Week 7 Scribe Notes

Uncertainty Bias

- **What is it?**
  - One group is under-represented in the sample, therefore there is more uncertainty associated with predictions about that group
- The algorithm is risk averse, so with other conditions remaining the same, it will prefer to make decisions based on predictions about which they are more confident

GDPR (General Data Protection Rule)

- **What is it?**- Legal framework that sets guidelines for processing and collection of information from individuals living in the European Union
  - Went into effect in April 2018, replacing previous 1995 EU Data Protection Directive
  - Carries large penalties (20M euros or 4% global revenue) for violation as it is law, rather than a directive.
- **Article 22** – Prohibits any automated decision making unless conditions are met
  - Claims algorithmic profiling is discriminatory
  - **Paragraph 71 –** specifically addresses discrimination from profiling that makes use of sensitive data
- **Article 13 -** Data subjects have a right to access information collected about them
- **Article 14** - Subjects should be notified of data being collected

Interpretations of GDPR Article 22 Paragraph 71

- **Two interpretations -** There exist two ways of viewing this section of GDPR and how it pertains to identifying/discriminatory features
- **Minimal Interpretation –** Just get rid of protected features
  - Pros: Simple to implement
  - Cons: Does not address correlations between protected and unprotected features
- **Maximal Interpretation** - Get rid of all features that might be correlated with protected features
  - Pros: You have a legal model
  - Cons: You might lose predictive accuracy
  - Cons: You might run into the problem of uncertainty bias
- **Uncertainty bias**
  - One group is underrepresented in the sample, so there is more uncertainty associated with predictions about that group
  - Algorithm is risk-averse, so will prefer to make decisions from higher confidence decisions

Critical Analysis of Post-Hoc Interpretable ML Models

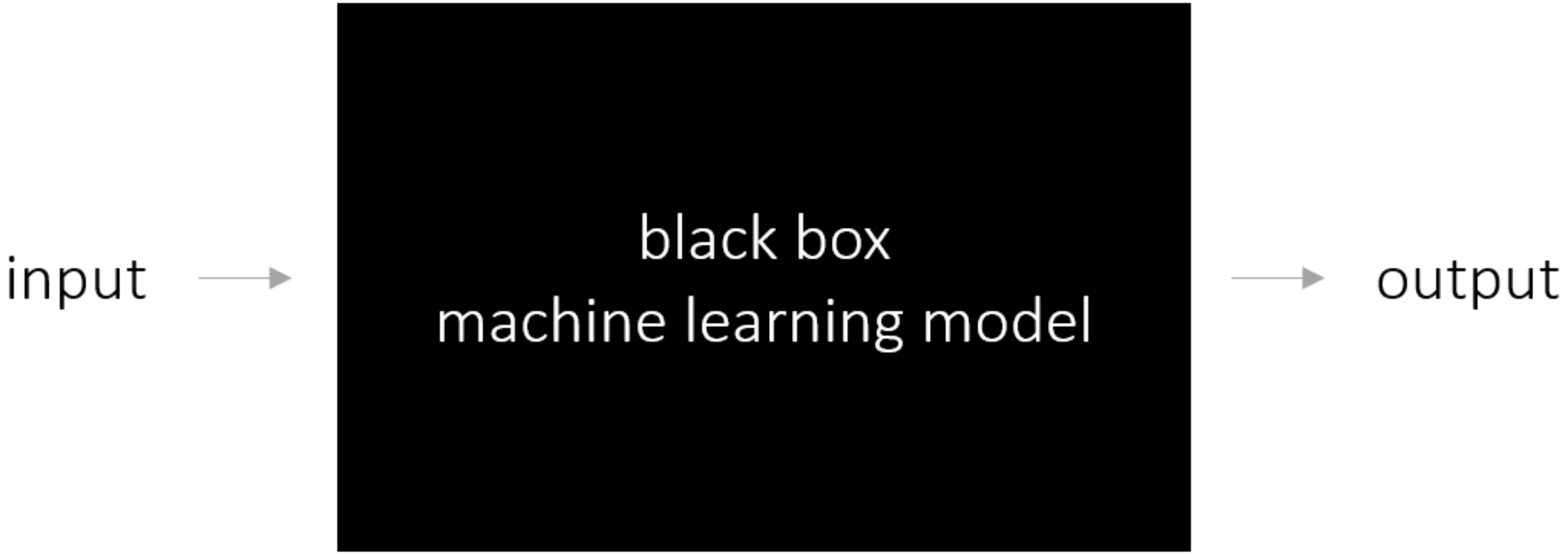
- **Post-hoc general characteristics**
  - Treat ML model as a black box
  - Design a system to explain what the black box learned/decided
- **Partial Dependence Plot (PDP)**
  - Pick a feature, vary the feature values across the data set
  - Observe how predictive accuracy changes as the values are varied
  - Pros: Easy to implement; intuitive
  - Cons: Assumption of independence; does not show changes at individual points in the data set
- **Individual Conditional Expectation (ICE)**
  - Instead of looking at the data set average, this looks at individual data points
  - Each line in the model's graph measures how changing a feature value affects a single data point
  - Pros: Easy to implement; intuitive
  - Cons: Assumption of independence

Critical Analysis of Post-Hoc Interpretable ML Models (cont.)

- **Permutated Feature Importance**
  - Shows the "feature importance" measuring how a single feature can affect the dataset
  - Pros:
    - Provides global insight
    - Comparable across platforms
  - Cons:
    - Assumption of independence
- **Global Surrogate**
  - Trains black box model to make a prediction
  - Train an interpretable model on the predictions from the black box model
  - Pros:
    - Any interpretable model can be used
    - Decision Tree, Linear model, etc
  - Cons:
    - Can only explain the model, not the data
    - Surrogates might not be accurate

Critical Analysis of Post-Hoc Interpretable ML Models (cont.)

- **Local Surrogate**
  - A model which mimics decision of a complicated model for one specific region of data
  - Will not be effective for all regions of the data, only one
  - **LIME**
    - Generates a dataset of perturbed instances by turning some of the interpretable components "off"
      - Perturbed instances are close to the original image, so they lie in the vicinity of the original data point
    - For each perturbed instance, one can use the complicated un-interpretable trained model to get the probability that a tree frog is in the model
  - Pros:
    - Model-agnostic
    - Human-friendly explanations
  - Cons:
    - Need to define a kernel to define the area in which data points are considered
      - Often open problem, no good solution



LIME MODEL

Critical Analysis of Post-Hoc Interpretable ML Models (cont.)

- **SHAP (Shapley Value)**
  - Prediction can be explained by assuming that each feature value of the instance is a player in a game
  - Differs from surrogates because it is not learning a new ML model
  - Contribution from each player is measured by adding and removing the player from all subsets of the rest of the players
  - The Shapley value for one player is the weighted sum of all its contributions
  - Current state-of-the-art, should be used today

  - **Cons:**
    - Very complex, computationally expensive

Intrinsically Interpretable Models

Decision Trees

- **Pros**
  - Graphical Structure
  - Work with a subset of attributes
  - Hierarchical representation dictates the order features are considered in
- **Cons**
  - Irrelevant attribute splits
  - Data Fragmentation due to irrelevant splits

Classification Rules

- *List of if-then-else logical statements**
- **Pros**
  - More interpretable
  - Easier to understand modular pieces
  - Enable aggressive pruning/limit number of conditions that are present inside these rules
    - Irrelevant splits can be removed
- **Cons:**
  - No ordering of clauses, how to figure out which ones are more important?
  - How to measure feature importance? Average number of rule conditions tested

Intrinsically Interpretable Models

Decision Tables

- *Tabular representation of data**
- **Pros**
  - Fairly interpretable
  - Legitimately a table
- **Cons**
  - Size of these tables can grow prohibitively long

Nearest-Neighbors

- *For any data point, look at nearby data points, whichever data point has majority, then the selected data point will match**
- **Pros**
  - Can be improved using **prototype points**
    - *Single points representing one class category*
- **Cons:**
  - Explanation is different for every data point
    - **Each data point** has a **different set of nearest neighbors**

Bayesian Network Classifiers

- **How to improve?**
  - Don't use model size as a single measure of interpretability
  - Introduce semantic monotonicity constraints in model building
    - Probability of someone having cancer is going to go up as the age of a person goes up