# Interpretability/Explainability

Machine Learning Theory (MLT)

Edinburgh

Rik Sarkar

# Interpretability/explainability

- Understanding the behaviour of complex models, black box models
  - There are many possible definitions

- How much can a human understand (the cause of) a decision made by a model ?
- How well can a human predict the model's decisions?

- Interpretability and Explainability are often used interchangeably
  - Some authors make a difference

# Why do we need interpretability?

- Because ML models are not perfect
  - We minimize loss on average
    - A particular prediction/decision can still be wrong
  - Sample data may not be perfectly representative
  - Interpretability lets us avoid critical decision errors
  - Interpretability lets us improve our models/algorithms
  - Think of it as debugging ML
- We want to understand the world. Not just blindly use models
  - Interpretability can help us understand nature by uncovering the pattern/relation identified by ML

# Interpretability makes it easier to check other properties

- Fairness
  - Why was the decision made?
- Privacy
- Robustness
- Causality
- Trust

# Interpretability/Explainability are subjective

- “Human understanding” is subjective
  - What is understanding?
  - Understandable to whom?
  - A statistical explanation may sense to a statistician, not to a child
- Many types of approaches, including visualization, causality, human study etc
- We will discuss some ideas suitable for our statistical methods

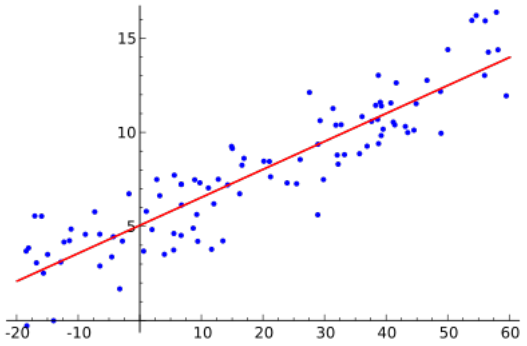
# What may an interpretation method produce?

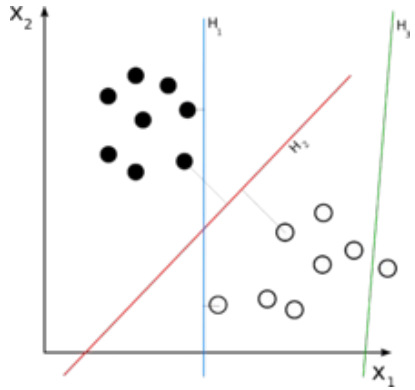
- Feature summary statistic (e.g. how important is a feature?)
- Model details e.g. learned weights
  - E.g. makes sense for linear models, not for NNs
- Data points: Certain predictions may be explained by predictions of nearby data points that differ in some features
  - Counterfactuals
- Approximate interpretable model
  - Approximate a black box model (e.g. NN) by an interpretable model e.g. linear model
  - Either global approximation, or local approximation in the neighborhood of a data point

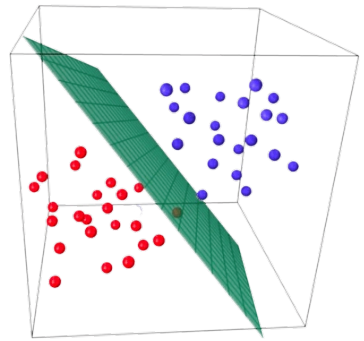
- Global explanations: Explains the whole model (for all possible inputs)
- Local explanations: Explains a particular input/output (and may be its neighborhood)

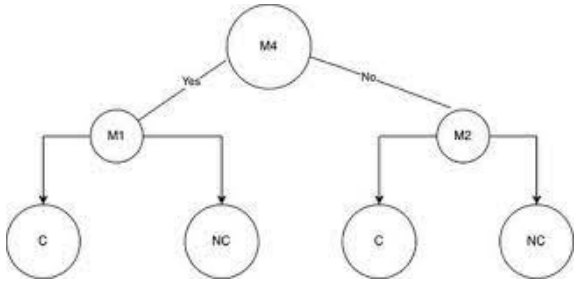
# Interpretable models

- Certain models are naturally interpretable
- Linear regression
- Linear classifiers
  - SVM, Logistic regression, perceptron…
- Decision trees
- k-NN



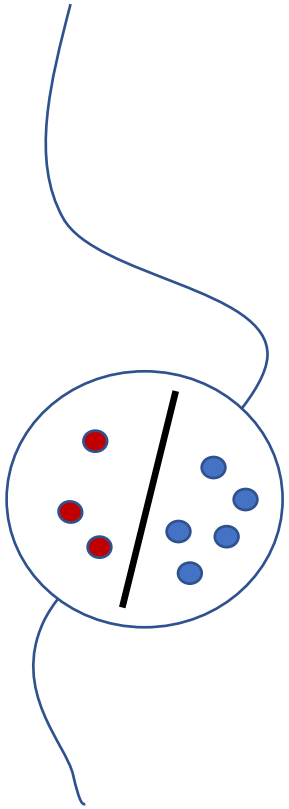


# Local explanations: How does model M treat data point x: which features are important?

- Models are complex
- Models are global
- What if we want to understand the prediction of a single data point
  - Or a group of similar points

# LIME (Local Interpretable model-agnostic Explanations)

- Suppose model $M$ is given
- We want to understand the output $M(x)$ at point $x$.
- LIME:
  - Take random samples $S$ in a small neighborhood of $x$
  - Get their label/predictions from $M$, that is $T = M(S)$
  - Build an interpretable model $M'$ using $(S, T)$



# Permutation feature importance: How important is a feature?

- Suppose test data $Z$ is given by a matrix, with columns as features
  - Where $Z_{i,j}$ is the value of feature $j$ for data point $i$
- First build the model, find its test error: $e$
- For each feature $j$
  - Take the $\mathrm{column}\ Z_{*,j}$ : all values of feature $j$
  - Permute the values randomly
  - Find test error $e_p$
  - Importance of $j$: $e_p - e$

# Permutation importance

- Computes the importance of a feature by nullifying its values by randomizing, and then observing how much it was contributing

- Can be applied to either test data or training data
  - Test data: Permute feature values in test data and compute $e_p$
  - Training data: permute feature values in training data and train new model
    - Then test on unmodified test data

- Suppose features $j$ and $j + 1$ are tightly correlated (say, exactly the same)
  - What are the error differences $e_p - e$ ?

# Feature importance

- The problem: Features may have complex interrelations
  - $x_1$ and $x_2$ may be correlated
  - Importance of $x_1$ may depend on the presence of $x_2$
- Approach: Shapley values from cooperative game theory
  - Addresses interrelated factors

# A simple idea: marginal contributions

- If the set $S = \{x_1, x_2, x_3, \dots\}$ can collaborate to obtain a value of $f(S)$

- The marginal contribution of $x_i$ is:
  - $f(S) - f(S \setminus \{x_i\})$

# A simple idea: marginal contributions

- If the set $S = \{x_1, x_2, x_3, \dots\}$ can collaborate to obtain a value of $v(S)$
- The marginal contribution of $x_i$ is:
  - $v\,(S) - v(S \setminus \{x_i\})$

- What are marginal contributions of $x_1, x_2$ in these cases:
  - $x_1$ and $x_2$ can collaborate on a project that has value $v$ when both are present and $0$ when any one is absent.
  - $x_1$ or $x_2$ could either (alone) complete the project of value $v$

# Fair division

- Consider $n$ players collaborating on a project
- The cost or profit should be shared in a “fair” manner

- Example: A, B and C share a taxi
  - A gets off first, when meter reads £12
  - B gets off next, when meter reads £14
  - C gets off when meter reads £16

  - How should they share the costs?

# Shapley value

- Suppose $v$ is the value function and $N$ is the set of possible players and $n = |N|$ (we say a game is defined by $(v, N)$)
- The Shapley value for player $i$ is:
  - $\varphi_i(v) = \frac{1}{n} \sum_{S \subseteq N \setminus \{i\}} \binom{n-1}{|S|}^{-1} (v(S \cup \{i\}) - v(S))$

- Take marginal contributions of $i$ to all possible subsets
  - Average them

$$\varphi_i(v) = \frac{1}{\text{number of players}} \sum_{\text{coalitions excluding } i} \frac{\text{marginal contribution of } i \text{ to coalition}}{\text{number of coalitions excluding } i \text{ of this size}}$$

- Check what happens in the cases on the previous slides

# Properties of Shapley value

- **Efficiency**: $\sum_{i \in N} \varphi_i(v) = v(N)$
  - Shapley values add up to the total value
- **Symmetry**: if $v(S \cup \{i\}) = v(S \cup \{j\})$ for every $S$ which contains neither $i$ nor $j$, then $\varphi_i(v) = \varphi_j(v)$
- **Linearity**: For games with gain functions $v$ and $w$:
  - $\varphi_i(v + w) = \varphi_i(v) + \varphi_i(w)$
- **Null player**: if $v(S \cup \{i\}) = v(S)$ for all $S$, then $\varphi_i(v) = 0$

- Shapley value is the only map from games to payoff vectors that satisfies all four axioms
- Applies for any kind of $v$ as long as it is defined for all subsets S

# Alternative (equivalent) definition of Shapley value

- Suppose $v$ is the value function and $N$ is the set of possible players and $n = |N|$ (we say a game is defined by $(v, N)$)
- Suppose $\Pi(N)$ is all possible permutations of players
  - $\pi \in \Pi(N)$ is a permutation.
  - $\pi(i)$ is position of player $i$ in $\pi$
- Suppose $\mathcal{P}_i^{\pi} = \{j \in N : \pi(j) < \pi(i)\}$ is the set of players before $i$ in $\pi$
- Then shapley value is
  - $\varphi_i(v) = \frac{1}{|\Pi(N)|} \sum_{\pi \in \Pi(N)} [v(\mathcal{P}_i^{\pi} \cup \{i\}) - v(\mathcal{P}_i^{\pi})]$
- Take all possible permutations. Compute average marginal contribution

# Computational Complexity

- In either definition, SV computation is inefficient

- All possible subsets of n elements numbers in $2^n$
- All permutations is $n!$

# Shapley value approximation

- Idea:

- Take $k$ random samples of permutations
- Compute marginal contribution of each $i$ in each permutation

- Take average

# Approximation algorithm

**Data:** $(\mathcal{N}, v)$ - Cooperative TU game.
$k$ - Number of sampled permutations.
**Result:** $\widehat{\phi}_i^{Sh}$ - Approximated Shapley value $\forall i \in \mathcal{N}$.

$\widehat{\phi}_i^{Sh} \leftarrow 0, \quad \forall i \in \mathcal{N}$
**for** $(1, \ldots, k)$ **do**
  $\pi \leftarrow$ Uniform Sample$(\Pi(\mathcal{N}))$
  **for** $i \in \mathcal{N}$ **do**
    $\mathcal{P}_i^{\pi} \leftarrow \{j \in \mathcal{N} \mid \pi(j) < \pi(i)\}$
    $\widehat{\phi}_i^{Sh} \leftarrow \widehat{\phi}_i^{Sh} + \frac{v(\mathcal{P}_i^{\pi} \cup \{i\}) - v(\mathcal{P}_i^{\pi})}{k}$
  **end**
**end**

**Algorithm 1:** Monte Carlo permutation sampling approximation of the Shapley value.

# Observation

- This is like other learning or approximation tasks.
- We are sampling permutations and recording value of $i$ in each

- General idea:
  - Whatever interrelations exist between players, are likely to be captured by some of the permutations if we take enough samples.

# Class exercise:

- In a company, employees work in small groups
- Suppose $i, j$ collaborate on a project of value 1 only when both are present, so each has value ½ (i.e. $\varphi_i = \frac{1}{2}$)
- Observe that in a random permutation, both are equally likely to occur first
  - The one occurring first has marginal value 0
  - The one occurring second has marginal value 1
- How many permutation samples (among n players) do we need to get an $\epsilon$ approximation for the value of $\varphi_i$ with probability $1 - \delta$?

- Hint: use Hoeffding's inequality:
  - $\mathbb{P}\left[\left|\frac{1}{k}\sum_{i=1}^{k}\theta_i - \mu\right| > \epsilon\right] \leq 2e^{-2k\epsilon^2}$

# Theoretical properties

- The estimate $\hat{\phi}_i^{Sh}$ is an unbiased estimator of true Shapley value
- With increasing sample size $k$, it converges to true value

- When each marginal contribution is bounded by $r$
- The number of permutation samples needed for $\epsilon, \delta$ accuracy is
  - $k \geq \frac{\ln(\frac{2}{\delta})r^2}{2\epsilon^2}$

# Local explanations – explaining point prediction

- Suppose a model $h$ is given
- We want to evaluate the importance of feature $i$ on a prediction $h(x)$
- Treat features of $x$ as players
- Pass different subsets of features of $x$ to $h$
  - Compute marginal differences in loss and thus Shapley value
- Note:
  - “Removing” features is not trivial
  - Usually achieved by setting to zero, or average of feature

# Global feature importances

- Method 1: Sum local valuations over all data points to get aggregate feature importances
- Method 2: Train models with feature subsets and compute their empirical or test losses
  - Use marginal gains to get Shapley value
- [optional reading: related paper: SHAP: Lundberg 2017].

# Other applications of shapley explanations

- Data valuation: Which training data points (or subsets) are valuable?
  - Train on different subsets with/without the point (or subset) to compute Shapley values
- Which neurons in a neural network are important?
  - ….
- Various other applications
  - [(optional reading) Recent survey: Rozemberczki  et al. 2022]

# A note on ML, ethics and dangers

- A lot of current debate on sentient AI taking control of everything
  - Not clear that will happen or can happen

- But dangers of AI/ML are not far away
- We have already seen problems in
  - Privacy
  - Fairness
- Our data is getting stolen and used right now!
- Models are making decisions
  - Not clear why
  - Not clear how fair
- Laymen, and powerful people trust AI without doubt
- Other powerful people tell us to trust AI without doubt...

# A note on ML, ethics and dangers

- AI is already dangerous
  - Yes, it is a useful technology, but still dangerous!

- Implications on privacy, fairness are not clear
- Safety is far from guaranteed.
  - See: self driving cars
- Implications of everyone using simple current AI is not clear!

- Our society is built on humans making human like decisions (both right and wrong)
- What happens when models make all those decisions – different right and wrong
  - E.g. On youtube/Netflix models determine what content is made, propagated
    - Changes our culture, beliefs, intuitions…
  - What happens when *all* financial decisions, career decisions, communication are made completely differently?