Interpretable Machine Learning

Nan Ye

School of Mathematics and Physics
The University of Queensland

Where Are We Heading to?

How to build good ML models

- Making use of a crowd ⇒ Week 7 Ensemble methods

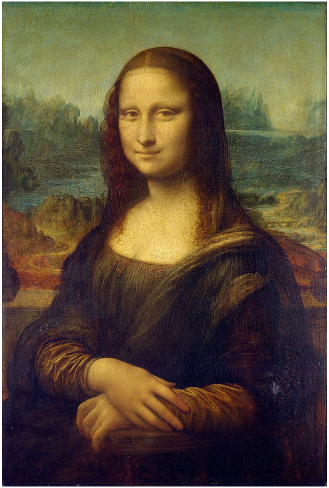
  each of us is a biological prediction model trained on different datasets...
- Using a neural network ⇒ Week 8 and 9 Neural networks

  brain-inspired models, some are good for images...
- Making a robust model ⇒ Week 10 Robust machine learning

  malicious users, outliers,...
- Asking for explanations ⇒ Week 11 Interpretable machine learning

  ...let's ask the machines for explanations...
- Exploiting prior beliefs ⇒ Week 12 Bayesian methods

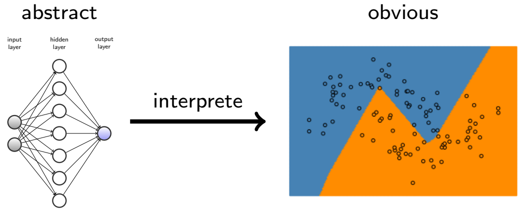
Interpretability



Everybody has his own interpretation of a ~~machine learning algorithm~~ painting he sees...
Francis Bacon

What is an interpretation?

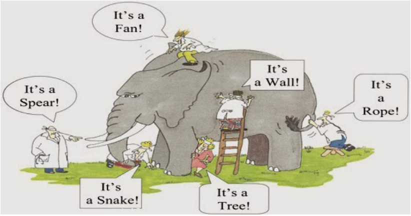
- An interpretation connects the abstract/unfamiliar to the obvious/familiar.
- Not new in this course.



- Many other ways of interpreting machine learning algorithms have been created — we are good at coming up with interpretations.
- Interpretations help come up with explanations for the predictions.

Interpretation = Misinterpretation?

- Each interpretation often tells us part of the truth, and we may need to use several methods to form a more complete picture.



- An algorithm designed to generate helpful interpretations may produce misinterpretations — understanding how it works helps us to avoid misinterpreting its output.

Interpretable Machine Learning

- We want to find intuitive descriptions for
  - the functional relationship represented by a model
  - each component of a model
  - effect of each input variable
  - ...
- Intuitive: visualizations, numerical summary, simple rules, ...
- Interpretations sometimes help explaining why a model makes a prediction.
- We discuss some interpretation methods and how they can be applied in this lecture.

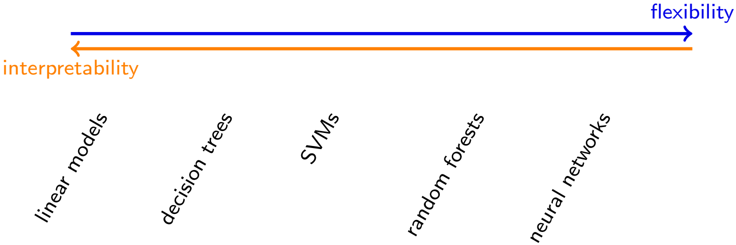
Approaches

- Various approaches have been taken to make machine learning models interpretable, and they can be categorized in various ways.
- Built-in vs post hoc
  - Built-in: models are designed to be interpretable (e.g. linear regression)
  - Post hoc: models are analyzed for interpretability (e.g. permutation importance)
- White-box vs black-box
  - White-box: everything about the model is needed (e.g. linear regression model weights)
  - Black-box: only partial information about the model is needed (e.g. permutation importance)

- Model specific vs model agnostic
  - Model-specific: designed for specific models only (e.g. linear regression model weights)
  - Model-agnostic: designed for generic learning approaches (e.g. permutation importance)
- We will cover some basic methods
  - Interpretable models: linear regression, logistic regression, decision trees
  - Surrogate model method
  - Variable importance: Gini importance, permutation importance
  - Low dimensional approximation

Interpretable Models

- More flexible/complex models often have better performance, but typically harder to interpret



- For a long time, interpretable models like linear models are strongly preferred.

Interpreting Linear Models

- We have seen a number of linear models in this course
  - Linear regression, logistic regression, SVM with linear kernels
- Linear models are simple and their parameters often have easily interpretable meanings

Linear regression

- A linear regression model $f_{\mathbf{w}}(\mathbf{x})$ has the form

$$f_{\mathbf{w}}(\mathbf{x}) = w_0 + \sum_{i=1}^{d} w_i x_i,$$

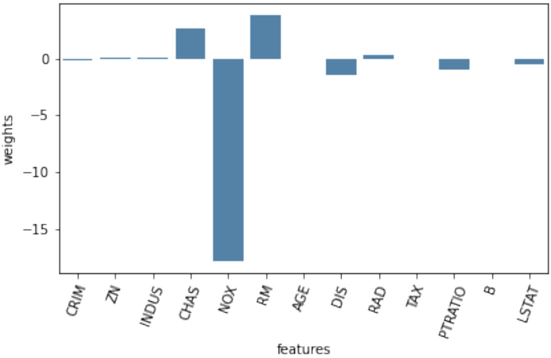
  where $\mathbf{w} = (w_0, w_1, \ldots, w_d)$, and $\mathbf{x} = (x_1, \ldots, x_d)$.
- Interpretation of the parameters
  - bias w_0: output when all features are 0
  - weight w_i: change in the output when x_i increases by one unit

- Boston house prices again: predict median house price in a using 13 features

```
- CRIM     per capita crime rate by town
- ZN       proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS    proportion of non-retail business acres per town
- CHAS     Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- NOX      nitric oxides concentration (parts per 10 million)
- RM       average number of rooms per dwelling
- AGE      proportion of owner-occupied units built prior to 1940
- DIS      weighted distances to five Boston employment centres
- RAD      index of accessibility to radial highways
- TAX      full-value property-tax rate per $10,000
- PTRATIO  pupil-teacher ratio by town
- B        1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
- LSTAT    % lower status of the population
- MEDV     Median value of owner-occupied homes in $1000's
```

- Which features are important to you?

- Generally, weights $\neq$ importance



- Weights of normalized features are much better indicators of feature importances

  - Normalization: scale each feature so that it has unit variance, then train a linear regression model
  - Magnitude measures importance
  - Sign reflects whether it has a positive or negative effect

- Weight of normalized feature $=$ weight of unnormalized feature $\times$ feature standard deviation
- Thus weight of normalized feature is a kind of *normalized weight*

(Binary) Logistic regression

- A binary logistic regression model defines a conditional class distribution

$$p(y \mid \mathbf{x}, \mathbf{w}) = \frac{1}{1 + e^{-y(w_0 + \sum_{i=1}^{d} w_i x_i)}},$$

where $\mathbf{w} = (w_0, w_1, \ldots, w_d)$, $\mathbf{x} = (x_1, \ldots, x_d)$, and $y \in \{0, 1\}$.

- Equivalently, the log-odds is linear

$$\ln \frac{p(y = 1 \mid \mathbf{x}, \mathbf{w})}{p(y = 0 \mid \mathbf{x}, \mathbf{w})} = w_0 + \sum_{i=1}^{d} w_i x_i.$$

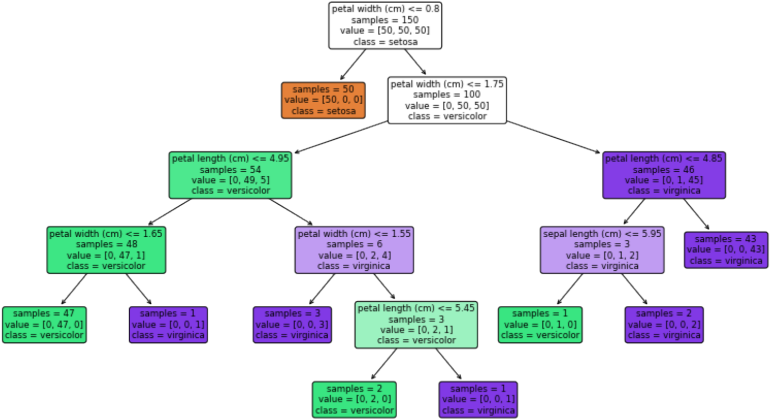
- Interpretation of the parameters
  - bias w_0: log-odds when all features are 0
  - weight w_i: change in the log-odds when x_i increases by one unit

- As in linear regression, the weights generally do not measure the importance of the features.
- However, the weights of the normalized features are much better indicators of feature importances.

Interpreting Decision Trees

- Decision trees can be converted into a set of rules
  - Each rule correspond to each path from the root to a leaf
- The rules can often be simplified (e.g. test conditions on the same feature can often be combined).

- A decision tree for iris data



- E.g.: petal width $\leq$ 0.8 cm $\Rightarrow$ setosa (in fact, we can replace $\Rightarrow$ by iff)
- E.g.: petal width $\in$ (0.8 cm, 1.65 cm) and petal length $\leq$ 4.95 cm $\Rightarrow$ versicolor (simplified from the path from root to the left-most green leaf)

Interpreting Complex Models

- For complex models, it is often hard (if not impossible) to interpret what they are doing by examining their internals.
- We can interpret them by querying their input-output relationships to find
  - surrogate interpretable models
  - variable importance scores

Surrogate Model Method

- A surrogate model M' for a given model M is one trained to fit the predictions of M on a dataset.
- The dataset is chosen depending on what you consider interesting
  - training set, test set, or a subset of them
  - a grid of points
- The surrogate model is chosen to be an interpretable model
- An interpretable model is simpler, and thus such a surrogate model inevitably oversimplies the original model.

Variable Importance

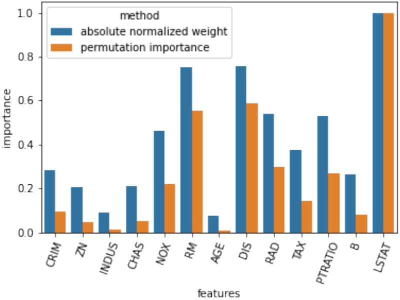
- We often have various ways to measure how much importance a model assigns to a variable
  - each metric only looks at a specific aspect of the model
  - sometimes the metrics may present conflicting pictures
- Many variable importance scores are model-specific (e.g. the normalized weights for a linear regression model).
- The permutation importance is a model-agnostic importance score.

Permutation importance

- Given: a dataset D, a model M, a performance score
- Computing the permutation importance of a variable in M
  - calculate the score s of M on D
  - create multiple *permuted* datasets $D_1, \ldots, D_k$
    - each is the same as D except that the values of the variable for all the instances are permuted
  - calculate the score s_i of M on D_i for each i
  - calculate the mean $\bar{s}$ and standard deviation σ of $s_1, \ldots, s_n$.
  - permutation importance: mean $= s - \bar{s}$, std $= \sigma$.
- The permutation importance is a random number.

Variable Importance for Linear Regression

- Comparison of two variable importance scores on the Boston house price dataset



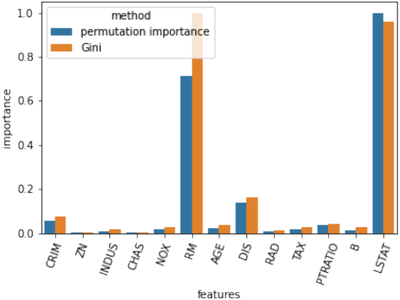
  - To make the two importance scores comparable, the scores are scaled such that the maximum is 1.
  - While the two scores are generally different, they rank the features similarly.

Variable Importance for Random Forest

Gini importance

- Random forest has a model-specific variable importance score known as the Gini importance.
- The Gini importance of a variable is the total weighted decrease in node impurities from splitting on the variable, averaged over all trees.
  - The weight is the number of examples involved in impurity decrease.
  - For classification, node impurity is the Gini index.
  - For regression, node impurity is the residual sum of squares.
- In an implementation, Gini importances may be normalized so that the sum of the importances of all variables sum to 1.

- Comparison of two variable importance scores on the Boston house price dataset

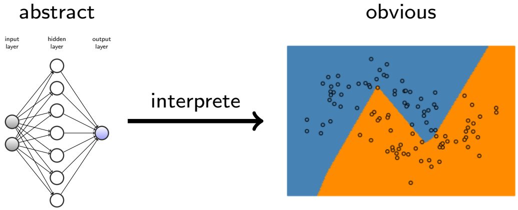


  - To make the two importance scores comparable, the scores are scaled such that the maximum is 1.
  - The two scores are somewhat similar, but they differ significant for the top 2 features.

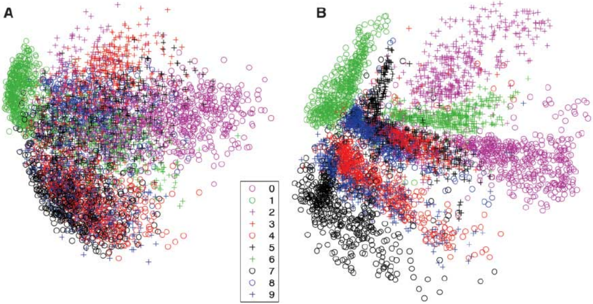
- Random forest and linear regression assign very different importance scores to the features.
- The importance scores are not the intrinsic importance scores of the features, but measures how important the models consider the features to be.

Low-dimensional Approximation

- For low dimensional data, we can directly visualize the functional relationship represented by a model



- For high dimensional data, we can approximately visualize the funtional relationship by performing dimension reduction first.
  - PCA is one way to do dimension reduction, but there are other ways, such as t-SNE, autoencoders (not covered in this course)



PCA and autoencoder codes for MNIST

Hinton and Salakhutdinov, Reducing the dimensionality of data with neural networks, 2006

Checking Your Understanding

Which of the following statement is correct? (Multiple choice)

(a) A more flexible model is generally easier to interpret.

(b) If two features are identical, then they always get the same importance scores.

(c) A post hoc interpretation method can only be applied to a black-box model.

What You Need to Know

- An interpretation
  - connects the abstract/unfamiliar to the obvious/familiar.
  - often tells part of the truth
- Approaches to make machine learning models interpretable
  - Built-in vs post hoc, white-box vs black-box, model-specific vs model-agnostic
- Some basic methods
  - interpretable models, surrogate method, variable importance, low dimensional approximation