data
  - Causal estimates can be very sensitive to data errors
- Also important to predictive modeling
  - Can improve the quality of predictions.
  - Must trade off improving prediction quality with human effort, which is costly.

Feature Engineering

- A step beyond data cleaning.
- Combining, adjusting existing data to create new variables that are likely to help predict outcomes.
- Requires human judgment.
- Also important to remove features that are known to have poor predictive power.
- Just like comparing models, you can experiment with which features work best.

Feature Engineering: Example

- Your company sells ski gear.
- You know that people who live in places where it snows buy new ski gear when it snows.
- However, people who live in places without snow buy new ski gear before holidays.
- Based on this knowledge you might want to create new a new variables that captures the probability of snow in a place.
- You might interact this variable with week of the year.
- ML could discover this relationship with enough data, but if you give it the right variables it can make use of the relationship faster.

Feature Engineering

Common data transformations

- "standardize" the variables
  - $z = \frac{x - mean(x)}{sd(x)}$
  - 'glmnet' does this automatically for all features by default.
- "log transformation"
  - $z = \log(x + 1)$
- These may help with linear models, but not tree-based models.
- Also, "interaction" of features:
  - $z = x_1 \cdot x_2$
  - $z = x \cdot \mathbb{I}\{t \in [t_1, t_2]\}$ features over specific time intervals that seem important, (e.g., seasons, times of the day).

Model Tuning

- Cross-validation (CV) can assist with model tuning.
  - Try out different combinations of parameters
  - Use CV to test which gives the best predictions
- Some models have more parameters than others
- Some models are more responsive to model tuning than others.
  - LASSO: responsive; built-in CV support to select λ.
  - Random Forest: less responsive, default settings work well.
  - XGBoost: very responsive

Model Tuning: Examples

- LASSO:
  - just one parameter, λ.
  - We tuned it using cross-validation.
- Random Forest tuning parameters:
  - Feature bagging criteria: share of variables considered for splitting each node.
  - Maximum depth of trees
  - Minimum samples allowed in a terminal leaf.
  - Number of trees.
  - Random forest often performs well with the default parameter settings.

Model Tuning: Examples

- LASSO:
  - just one parameter, λ.
  - We tuned it using cross-validation.
- Random Forest tuning parameters:
  - Feature bagging criteria: share of variables considered for splitting each node.
  - Maximum depth of trees
  - Minimum samples allowed in a terminal leaf.
  - Number of trees.
  - Random forest often performs well with the default parameter settings.

Neural Networks

- In the news as producing some of the best predictions
- Can handle complex data, like images, sounds, translations.
- Big requirements.
  - Million+ observations for training
  - Lot's of tuning.
  - Slow to run.
- With a large data set and properly tuned, it can outperform other methods by a wide margin.

Broad Areas of Machine Learning

- **Supervised Learning**
  - What we have done is this course
  - Given data on y and X, learn f to predict $y = f(X)$.
- **Unsupervised Learning**
  - Find clusters of similar features in X
  - Learn to mimic the relationships in X
  - No y values.
  - E.g., cluster methods like hierarchical clustering or k-means
    - These can be used for segmentation analysis in marketing
- **Reinforcement learning**
  - Next slide

Reinforcement Learning

- Algorithms that learn to optimize playing a game.
- Anticipates how current decisions might affect payouts later on.
- Conceptually similar to multi-armed bandits
- Experiment to determine how to maximize long-run rewards
- Could be applied to
  - Repeated interactions with a customer to maximize CLV
  - Dynamically changing ads over time.

Privacy Considerations: GDPR

- From 2018, the General Data Protection Regulation (GDPR) took effect in the European Union (EU).
- Regulates:
  - Use of personal data in the EU.
  - Transfer of personal data out of the EU.
  - Individuals may prevent use of personal data for marketing purposes.
  - Individuals must opt-in to have their data used.
- Similar laws followed in several other countries:
  - California Consumer Privacy Act from 2018
  - Personal Information Protection Law (PIPL) in China came into force from 2021.
  - Also, Turkey, Brazil, Chile, Argentina, Japan, South Korea, South Africa.
- Laws limit where and when you can use personal data for business applications.

Privacy Considerations

Key points for the business to know

- Laws differ by region.
  - This may imply different data handling for each region you work in.
- Personally Identifiable Information (PII) is not what you might intuitively think according to laws.
  - E.g., IP address are PII
- Businesses need data governance procedures that cover:
  - Collection of data
  - Use of data
  - Sharing of data

Privacy Considerations: Synthetic Data

- Synthetic data may provide balance between privacy and business optimization.
- Synthetic data is based on real data and retains its statistical properties, but has been simulated.
- Synthetic data can act as a proxy for real data.
- Balancing act: fully reflect the relationships in the original data without revealing the original people in the data.
- Potential to manage privacy concerns:
  - No GDPR implications
  - No liability if the data a breached.

AI Bias

- Prediction technology allows for personalization of business activities
  - Offer loans to those least likely to default
  - Show ads to those most likely to click.
- Many countries have laws against discriminating based on some **protected** attributes:
  - E.g., gender, race, religion, age.
  - Many consumers find such practices unfair.

AI Bias

- Algorithms will use **protected** attributes if they are predictive
- Algorithms will find ways to proxy for **protected** attributes if they are left out but predictive.
  - Sometimes creating more bias than if the protected attributes were included in the data.
- Current work on how to prevent bias in algorithms.
  - Ascarza and Israeli 2021 build on Causal Forest
  - "Bias-Eliminating Adaptive Trees"

AI Bias and Synthetic Data: Chanel Case Study

- Chanel wanted to develop an iPhone app that would:
  - Start with a photo of any color provided by the user
  - Find the Chanel lipstick that matched that color the best
  - Let the user "try on" the lipstick using augmented reality.
- Chanel already had a large number of photos with lipstick that could be used as training data.
  - This allowed them to avoid scraping images from the web, "non-consensual" images.
- Challenge: Chanel's photos included a disproportionately large number of white people.
  - Non-whites would have less training data and therefore worst performance.
- Solution: Synthetic data
  - Create simulated images of people to train.
  - Chanel doesn't collect any pictures taken by users, which further protects them from privacy concerns.