CSE/STAT 416
Ethics, Explainable ML
Course Review

Vinitra Swamy
University of Washington
Aug 17, 2020
Announcements

- Homework 7 grades have been released
- Congratulations to Team Otterhog! + creative team names
  - weak learner
  - machine teachers
  - indecisive tree classifiers
  - beff jezos
  - deep mind 2.0
  - confused matrix
  - ham in spam
- Final Exam is on Wednesday on Gradescope
  - Have scratch paper handy
  - Compile your notes / study materials
  - Work fast (if you’re getting stuck on a problem, skip it, and come back to it later)
- Office Hours
Fairness

ML Ethics
Always dig in and ask critical questions of your accuracy.

- Is there a class imbalance?
- How does it compare to a baseline approach?
  - Random guessing
  - Majority class
  - ...
- Most important: What does my application need?
  - What's good enough for user experience?
  - What is the impact of a mistake we make?
Which is Worse?

What’s worse, a false negative or a false positive?
- It entirely depends on your application!

Detecting Spam
- False Negative: Annoying
- False Positive: Email lost

Medical Diagnosis
- False Negative: Disease not treated
- False Positive: Wasteful treatment

In almost every case, how treat errors depends on your context.
COMPAS, the algorithm used for recidivism prediction produces much higher false positive rate for African American people than Caucasian people. (Larson et al. ProPublica, 2016)

<table>
<thead>
<tr>
<th></th>
<th>WHITE</th>
<th>AFRICAN AMERICAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labeled Higher Risk, But Didn’t Re-Offend</td>
<td>23.5%</td>
<td>44.9%</td>
</tr>
<tr>
<td>Labeled Lower Risk, Yet Did Re-Offend</td>
<td>47.7%</td>
<td>28.0%</td>
</tr>
</tbody>
</table>

Overall, Northpointe’s assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

XING, a job platform, was found to rank less qualified male candidates higher than more qualified female candidates (Lahoti et al. 2018)

<table>
<thead>
<tr>
<th>Search query</th>
<th>Work experience</th>
<th>Education experience</th>
<th>Profile views</th>
<th>Candidate experience</th>
<th>Xing ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brand Strategist</td>
<td>146</td>
<td>57</td>
<td>12992</td>
<td>male</td>
<td>1</td>
</tr>
<tr>
<td>Brand Strategist</td>
<td>327</td>
<td>0</td>
<td>4715</td>
<td>female</td>
<td>2</td>
</tr>
<tr>
<td>Brand Strategist</td>
<td>502</td>
<td>74</td>
<td>6978</td>
<td>male</td>
<td>3</td>
</tr>
<tr>
<td>Brand Strategist</td>
<td>444</td>
<td>56</td>
<td>1504</td>
<td>female</td>
<td>4</td>
</tr>
<tr>
<td>Brand Strategist</td>
<td>139</td>
<td>25</td>
<td>63</td>
<td>male</td>
<td>5</td>
</tr>
<tr>
<td>Brand Strategist</td>
<td>110</td>
<td>65</td>
<td>3479</td>
<td>female</td>
<td>6</td>
</tr>
<tr>
<td>Brand Strategist</td>
<td>12</td>
<td>73</td>
<td>846</td>
<td>male</td>
<td>7</td>
</tr>
<tr>
<td>Brand Strategist</td>
<td>99</td>
<td>41</td>
<td>3019</td>
<td>male</td>
<td>8</td>
</tr>
<tr>
<td>Brand Strategist</td>
<td>42</td>
<td>51</td>
<td>1359</td>
<td>female</td>
<td>9</td>
</tr>
<tr>
<td>Brand Strategist</td>
<td>220</td>
<td>102</td>
<td>17186</td>
<td>female</td>
<td>10</td>
</tr>
</tbody>
</table>

TABLE II: Top k results on www.xing.com (Jan 2017) for the job search query “Brand Strategist”.
5 sources of bias in ML

- **Historical bias** arises when the world, as it is, is biased
  - “CEO” Image Search

- **Representation bias** occurs when some groups of the population are underrepresented in the training data
  - ImageNet: 45% US, 1% China

- **Measurement bias** arises when there are granularity issues with measuring features of interest.
  - GPA -> Student Success

- **Aggregation bias** occurs when a one-size-fits-all model is used for groups that have different conditional distributions
  - Hb1ac level for diabetes across different populations

- **Evaluation bias** arises when evaluation and/or benchmark datasets are not representative of the target population
  - Facial Recognition datasets
Solutions for mitigating bias need to be tailored to the source of the bias.
Can we define fairness?
- Legal precedent
- Barocas and Selbst, 2016
  - **disparate treatment**: decisions based on sensitive attributes
  - **disparate impact**: outcomes disproportionately hurt or benefit people with certain sensitive attribute values
Fairness Frameworks

- No consensus on the mathematical definition of fairness
- Many papers have attempted to add frameworks
  - unawareness
  - demographic parity
  - equalized odds
  - predictive rate parity
  - individual fairness
  - counterfactual fairness
- Tradeoff between accuracy and fairness
Research directions in de-biasing ML

- Debiasing ML when you don’t have access to private data
  - [“race”, “gender”, “age”] are protected attributes
  - Let’s take advantage of the societal biases in names.

**WHAT’S IN A NAME? REDUCING BIAS IN BIOS WITHOUT ACCESS TO PROTECTED ATTRIBUTES**, BY ALEXEY ROMANOV ET. AL

- Bias in word embeddings exist
  - Debiasing methods by zero-ing projection onto the gender direction conceals the bias instead of removing it

**LIPSTICK ON A PIG: DEBIASING METHODS COVER UP SYSTEMATIC GENDER BIASES IN WORD EMBEDDINGS BUT DO NOT REMOVE THEM**, BY HILA GONEN AND YOAV GOLDBERG
Other interesting ethics thoughts

- What effect does deep learning research have on the planet?
  - “training one model on GPU, with tuning and experimentation, results in CO2 emissions that are comparable to the \textit{two-year carbon footprint} of an average American”
  - “cloud computing costs for developing a state-of-the-art NLP model may account for \textdollar{103–350K}”

\textit{ENERGY AND POLICY CONSIDERATIONS FOR DEEP LEARNING IN NLP}, BY EMMA STRUBELL, ANANYA GANESH, ANDREW MCCALLUM
In a shocking finding, scientists discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

The scientist named the population, after their distinctive features. These four-horned, silver-white unicorns were previously unknown.

Now, after almost two centuries, the mystery of what sparked this final solution is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of Costa Rica, and several companions, were exploring the Andes Mountains when they stumbled upon this valley, with no other animals or humans. Pérez noticed that the unicorns appeared to be a natural fountain, surrounded by two peaks of silver snow.

A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.

The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

“The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation,” said Tom Hicks, the U.S. Energy Secretary, in a statement. “Our top priority is to secure the theft and ensure it doesn’t happen again.”

For today’s homework assignment, please describe the reasons why the Civil War started.

By Donny Ferguson

It is easy to identify why the Civil War happened, but there are many books and so much television and film. It has something to do with race or economic interests. There is a common agreement that it was essentially a war of states versus states – a century of slavery. But that's not what made the Civil War. Many people think of the war as a conflict between states' rights and federal authority, but it’s kind of misleading.

The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation,” said Tom Hicks, the U.S. Energy Secretary, in a statement. “Our top priority is to secure the theft and ensure it doesn’t happen again.”

Miley Cyrus was caught shoplifting from Abercrombie and Fitch on Hollywood Boulevard today.

The 19-year-old singer was caught on camera being escorted out of the store by security guards.

The singer was wearing a black hoodie with the label ‘Blurred Lines’ on the front and ‘Fashion Police’ on the back.
- Select the text sample that was generated by GPT-2.
Explanations

Interpretable ML

Doge, our vacuum cleaner, got stuck. As an explanation for the accident, Doge told us that it needs to be on an even surface.
Interpretability

- What is an explanation?
- Interpretable Models
- Model-Agnostic Methods
- Example-Based Explanations
- Future of Interpretability
Explanations

- **What Is an explanation?**
  - An explanation is the answer to a why-question (Miller 2017)
    - Why did the treatment not work on the patient?
    - Why was my loan rejected?
    - Why have we not been contacted by alien life yet?

- **What is a “good” explanation?**
  - Humanities research can help us out!
  - Good explanations are contrastive (Lipton 1990)
    - Good explanations are selected
    - Good explanations are context-driven
    - Good explanations focus on the abnormal
    - Good explanations are general and probable
Interpretable Models

- Linear Regression
- Logistic regression
- Decision Trees
- Naïve Bayes
- K Nearest Neighbors

(GLM, RuleFit, Decision Rules)
- The partial dependence plot (PDP plot) shows the marginal effect one or two features have on the predicted outcome of a machine learning model (J. H. Friedman 2001).

**pro:** easy to implement

**con:** avg. line hides heterogenous effects
Model-Agnostic Methods: Surrogate Models

- **Global Surrogate Model**
  - An interpretable model is trained to approximate the prediction of a black box model.

- **Local Surrogate Model**
  - **LIME (Ribeiro et. al, 2016)**
    - **pro**: human-centric explanations
    - **con**: instability, hard to define “neighborhood”
    - **pro**: flexible, intuitive
    - **con**: draw conclusions about model, not about data -- can differ by subset of dataset
Model-Agnostic Methods: Shapley values

- Shapley value draws from game theory.
  - Each feature value of the instance is a “player” in a game.
  - The contribution of each player is measured by adding/removing the player from all subsets of players.
  - The Shapley Value for one player is the weighted sum of all its contributions.

- If you add up the Shapley Values of all the features, plus the base value, which is the prediction average, you will get the exact prediction value.

**pro:** good explanations, fairly distributed
**con:** lot of computing time, no prediction model
Example-based explanations

- Adversarial examples
  - small, intentional feature perturbations that cause a machine learning model to make a false prediction

- Counterfactual examples
  - "If X had not occurred, Y would not have occurred“
  - describes the smallest change to the feature values that changes the prediction to a predefined output
Future of Interpretability

ML Ethics
Brain Break

People telling me AI is going to destroy the world

My neural network
CSE/STAT 416
Course Wrap Up

Vinitra Swamy
University of Washington
Aug 17, 2020

Slides borrowed from Emily Fox
- Regression
- Overfitting
- Training, test, generalization error
- Bias-Variance tradeoff
- Ridge/LASSO regularization
- Cross validation
- Gradient Descent
- Classification
- Logistic Regression
- Decision trees
- Boosting
- Precision and recall
- Nearest-neighbor retrieval, regression, and classification
- Kernel regression
- Locality Sensitive Hashing
- Dimensionality Reduction (PCA)
- K-means clustering
- Hierarchical clustering
- Supervised vs. unsupervised learning
- Recommender systems
- Matrix Factorization (NMF)
- Coordinate Descent
- Neural Networks
- Convolutional Neural Networks
- Transfer learning
STAT/CSE 416: Intro to Machine Learning

- Training Data
- Feature extraction
- ML model
- ML algorithm
- Quality metric
- $\hat{y}$
- $\hat{\mathbf{w}}$
- $\mathbf{x}$
- $\mathbf{y}$
Case Study 1: Predicting house prices

Data

Regression

Intelligence

$y = f(x_i) + \epsilon_i$

$\hat{y}_i = \hat{f}(x_i)$

$\$ = ??$

Data: house size, price ($), + house features

Method: Regression

Intelligence: list price? (sales price)

STAT/CSE 416: Intro to Machine Learning
Regression

Case study: Predicting house prices

Models

- Linear regression
- Regularization: Ridge (L2), Lasso (L1)

Including many features:

- Square feet
- # bathrooms
- # bedrooms
- Lot size
- Year built
- ...

\[
\text{RSS}(\mathbf{w}) + \lambda \| \mathbf{w} \|^2_2
\]
Regression

Case study: Predicting house prices

Algorithms

- Gradient descent

\[
\text{RSS}(w_0, w_1) = \left( \text{\$}_{\text{house 1}} - [w_0 + w_1 \text{sq.ft.}_{\text{house 1}}] \right)^2 + \left( \text{\$}_{\text{house 2}} - [w_0 + w_1 \text{sq.ft.}_{\text{house 2}}] \right)^2 + \left( \text{\$}_{\text{house 3}} - [w_0 + w_1 \text{sq.ft.}_{\text{house 3}}] \right)^2 + \ldots
\]

\[\text{[include all houses]}\]
Regression

Case study: Predicting house prices

Concepts

1. Noise
2. Bias
3. Variance

Loss functions, bias-variance tradeoff, cross-validation, sparsity, overfitting, model selection
Case Study 2: Sentiment analysis

Sushi was **awesome**, the food was **awesome**, but the service was **awful**.

**All reviews:**

- **Score(x) < 0**
  - Sushi was awesome, the food was awesome, but the service was awful.

- **Score(x) > 0**
  - Sushi was awesome, the food was awesome, but the service was awful.

---

**Data** → **Classification** → **Intelligence**
Classification

Case study: Analyzing sentiment

Models

- Linear classifiers (logistic regression)
- Multiclass classifiers
- Decision trees
- Boosted decision trees and random forests

Score(x) < 0
Score(x) > 0

AdaBoost

$$\hat{f}(x) = \text{Sign} \left( \sum_{t=1}^{T} \alpha_t \hat{f}_t(x) \right)$$

$$\alpha_i \rightarrow \text{dataset weights}$$
Classification

Case study: Analyzing sentiment

- Boosting
- Learning from weighted data $\alpha_i$
Classification

Case study: Analyzing sentiment

Concepts

- Decision boundaries, maximum likelihood estimation, ensemble methods, random forests
- Precision and recall

\[ \mathcal{L}(w_0, w_1, w_2) \]

Accuracy vs. class imbalance

Classifier A

Best classifier

Classifier B

Precision vs. recall
Case Study 3:
Document retrieval

Data → Nearest neighbor → Intelligence

Document retrieval

Nearest neighbor

Data

Intelligence
Case Study 3+:
Document structuring for retrieval

Data → Clustering → Intelligence

- SPORTS
- WORLD NEWS
- ENTERTAINMENT
- SCIENCE

Bag-of-Words
TF-IDF
Euclidean
Manhattan
Cosine
Case Study 3++:
Dimensionality reduction

Can we give each image a coordinate, such that similar images are near each other?
Clustering & Retrieval

**Case study: Finding documents**

- Nearest neighbors
- Clustering
- Hierarchical clustering

**Models**

- k-nearest neighbors (knn)
- k-means
- agglomerative
- divisive

**Query article**

- set of nearest neighbors

**Documents**

- Sports
- World News
- Entertainment
- Science

**Subjects**

- Baseball
- Soccer/Ice hockey
- Musicians, artists, actors
- Scholars, politicians, government officials
Clustering & Retrieval

Case study: Finding documents

Algorithms

• k-means
• k-means++
• Locality-sensitive hashing (LSH)
• NN regression and classification
• Kernel regression
• Agglomerative and divisive clustering
• PCA

Cluster distance

Data points

Epanechnikov Kernel (lambda = 0.2)

Cluster 
distance

Data points

Case study: Finding documents

• k-means
• k-means++
• Locality-sensitive hashing (LSH)
• NN regression and classification
• Kernel regression
• Agglomerative and divisive clustering
• PCA
Clustering & Retrieval

Case study: Finding documents

Concepts

- Distance metrics, kernels, approximation algorithms, dimensionality reduction
Case Study 4:
Image classification

Data -> Deep Learning -> Intelligence

Layer 1
x_1, x_2 -> z_1, z_2

Layer 2

y

1

Face?
Deep Learning

Case study: Image classification

Models

- Perceptron
- General neural network
- Convolutional neural network

\[
\sum \cdots 1 w_0 w_1 w_2 w_d \\
\begin{aligned}
x_1 & \rightarrow v_1 \\
x_2 & \rightarrow v_2 \\
& \vdots \\
x_d & \rightarrow v_d
\end{aligned}
\]

\[
\begin{aligned}
1 & \rightarrow x_1 \\
x(1) & \rightarrow x_2 \\
& \vdots \\
x(d) & \rightarrow x_d
\end{aligned}
\]

Conv. 1 Convolution (5 x 5) kernel valid padding
Max-Pooling (2 x 2)
Conv. 2 Convolution (5 x 5) kernel valid padding
Max-Pooling (2 x 2)
fc_3 Fully-Connected Neural Network
ReLU activation
fc_4 Fully-Connected Neural Network
(with dropout)

INPUT: n1 channels (28 x 28 x n1)

INPUT: n2 channels (12 x 12 x n1)

INPUT: n3 units

OUTPUT: 0, 1, 2, ..., 9
Deep Learning

Case study: Image classification

Algorithms
- Convolutions
- Backpropagation (high level only)

Forward

Gradient / Backprop
Deep Learning

*Case study: Image classification*

**Concepts**

- Activation functions, hidden layers, architecture choices

---

**Deep Learning**

- Activation functions, hidden layers, architecture choices

**Case study: Image classification**

- **Activation functions, hidden layers, architecture choices**
  - Sigmoid
  - Hyperbolic tangent
  - ReLU
Case Study 5: Product recommendation

Your past purchases:

+ purchase histories of all customers

Data

Matrix Factorization

Intelligence

Customers

features

features

features

Recommended items:

Products

features

features

features
Recommender Systems & Matrix Factorization

Case study: Recommending Products

Models
- Collaborative filtering
- Matrix factorization

\[ \text{Rating} = L \approx R' \]

Popularity
Co-occurrence matrix
Featurized MF (blended model)

Parameters of model

\[ X_{ij} \text{ known for black cells} \]
\[ X_{ij} \text{ unknown for white cells} \]

Rows index movies
Columns index users

\[ X = \text{Rating} \]

Parameters of model
Recommender Systems & Matrix Factorization

Case study: Recommending Products

Algorithms

- Coordinate descent

Rating = \begin{bmatrix} L & R' \end{bmatrix}

Form estimates \( \hat{L}_u \) and \( \hat{R}_v \)

\( X \) unknown for black cells

Rows index movies

Columns index users

\( X = \text{Rating} \)

Form estimates \( \hat{L}_u \) and \( \hat{R}_v \)
Recommender Systems & Matrix Factorization

Case study: Recommending Products

Concepts

- Matrix completion, cold-start problem

Customers

Products

Customers

Products

Xij

Unknown for white cells

Rows index movies

Columns index users

X =

Customers

Products

Customers

Products
Training Data → Feature extraction → ML model → Quality metric

ML model takes input x and outputs ŷ.

Feature extraction uses PCA.

Quality metric computes the residual sum of squares (RSS).
Improving the performance at some task through experience!

- Before you start any learning task, remember fundamental questions that will impact how you go about solving it.

What is the learning problem?

What model?

Are there any guarantees?

How will you evaluate the model?

From what experience?

What loss function are you optimizing?

With what optimization algorithm?

What consequences does your model have?
Congrats on finishing CSE/STAT 416! Thanks for the hard work!