Lecture 10: Explainability & Transparency

So far…

- Principles in human-centred AI
- Privacy, fairness, justice..

- Today, explainability and transparency
  - Make clear what the system can do
  - Make clear why the system did what it did
  - Allow formation of mental models

Question

- Suppose you head an IT team, and suppose you have to manage a software that goes with critical hardware for your organization.
- There are two alternatives:
  - a proprietary software that an OEM built and you are eligible to get free-of-cost
  - an open source equivalent—slightly worse than OEM in performance, but open source
- Which one would you choose? What would your considerations be?

Explainability, transparency, interpretability

- Transparency → Inner workings (code), training data, etc. are open/available for anyone to read, understand, evaluate (not black box)
  - Openness about what, why, how (code, data, training parameters,…)
  - Comprehensibility of #1 (anyone should understand what happened)
  - Forthcoming about #1 and #2 (e.g., inspecting code isn't too hard)

- Important as AI makes crucial decisions
  - Crucial to users, to people responsible for the function AI is helping with, to legal system

Example

- Diagnostic AI → doctors need to know why the AI made the decision it did, if things go wrong insurance/lawyers will need it, informed / curious patients will ask to see details

- Bank loans / insurance rates → customers will ask “why the rate/difference”, underwriters want to know the risk, legalities,..

Interpretability

- Interpretability means that the workings of the model are clear and comprehensible to the people that build / deploy it, so they can fine tune, debug, contribute to it.
  - This is about the entire lifecycle of the model during developing to maintenance

- Example: understanding and debugging a decision tree is far easier than doing so for deep neural networks, or even explaining exactly what each layer does

Explainability

- Explainability simply means the decisions of the AI can be explained / justified to a user (not necessarily tech savvy) in a manner they can understand.
  - This is about the end users of models getting post-hoc explanations

- Example:
  - Diagnostic AI → doctors are users → they should be able to ask “why did the AI say so”, to assess the AI’s reasonableness / accuracy

Explainable AI

- Data explainability
- Model explainability

Data explainability

- Exploratory analysis s
  - Eyeball data, look for outliers, etc. → typically summary statistics & visualizations

- Dimensionality reduction
  - Too many dimensions → reduce to a smaller set
  - Needs to be explained, since some variance in data is lost
  - PCA → clearly explains what directions of variance & how much
  - t-SNE / UMAP → non-linear (diff. transformations in different parts), capture various level of local and global trends (e.g., related categories closer or not), but un-interpretable due to hyperparameters

Example

- https://pair-code.github.io/understanding-umap/

Model explainability

- Explain the model’s predictions after made
- How this is done:
  - Techniques to explain model
  - Also tested (by humans, or automated) during testing phase
  - Via usability studies, too

What goes into model explainability

- Global explanations → training, decision makers, auditors
  - INPUT: What kind of data did you learn from?
  - OUTPUT: What output does it give?
  - PERFORMANCE: How accurate/reliable/precise are the predictions?
  - HOW: How does the system make its predictions
- Local explanations (local to one instance of prediction) → users
  - WHY → why/how did this instance result in this prediction?
  - WHY NOT → why/how did this instance not result in that other prediction?
  - WHAT IF → what would the prediction change to, if I made this change on input?
  - HOW TO BE THAT → what should have happened to be that other case?
  - HOW TO STILL BE THIS → how much variations will still result in this same prediction?
  - Others → drift over time, new training data, etc.

Simple models are easily explainable

- Logistical regressions → show the decision boundary
- Linear regression → show the regression equation
- Decision tree → show the hierarchy of decisions
- More complex models are hard to explain:
  - Neural networks
  - Generative models (LMs, GANS)

NEEDS
INPUT
OUTPUT
PERFORMANCE
WORKING
WHY
WHY NOT
WHAT IF
HOW TO BE THAT
HOW TO REMAIN HERE

Exercise

- How can you meet the user needs on right with:
  - Simple linear regression
  - Simple logistical regression
  - Decision trees

NEEDS
INPUT
OUTPUT
PERFORMANCE
WORKING
WHY
WHY NOT
WHAT IF
HOW TO BE THAT
HOW TO REMAIN HERE

Explainability techniques

- Model-agnostic (LIME, SHAP)
- Model-specific (decision tree interpreters)

LIME: Local Interpretable Model-agnostic Explanations

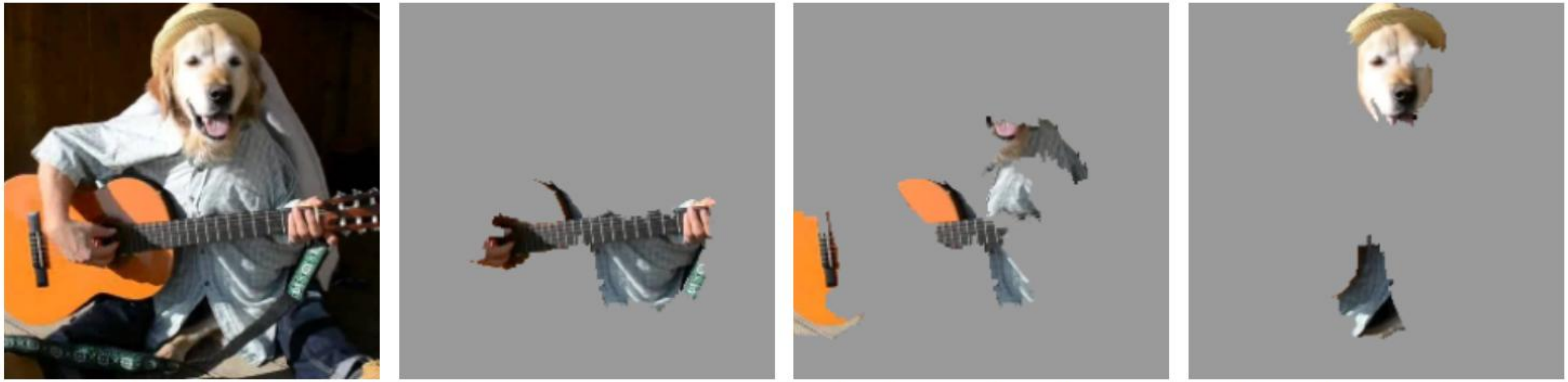
- Local → Why a prediction for this instance
- Model agnostic → works for any (black box) model
- Useful for complicated models

- Given a point & prediction that you want to explain:
  - Create a set of instances of the point, with the input slightly perturbed
  - See what the model's predictions are
  - Train a simple, more explainable model for this small set of points

LIME: Pros and cons

- Limitations:
  - Hard to form mental models since explained “models” are local, and not the real models
  - One time we might say go upto 1.5x, another time 0.8x for a feature
  - Misses nuances, oversimplifies
- Works great in some cases → continuous data, images, text classification

- [Image from Ribeiro et al. , 2016]



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), **"Acoustic guitar"** ($p = 0.24$) **and "Labrador"** ($p = 0.21$)

Assignment

- Build a UI for a user to use the model you had built earlier
- Explain its local predictions
- Feel free to reduce dimensions for convenience but not required. Do a good job, so accuracy is about 80% of original. Simply report overall accuracy for original + reduced dimensionality, if you choose to do it.

Next class

- SHAP
- Presenting explanations
- By then, read:
- "Why Should I Trust You?": Explaining the Predictions of Any Classifier [The original LIME paper by Ribeiro et al.]
  - https://arxiv.org/abs/1602.04938