STAT 214 Spring 2025 Week 9

Austin Zane

Lab 2

- Due at midnight. **No exceptions!**
- Make sure you have a lab2 directory. All group members must submit the lab.
- Teammate evaluations will play large role in grades.
- Should we reshuffle groups?
    - PollEv.com/austinzane912

Lab 3

- Will be released on Monday
- Text-fMRI problem to understand the brain:
  - 1. Using classical linear models to predict voxel response.
  - 2. Transformers for prediction.
  - 3. Interpreting transformers.
- Can read this paper to get started.

Incorporating Context into Language Encoding Models for fMRI

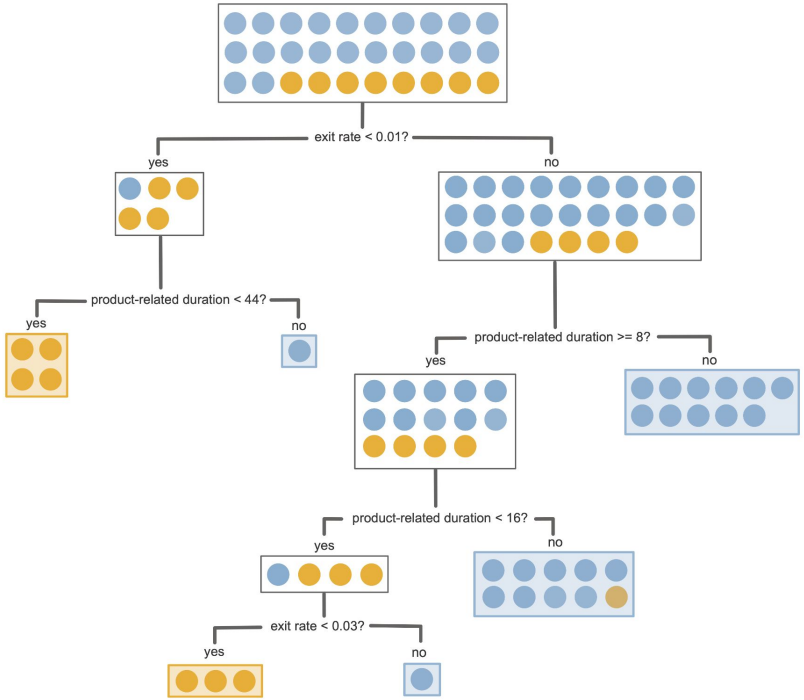
Shailee Jain[1] **Alexander G Huth**[1,2]
Departments of [1]Computer Science & [2]Neuroscience
The University of Texas at Austin
Austin, TX 78751
{shailee, huth}@cs.utexas.edu

Interpretability

Tree models

Decision tree:

- Just plot it as a tree and look at it (scikit-learn has built-in ways to do this, or do it manually and make it look nice)
- Hopefully it's not too big
- Can maybe represent how the training data is filtered by the tree?

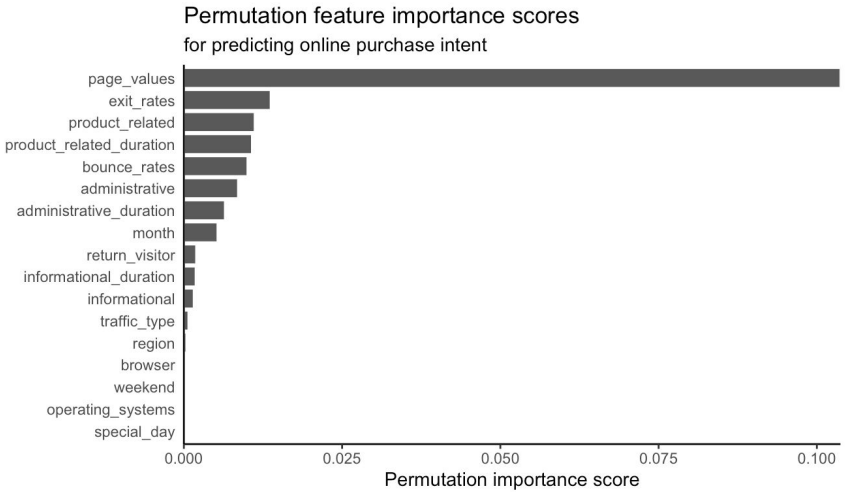


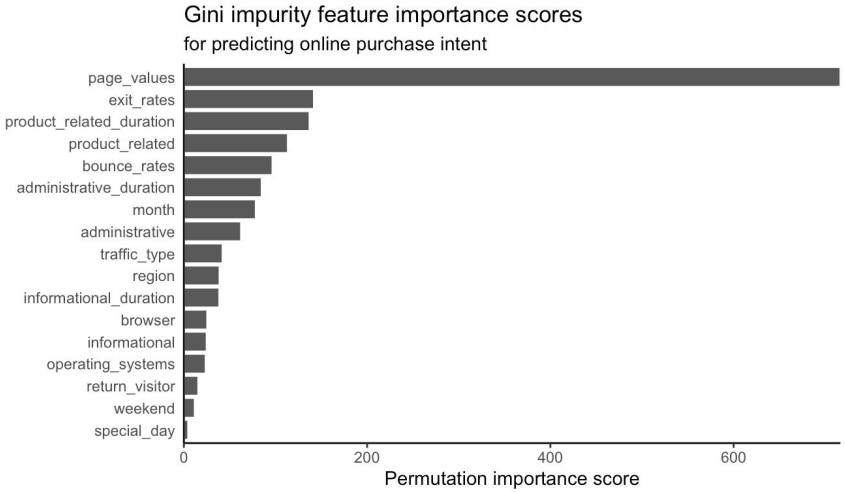
https://vdsbook.com/12-rf#the-classification-and-regression-trees-algorithm

Tree models

Random forest

- Too many trees to plot them!
- Measure **feature importance**
  - **permutation**: measure how performance (e.g. out of bag error) changes if you randomly permute one of the features.
  - **Gini impurity**: measure how much the “impurity” decreases in splits involving each variable
  - **Mean Decrease in Impurity**





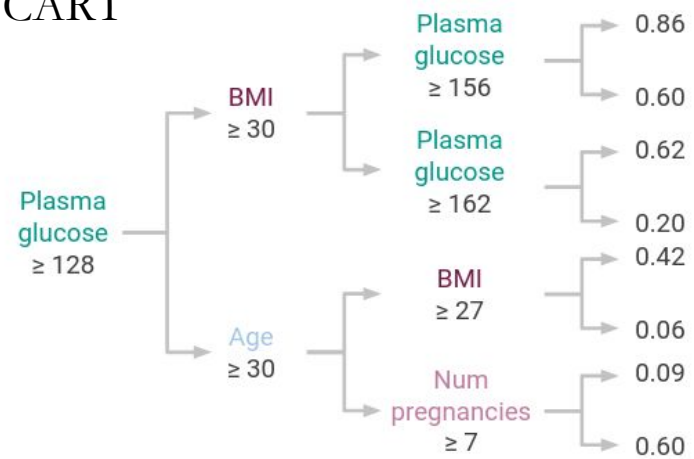
https://vdsbook.com/12-rf#random-forest-feature-importance-measures

Tree models

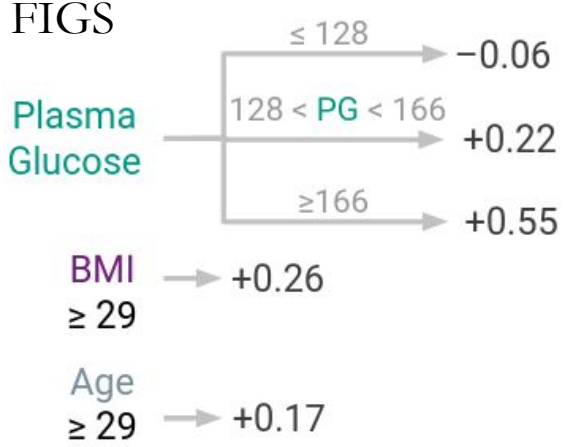
Remember that there are also other tree models, some of which are very interpretable!

https://github.com/csinva/imodels ← good stuff from Bin's group

CART



FIGS



https://arxiv.org/pdf/2201.11931

Linear models

Interpret by **looking at coefficients**

- **Intercept:** the prediction you'll make for an observation where all variables are zero (???). This usually doesn't make much sense. To improve interpretability, can center all the variables. Then the intercept will be the prediction for the "average" observation.
- **Other coefficients:** "holding all other variables constant, an increase of 1 in this variable changes our prediction by [coefficient]". Does this make sense when variables are correlated? Also won't work if you have nonlinear terms or interaction terms.
- **Comparing coefficients:** It does not make sense to compare the magnitude of coefficients to measure their "importance" unless the variables are on the same scale. Still doesn't make much sense if variables are correlated though.
- **Significance:** a statement about the magnitude of the coefficient relative to its variability. Hard to interpret usefully in a domain context.

https://vdsbook.com/10-ls_continued#linear-fits-with-multiple-predictive-features

Linear models: logistic regression

Interpretation becomes a little more tricky.

We are linearly modeling the log-odds: log(p/(1-p)). So “with an increase of 1 in ___ while holding all other variables constant, the model predicts a change in the log-odds of [coefficient]”

Deep learning models

Key idea: **attributions**

- **Feature attributions:** how the output of the model changes when the input changes. Understand the DL model as a whole.
- **Neuron attributions:** how individual neurons in the model change when the input changes. Understand the internals of the DL model.
    - Will discuss this further in future discussions.

Feature attributions

Two general approaches:

- **Gradient**-based: compute the gradient of the output of the model (i.e. the prediction) with respect to the inputs
- **Perturbation**-based: change the inputs of the model, plug them through, and see how the output changes.

You can make very nice plots if your inputs are images :)

Feature attributions: Integrated Gradients

3. Our Method: Integrated Gradients

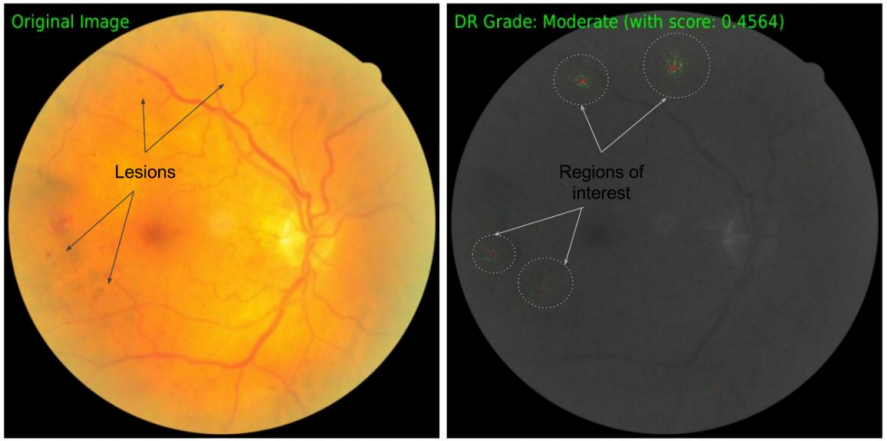
We are now ready to describe our technique. Intuitively, our technique combines the Implementation Invariance of Gradients along with the Sensitivity of techniques like LRP or DeepLift.

Formally, suppose we have a function $F : \mathsf{R}^n \rightarrow [0, 1]$ that represents a deep network. Specifically, let $x \in \mathsf{R}^n$ be the input at hand, and $x' \in \mathsf{R}^n$ be the baseline input. For image networks, the baseline could be the black image, while for text models it could be the zero embedding vector.

We consider the straightline path (in R^n) from the baseline x' to the input x, and compute the gradients at all points along the path. Integrated gradients are obtained by cumulating these gradients. Specifically, integrated gradients are defined as the path intergral of the gradients along the straightline path from the baseline x' to the input x.

The integrated gradient along the i^{th} dimension for an input x and baseline x' is defined as follows. Here, $\frac{\partial F(x)}{\partial x_i}$ is the gradient of $F(x)$ along the i^{th} dimension.

$$\mathsf{IntegratedGrads}_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha$$



how many townships have a population above 50 ? [prediction: NUMERIC]
what is the difference in population between fora and masilo [prediction: NUMERIC]
how many athletes are not ranked ? [prediction: NUMERIC]
what is the total number of points scored ? [prediction: NUMERIC]
which film was before the audacity of democracy ? [prediction: STRING]
which year did she work on the most films ? [prediction: DATETIME]
what year was the last school established ? [prediction: DATETIME]
when did ed sheeran get his first number one of the year ? [prediction: DATETIME]
did charles oakley play more minutes than robert parish ? [prediction: YESNO]

https://arxiv.org/pdf/1703.01365

Feature attributions

- Feature attributions most obviously make sense for interpreting **individual** predictions, and not necessarily the model as a whole
- You can do some obvious things (e.g. average over observations) to try to understand the model as a whole.

Fill out the teammate review form when we send it out!!!