https://ttpoll.com/p/744372

COSC 325: Introduction to Machine Learning

Dr. Hector Santos-Villalobos

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

Lecture 18: Explainability

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

Class Announcements

Homework
Homework #5 due 11/06
Homework #6 due 11/13

Course Project:
Course Project Presentation due 11/26
- ***Option #1: Youtube 10-min video and 3-min in-class presentation***
- ***Option #2: in-class poster/demo***

Lectures:
No class on Tuesday 11/05 (Election Day)
Expect code walkthrough videos by Tuesday
11/25 Lecture: No attendance record. Thanksgiving week.

Quizzes:
Weekly quiz as usual.

Exams:
Next exam 11/21. Same format.

Pop Quiz

What knowledge dissemination mechanism do you prefer for the course project final report?

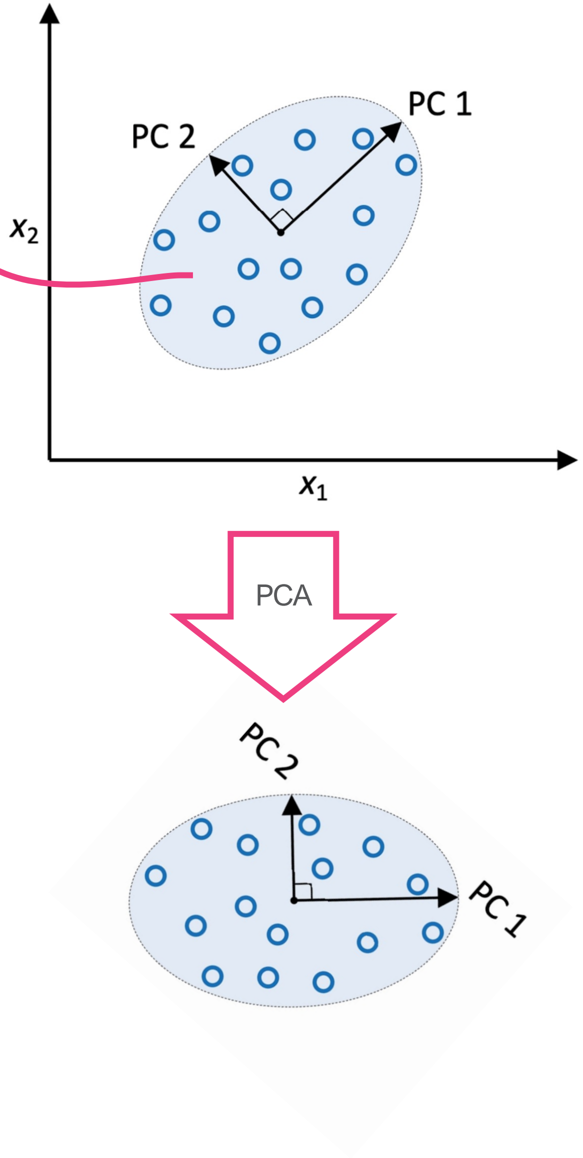
A. 10-minute YouTube video and 3-minute in-class presentation

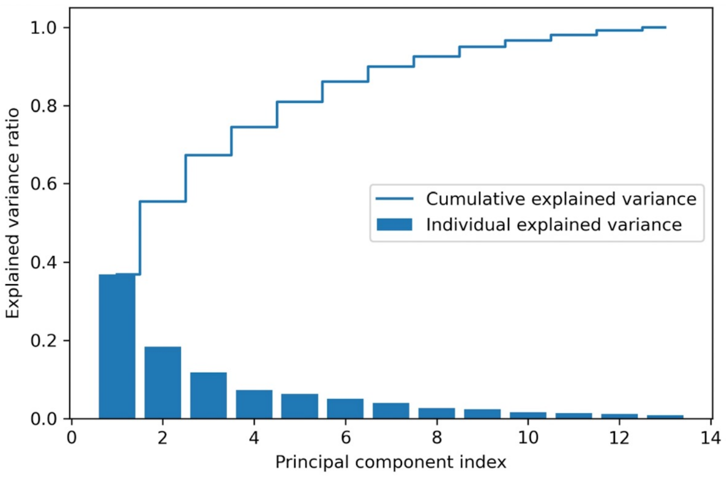
B. In-class poster [With optional demo]

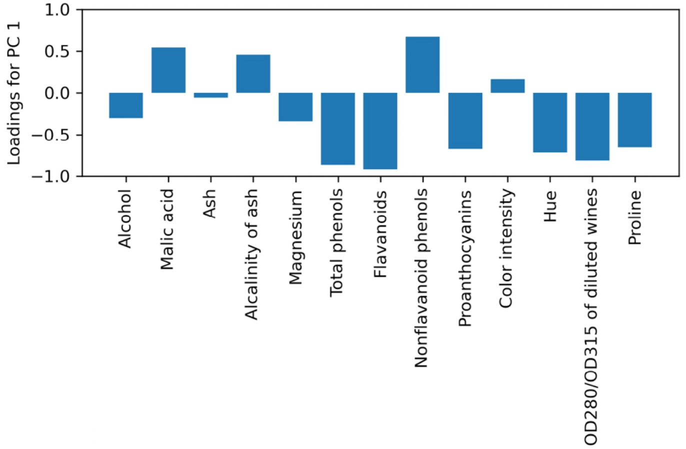
Review

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}$$

- Feature Extraction
  - Principal Component Analysis (PCA)
    - Unsupervised technique
    - Compute features covariance matrix
    - The eigenvectors (or principal components) project samples into a lower dimensional space and point toward the data's largest variance.
    - Explainable Variance Ratio: Normalize and sort eigenvalues to measure the contribution of the PCs on the data variance.
    - Loadings: Correlation between original features and principal components, indicating the features contributing more to the data variance.
  - Linear Discriminant Analysis (LDA)
  - t-distributed stochastic neighbor embedding (t-SNE)

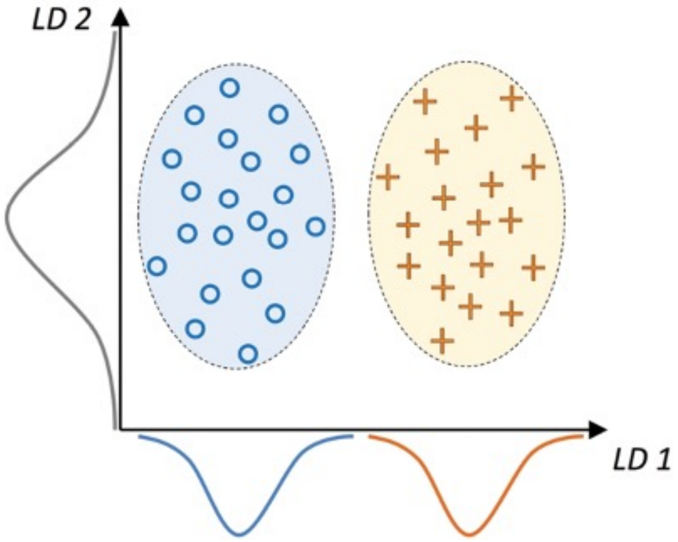






Review

- Feature Extraction
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
    - Supervised technique
    - Compute classes scatter matrices
    - The eigenvectors (linear discriminants) project samples into a lower-dimensional space and point toward the data direction that maximizes class discrimination.
  - t-distributed stochastic neighbor embedding (t-SNE)

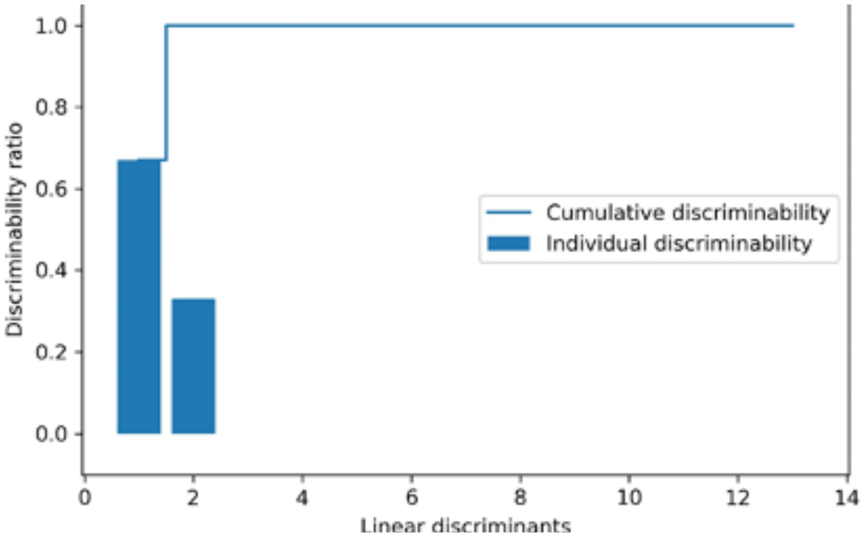


$S_W = \sum_{i=1}^{C} S_i$, where

$$S_i = \sum_{x \in D_i} (x - \bar{v}_i)(x - \bar{v}_i)^T$$

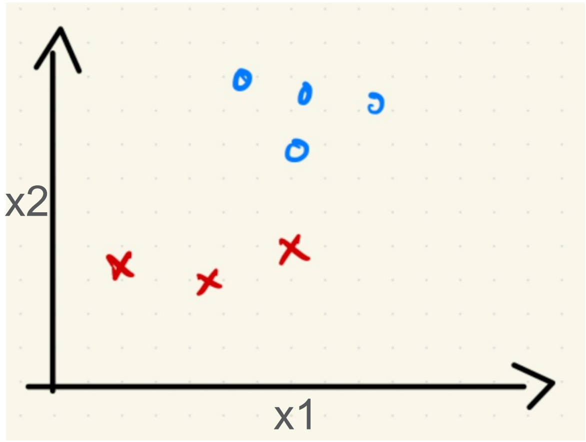
$$S_B = \sum_{i=1}^{C} n_i (\bar{v}_i - \bar{v})(\bar{v}_i - \bar{v})^T$$

Find LDs in $S_W^{-1} S_B$



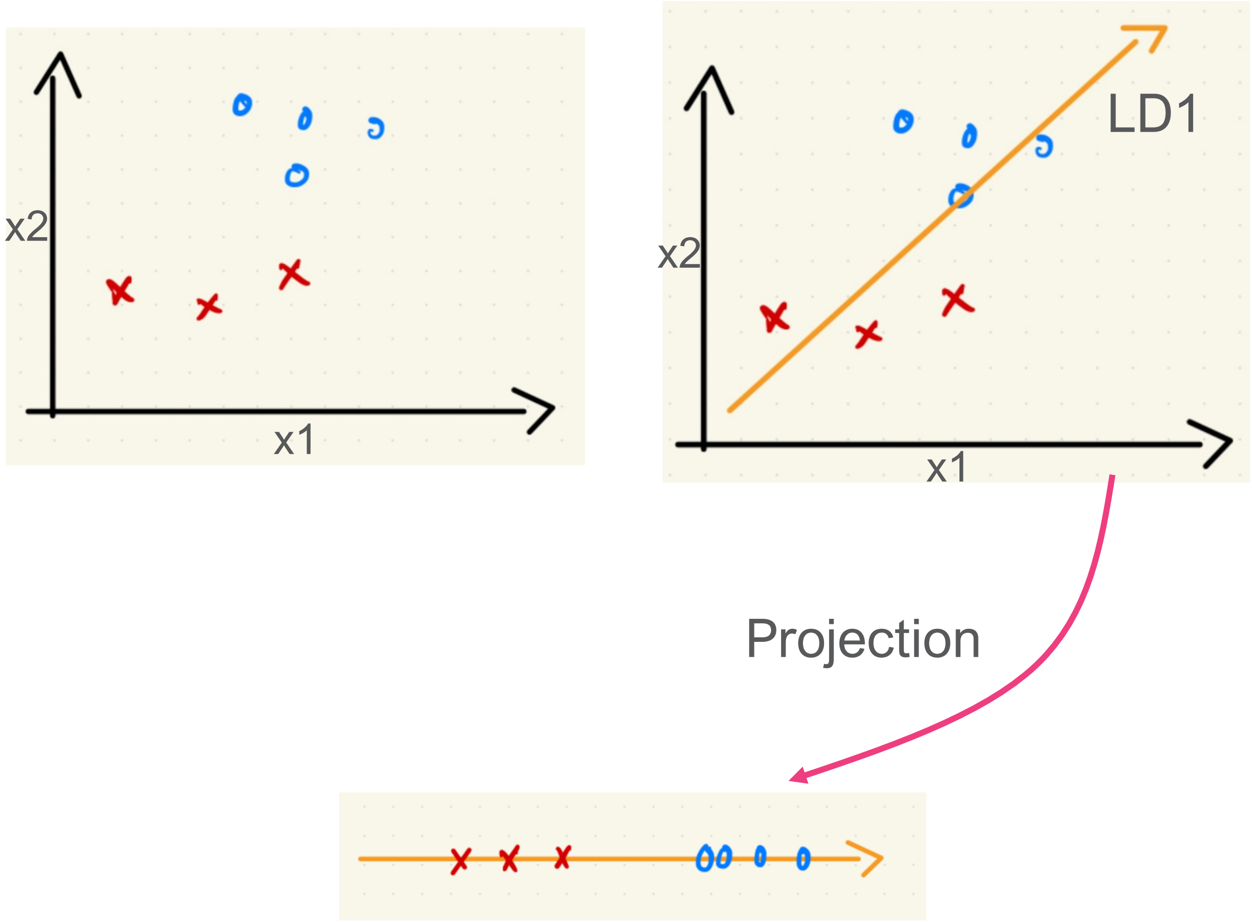
Review

- Feature Extraction
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
    - Supervised technique
    - Compute classes scatter matrices
    - The eigenvectors (linear discriminants) project samples into a lower-dimensional space and point toward the data direction that maximizes class discrimination.
  - t-distributed stochastic neighbor embedding (t-SNE)



Review

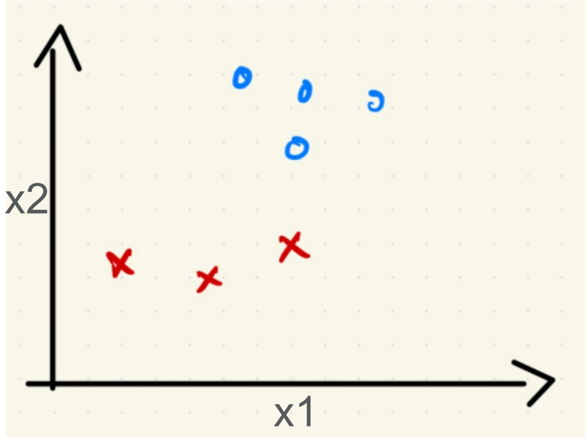
- Feature Extraction
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
    - Supervised technique
    - Compute classes scatter matrices
    - The eigenvectors (linear discriminants) project samples into a lower-dimensional space and point toward the data direction that maximizes class discrimination.
  - t-distributed stochastic neighbor embedding (t-SNE)

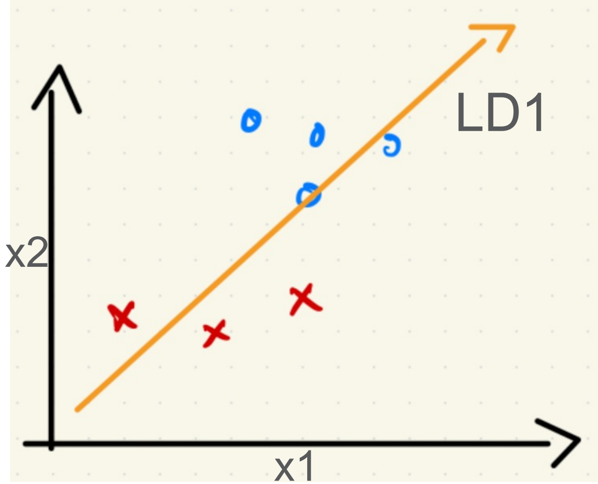


https://ttpoll.com/p/744372

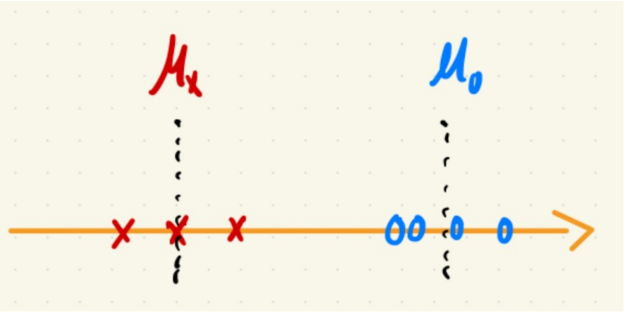
Review

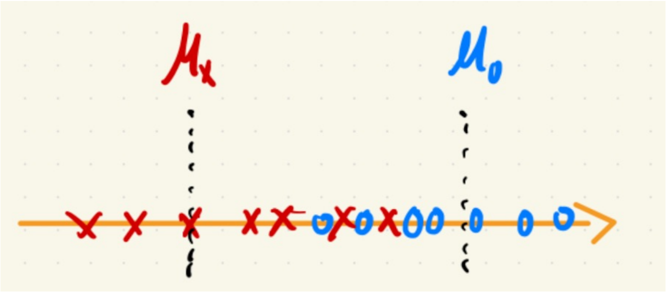
- Feature Extraction
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
    - Supervised technique
    - Compute classes scatter matrices
    - The eigenvectors (linear discriminants) project samples into a lower-dimensional space and point toward the data direction that maximizes class discrimination.
  - t-distributed stochastic neighbor embedding (t-SNE)





If we optimize only on the mean

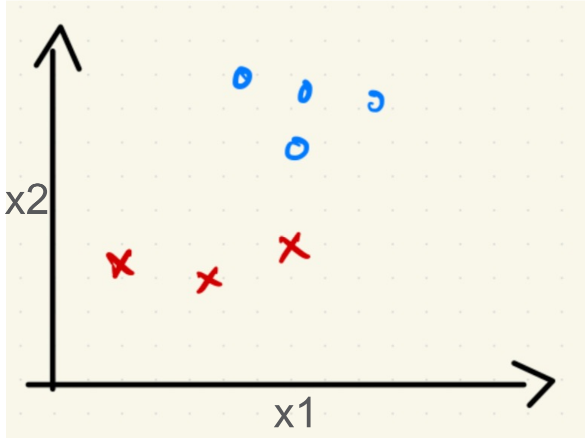


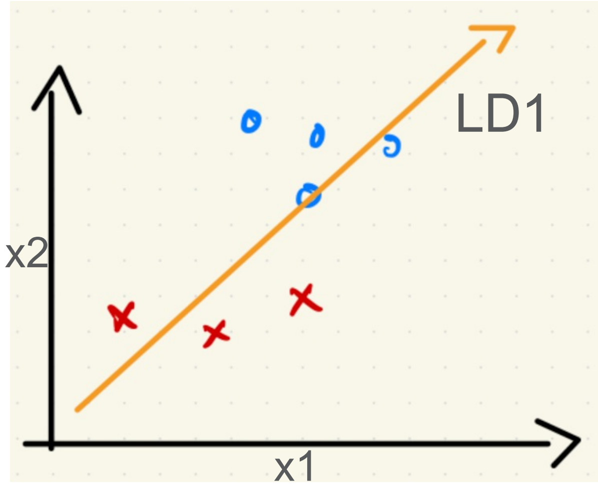


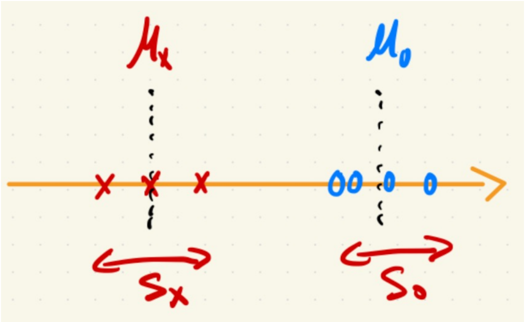
The spread of data may create an overlapping between classes.

Review

- Feature Extraction
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
    - Supervised technique
    - Compute classes scatter matrices
    - The eigenvectors (linear discriminants) project samples into a lower-dimensional space and point toward the data direction that maximizes class discrimination.
  - t-distributed stochastic neighbor embedding (t-SNE)





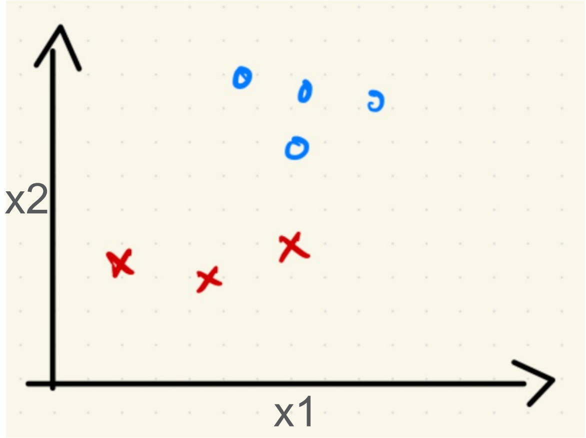


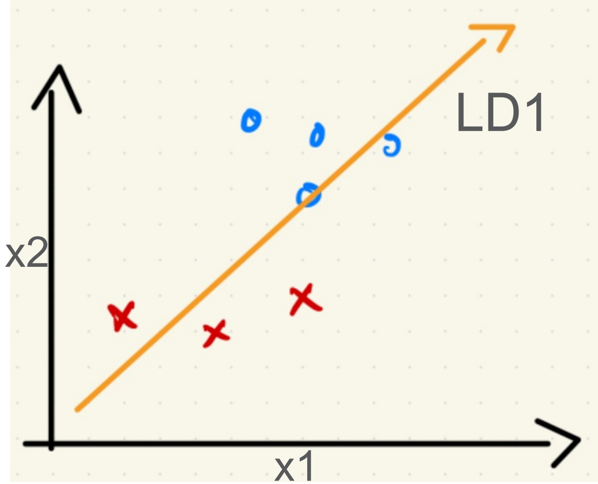
We maximize mean distance and minimize spread.

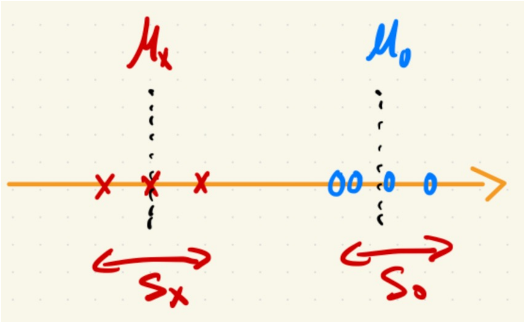
$$\frac{(\mu_x - \mu_0)^2}{s_x + s_0}$$

Review

- Feature Extraction
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
    - Supervised technique
    - Compute classes scatter matrices
    - The eigenvectors (linear discriminants) project samples into a lower-dimensional space and point toward the data direction that maximizes class discrimination.
  - t-distributed stochastic neighbor embedding (t-SNE)







We maximize mean distance and minimize spread.

$$\frac{(\mu_x - \mu_0)^2}{s_x + s_0}$$

$S_W^{-1} S_B$

THE UNIVERSITY OF TENNESSEE KNOXVILLE

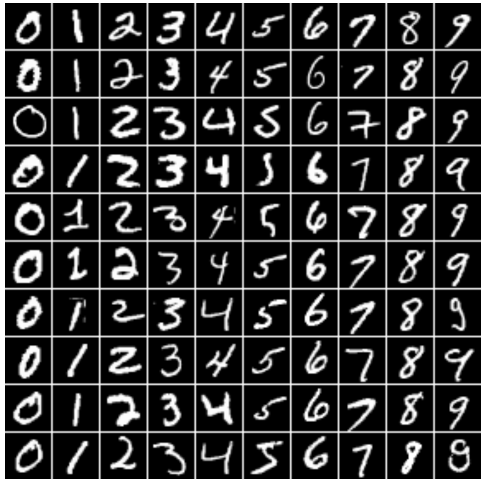
https://www.youtube.com/watch?v=azXCzI57Yfc

LDA Tutorial

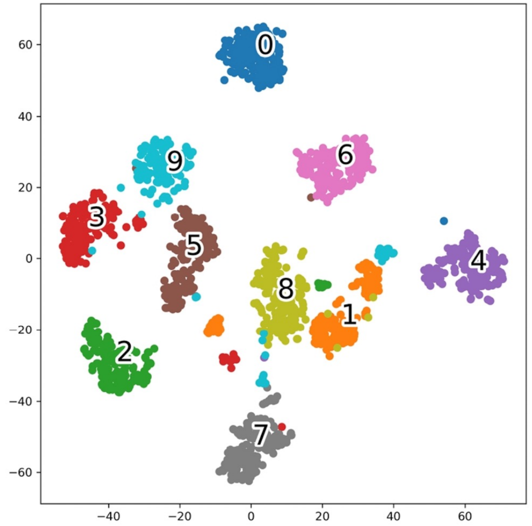
Review

- Feature Extraction
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
  - t-distributed stochastic neighbor embedding (t-SNE)
    - Unsupervised technique
    - Non-linear dimensionality reduction

MNIST Digits Dataset



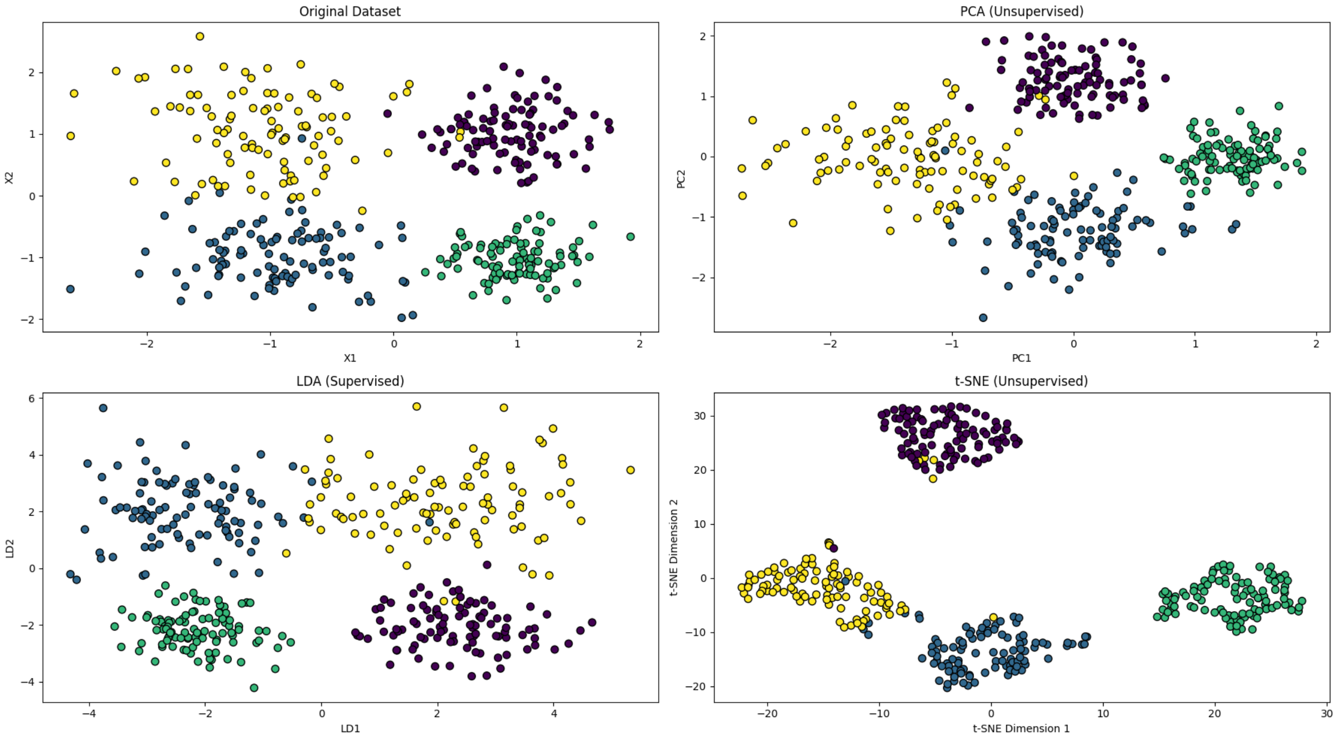
Digits t-SNE Projection



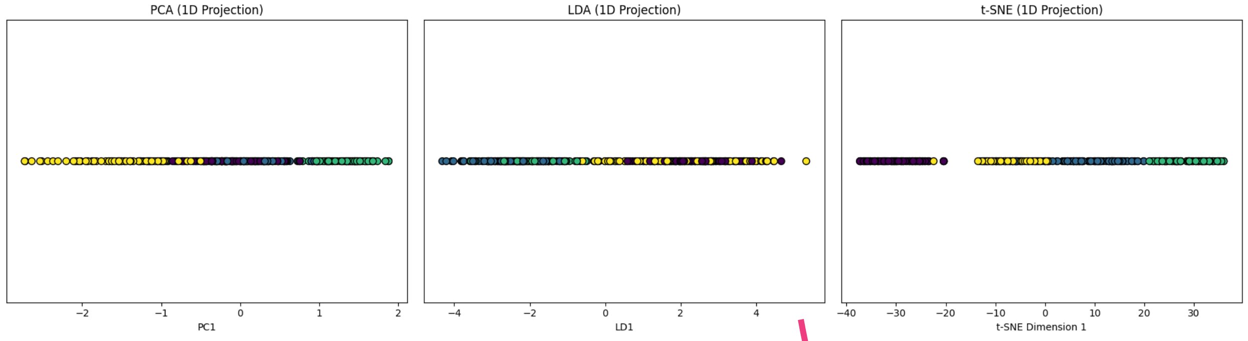
THE UNIVERSITY OF TENNESSEE KNOXVILLE

Review

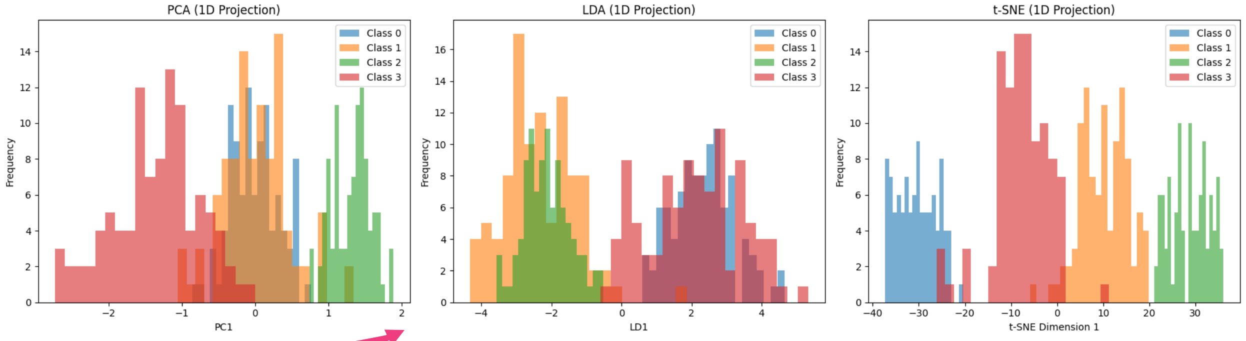
- Feature Extraction
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
  - t-distributed stochastic neighbor embedding (t-SNE)
    - Unsupervised technique
    - Non-linear dimensionality reduction



2-Components



1-Component



Histograms

Today’s Topics

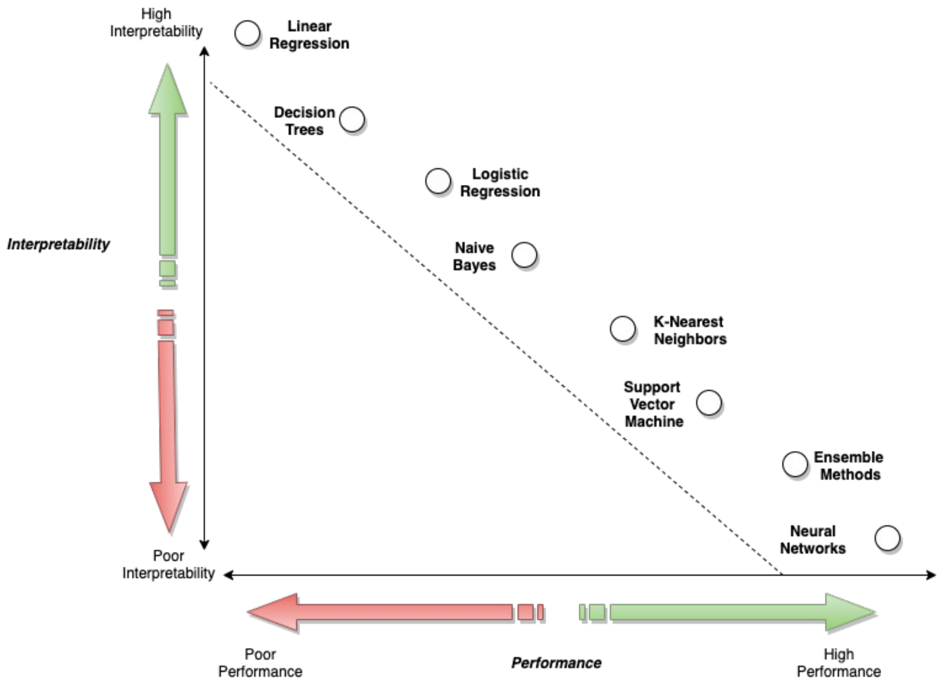
Explainability

Model Explainability (XAI)

- Methods that explain model decisions in human terms.
  - Connect patterns in the inputs to model decisions.
- **Interpretability:** the method explains the model predictions (i.e., the why and how).
  - Complete understanding of the inner model mechanics (contribution of model parameters)
  - Explain the relationship between inputs, model parameters, and predictions.

Model Explainability (XAI)

- Methods that explain model decisions in human terms.
  - Connect patterns in the inputs to model decisions.
- **Interpretability:** the method explains the model predictions (i.e., the why and how).
  - Complete understanding of the inner model mechanics (contribution of model parameters)
  - Explain the relationship between inputs, model parameters, and predictions.



Source: https://docs.aws.amazon.com/whitepapers/latest/model-explainability-aws-ai-ml/interpretability-versus-explainability.html

Terminology Check

- Local explainability
  – Method explains the relationship between the features and a prediction

What are the risks of patient A developing disease X?

- Global explainability
  – Method explains the relationship between the features and all model predictions

What are the risks of USA patients developing disease X?

Popular XAI Methods

- Feature importance (We already saw these)
  - E.g., Random Forest and Permutation feature importance
- SHapley Additive exPlanations (SHAP)
- Local Interpretable Model-Agnostic Explanations (LIME)
- Partial Difference Plot (PDP)
  - Individual Conditional Expectation (ICE)
- Counterfactual Explanations
  - What feature perturbations change the prediction?

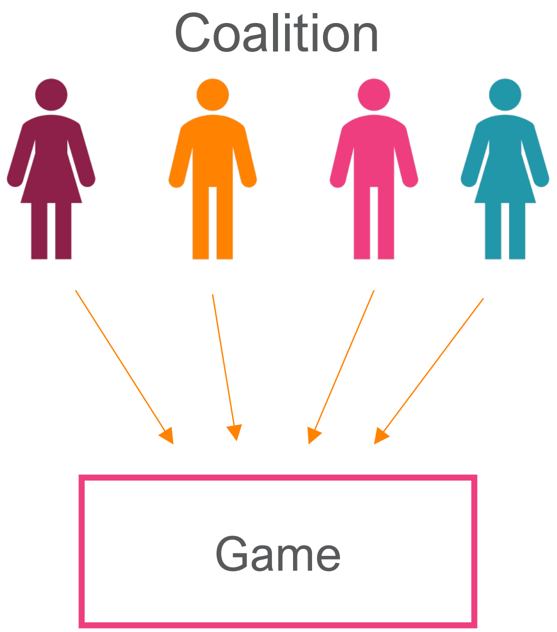
Popular XAI Methods

- Feature importance (We already saw these)
  - E.g., Random Forest and Permutation feature importance
- SHapley Additive exPlanations (SHAP)
- Local Interpretable Model-Agnostic Explanations (LIME)
- Partial Difference Plot (PDP)
  - Individual Conditional Expectation (ICE)
- Counterfactual Explanations
  - What feature perturbations change the prediction?

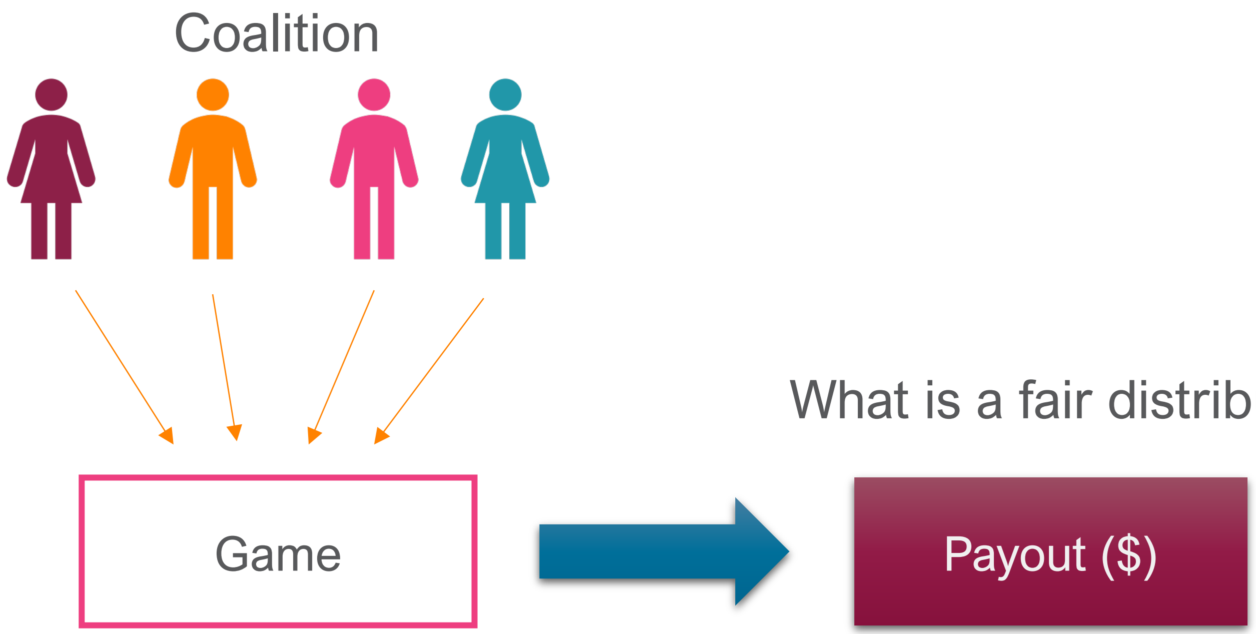
SHapley Additive exPlanations (SHAP)

- Model-Agnostic
- Based on game theory (Shapley, 1953).
- Explain the contributions of each feature to a specific prediction by estimating the features Shapley’s values.
- Provides local (individual predictions) and global (overall feature importance) explanations.
- Link: https://shap.readthedocs.io/en/latest/

Shapley Values

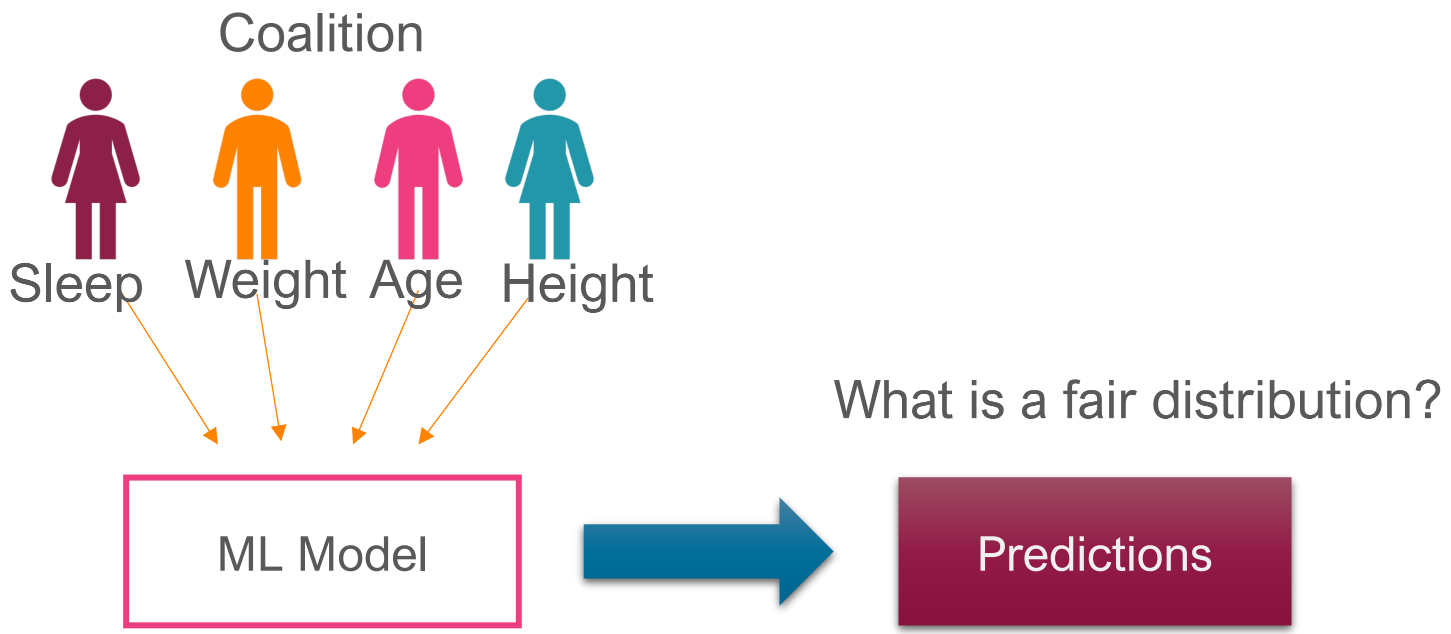


Shapley Values



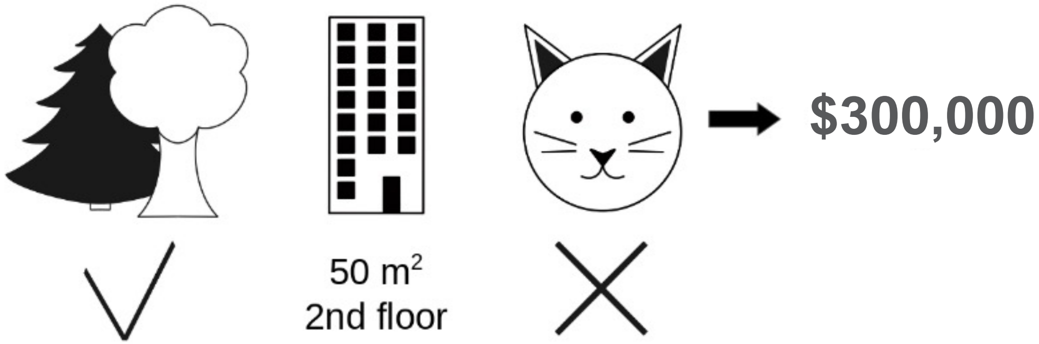
What is a fair distribution?

Shapley Values



SHapley Additive exPlanations (SHAP)

- Use case: Predict apartment prices.



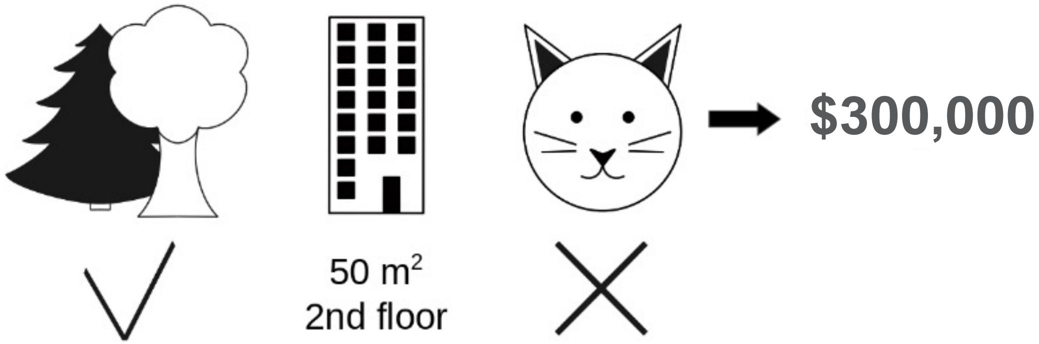
https://christophm.github.io/interpretable-ml-book/shapley.html

Average prediction for all apartments is $310k.

How much did each feature contribute to the prediction **compared to the average prediction**?

SHAP

- Use case: Predict apartment prices.



https://christophm.github.io/interpretable-ml-book/shapley.html

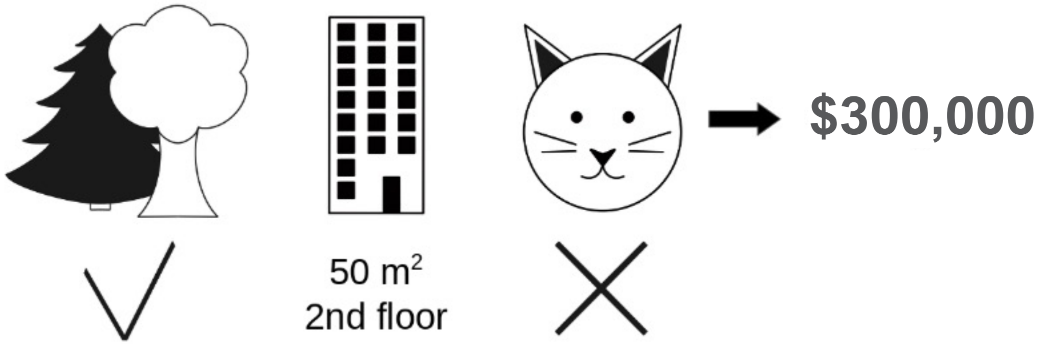
Average prediction for all apartments is $310k.

How much did each feature contribute to the prediction compared to the average prediction?

"The Shapley value is the average marginal contribution of a feature value across all possible coalitions."

SHAP

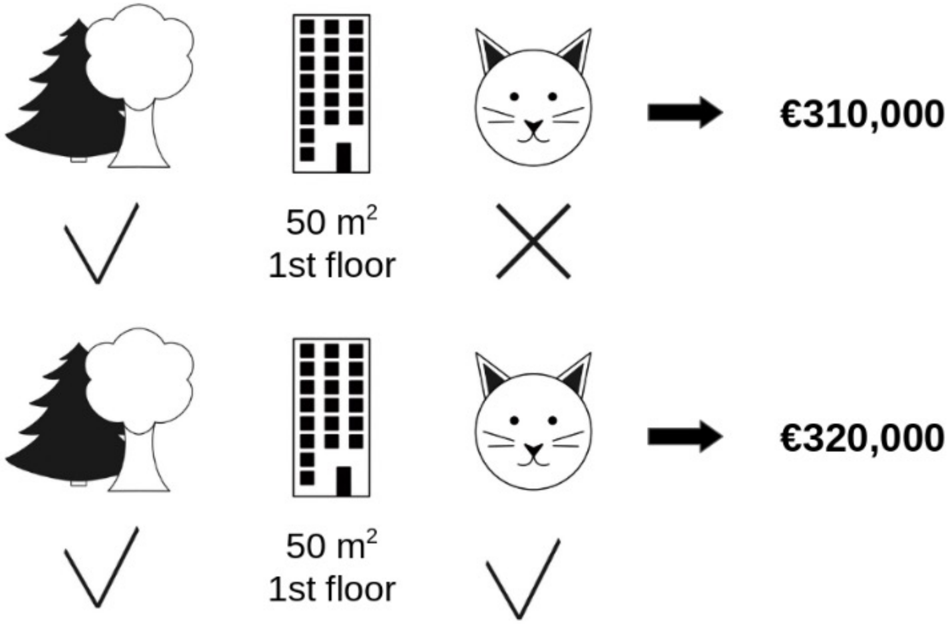
- Use case: Predict apartment prices.



https://christophm.github.io/interpretable-ml-book/shapley.html

Average prediction for all apartments is $310k.

"The Shapley value is the average marginal contribution of a feature value across all possible coalitions."

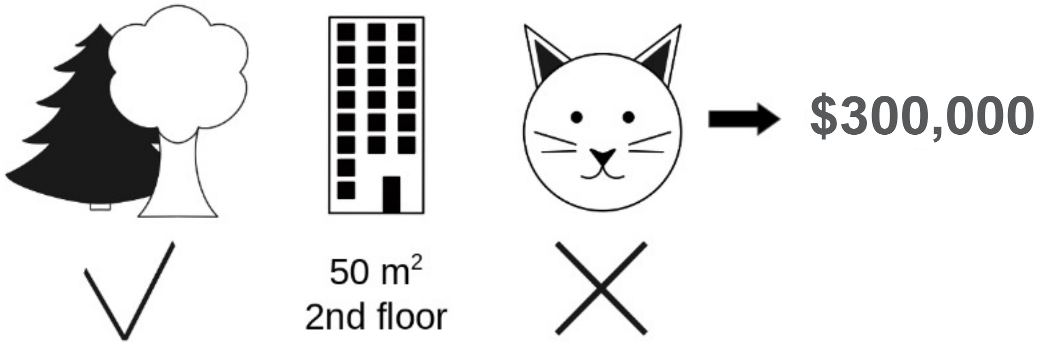


The contribution of "banned cats" is -$10,000

SHAP

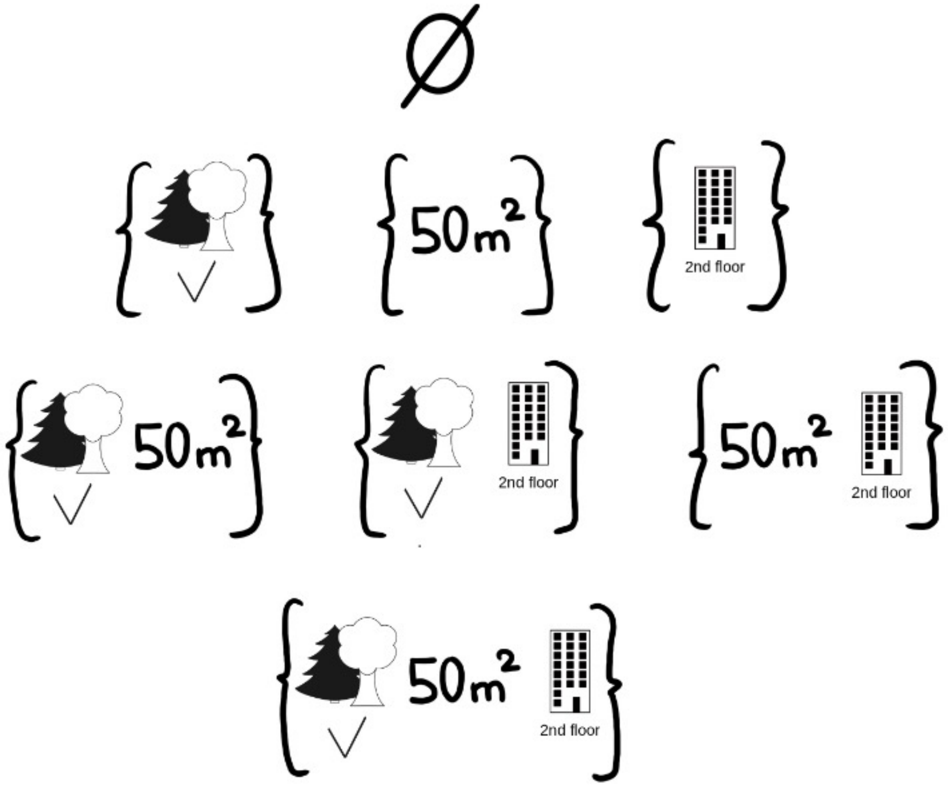
- Use case: Predict apartment prices.

"The Shapley value is the average marginal contribution of a feature value across all possible coalitions."



https://christophm.github.io/interpretable-ml-book/shapley.html

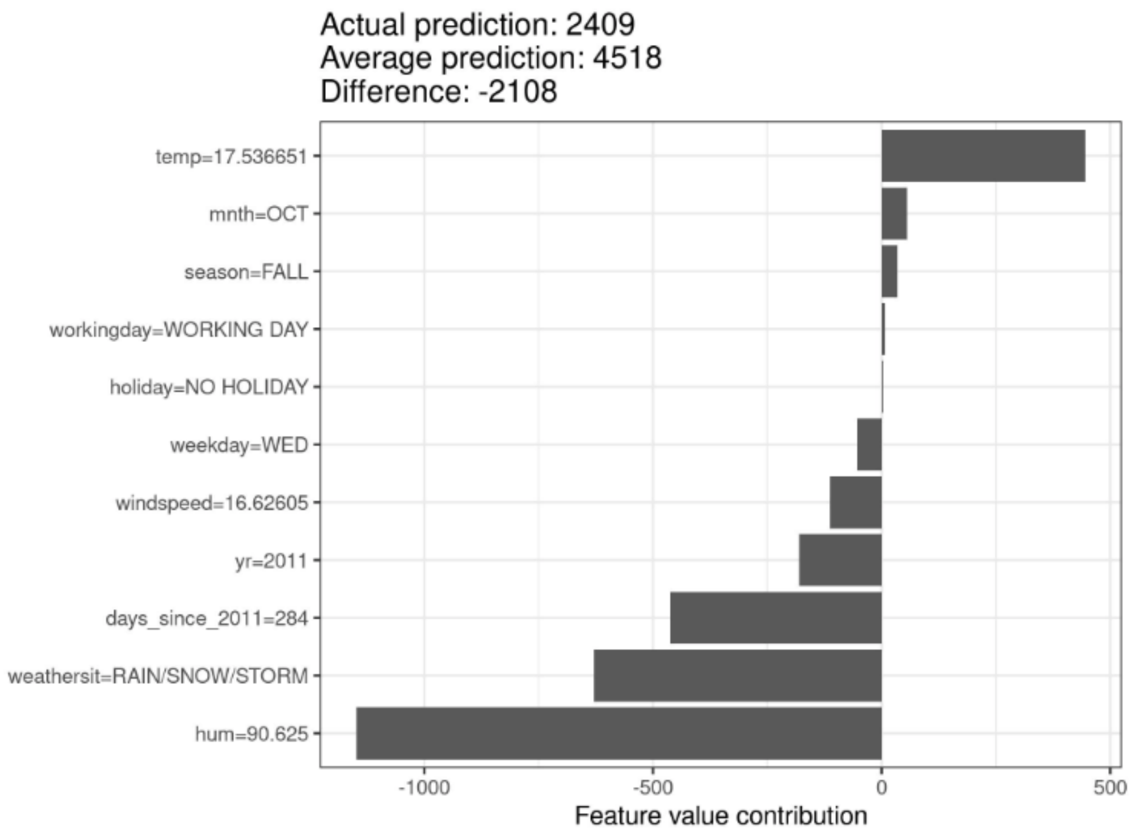
Average prediction for all apartments is $310k.



Feature of interest is replaced with random values.

SHAP

- Use case: Number of bike rentals

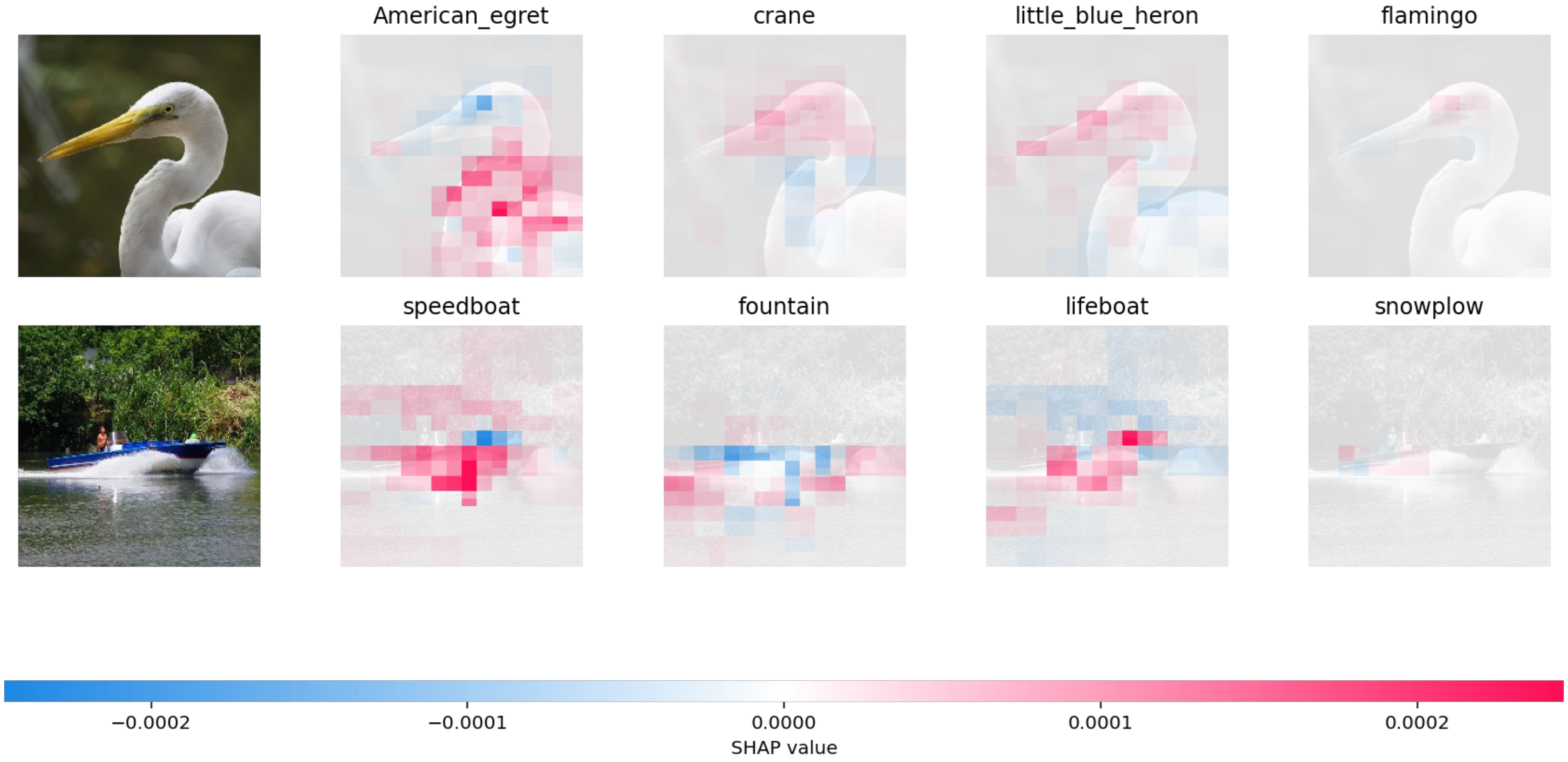


"The Shapley value is NOT the difference in prediction when we would remove the feature from the model."

"The Shapley value is the average contribution of a feature value to the prediction in different coalitions."

https://christophm.github.io/interpretable-ml-book/shapley.html

SHAP Image Examples



Source: https://shap.readthedocs.io/en/latest/example_notebooks/image_examples/image_classification/Explain%20ResNet50%20using%20the%20Partition%20explainer.html

SHAP Fair Payout Properties

- **Efficiency:** The contributions of a set of features must add up to the difference between the prediction and the average prediction values.
- **Symmetry:** The contributions of two features should be the same if they equally contribute to all possible coalitions.
- **Dummy:** A feature that does not change predicted value regardless of the coalition should have Shapley value of zero
- **Additivity:** You can add Shapley values from different games (E.g., Random Forest Model)

SHAP Advantages

- Guarantees fair distribution of feature contributions
- Contrastive explanations
  - Compares to single sample, subset, or whole dataset
- Solid theory
- Model agnostic
  - New Shapley Value approximations are not model-agnostic
    - Kernel SHAP, Tree SHAP, Deep SHAP

SHAP Disadvantages

- Work best for the complete feature set
- A lot of compute time with 2^m possible coalitions
- Misinterpretations: E.g., Loss in precision if the feature is removed.
  - Given a current set of feature values, the Shapley value measures the contribution of a feature to the difference between the current and mean prediction.
- Cannot be used to make statements about changes in prediction for changes in the feature values.
- Needs access to the data and uncorrelated features

https://www.youtube.com/watch?v=9haIOplEIGM

SHAP Tutorial

Pop Quiz

In the context of SHAP (SHapley Additive exPlanations), which of the following best describes the purpose of Shapley values?

A. To visualize the distribution of each feature in the dataset.

B. To measure the prediction error of a model on test data.

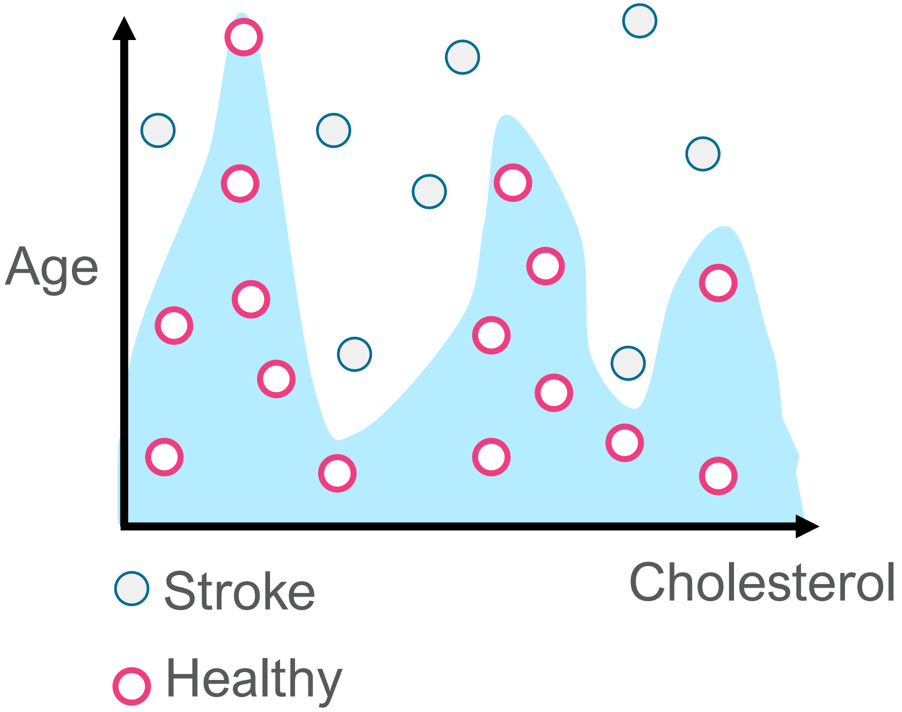
C. To assign a fair "contribution score" to each feature.

D. To standardize the features to ensure equal scaling across all features.

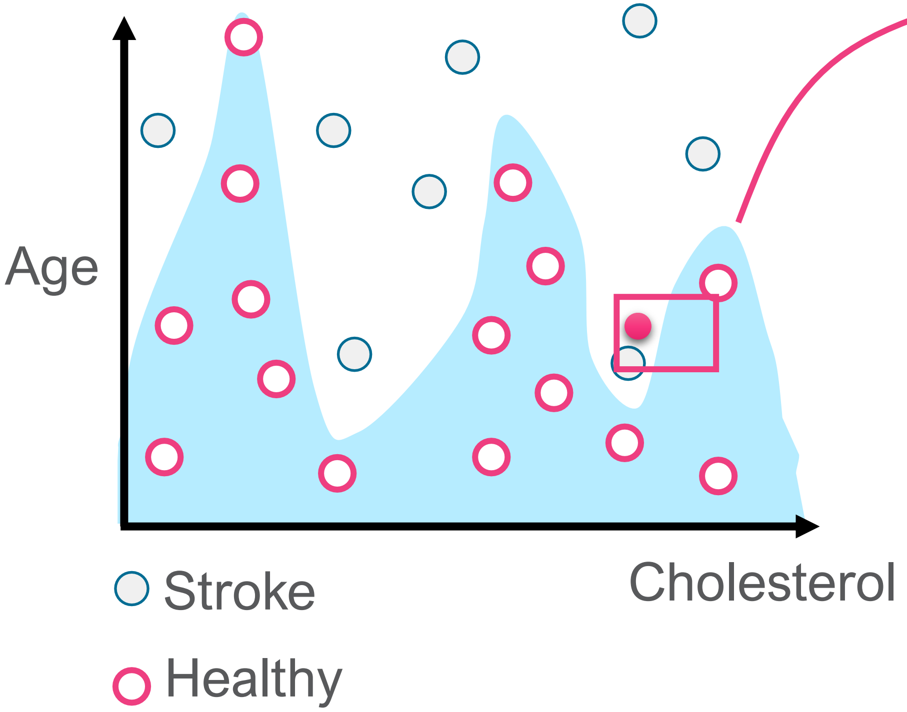
Local Interpretable Model-Agnostic Explanations (LIME)

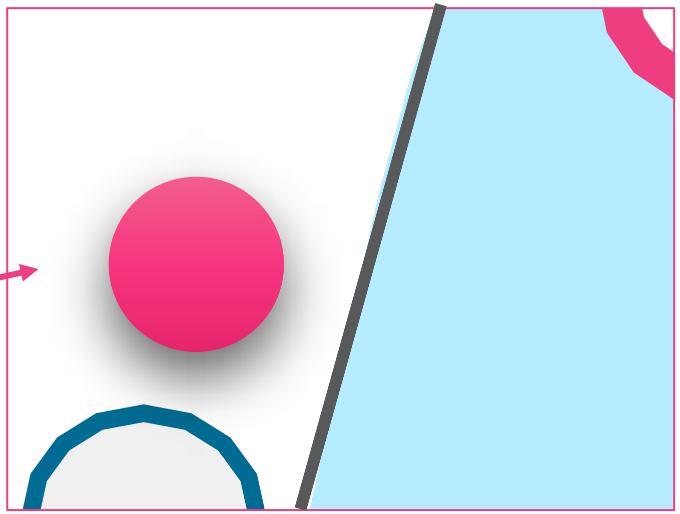
- Model-Agnostic
- Explain individual predictions by fitting a surrogate, interpretable model to a small neighborhood near the decision boundary of a more complex model.
- Provides local explanations
- Link: https://github.com/marcotcr/lime

LIME Intuition



LIME Concept





Locally, age is the only feature impacting prediction.

LIME Theory

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

LIME Theory

Complex Model

Input sample

Surrogate, Interpretable Model

Regularization term to ensure the sparsity of surrogate model parameters

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Proximity Weights

Family of surrogate models

LIME Theory

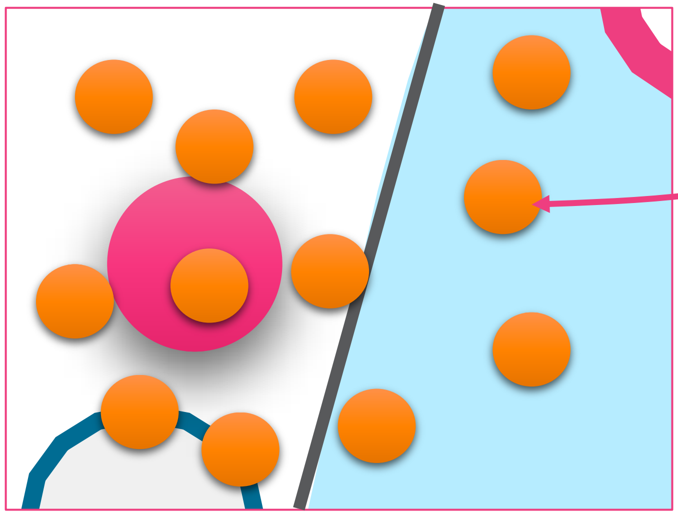
$$\xi(x) = \arg\min_{g\in G} \overbrace{\mathcal{L}(f, g, \pi_x)}^{\text{Good Approximation}} + \underbrace{\Omega(g)}_{\text{Stay Simple}}$$

We are searching for a good surrogate g.

Computing the Loss

$$\xi(x) = \arg\min_{g \in G} \boxed{\mathcal{L}(f, g, \pi_x)} + \Omega(g)$$

Randomly Generated Samples D

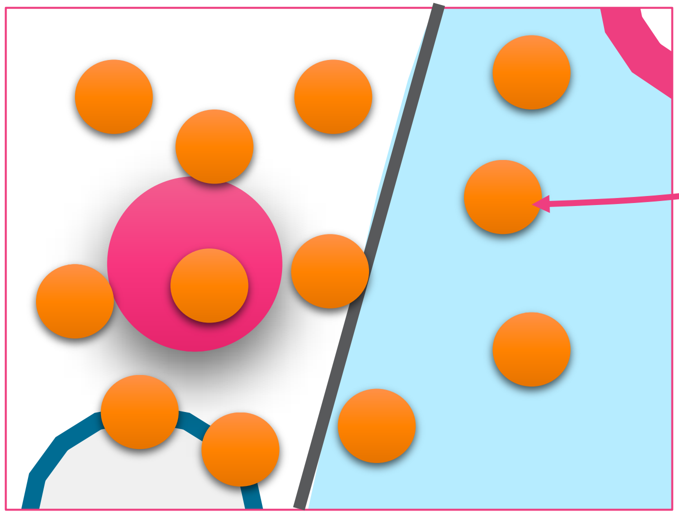


1. Get predictions from $f(D)$
2. Use predictions as labels for the new dataset D
3. Train g with new dataset D

Computing the Loss

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Randomly Generated Samples D



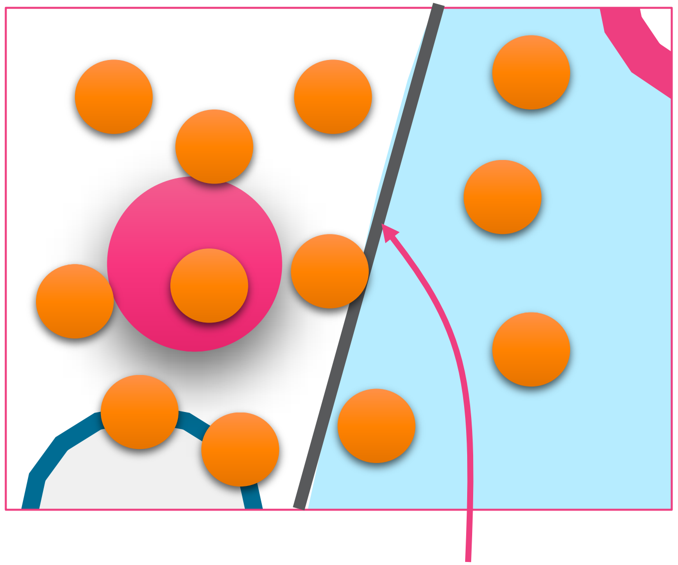
1. Get predictions from $f(D)$
2. Use predictions as labels for the new dataset D
3. Train g with new dataset D
4. Use π_x to penalize the loss from samples far away from the sample under inspection.

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in Z} \pi_x(z)\big(f(z) - g(z')\big)^2$$

THE UNIVERSITY OF TENNESSEE KNOXVILLE

Computing the Loss

$$\xi(x) = \underset{g \in G}{\arg\min}\, \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

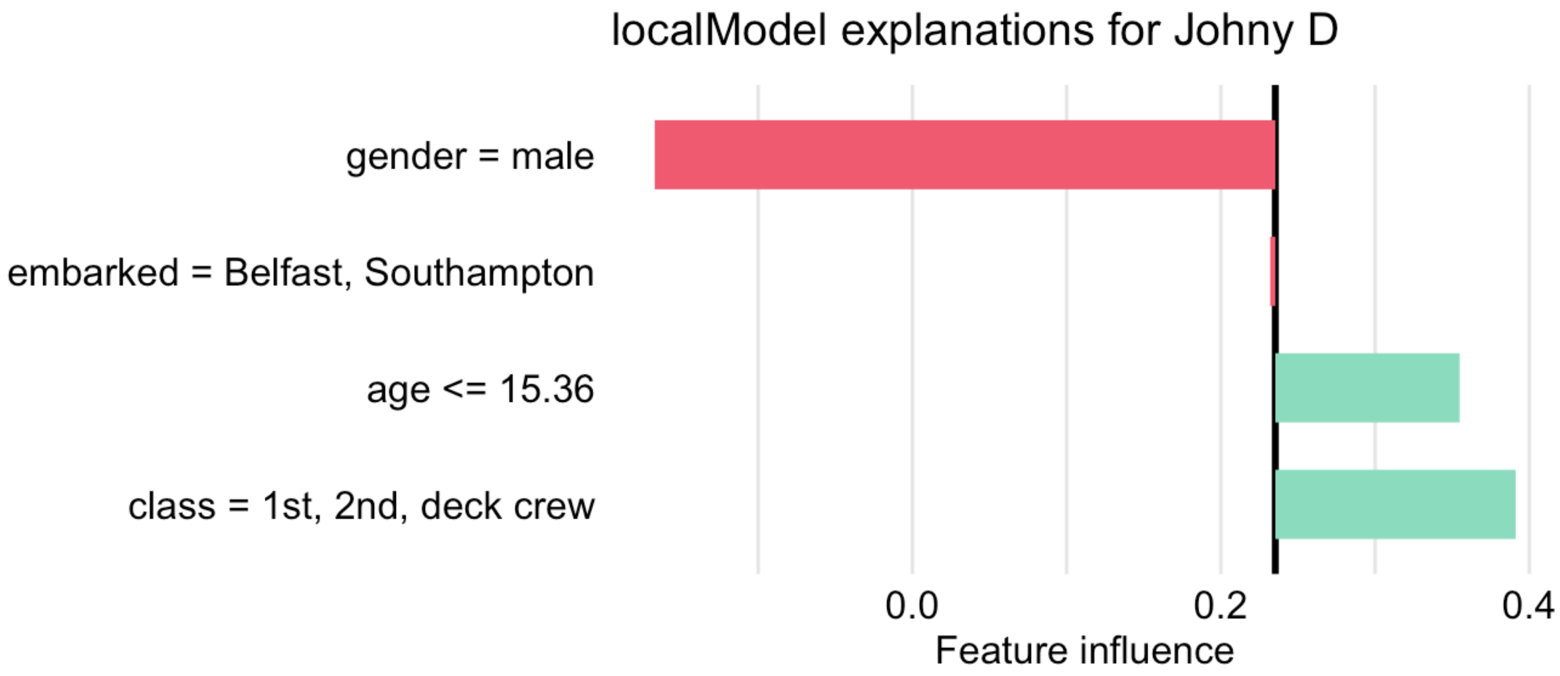


$y = w_0 + w_1 Cholesterol + w_2 Age$

1. Get predictions from $f(D)$
2. Use predictions as labels for the new dataset D
3. Train g with new dataset D
4. Use π_x to penalize the loss from samples far away from the sample under inspection.
5. After finding our g, we can use its weights for a local explanation of the features' influence on the prediction.

Tabulated Data Example

Titanic Dataset Johny D Sample



Feature influence values are the coefficients of the surrogate model.

Image Examples

Google's Inception v3 predictions

(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" ($p = 0.21$)

Source: https://velog.io/@tobigs_xai/1%EC%A3%BC%EC%B0%A8-LIME-%EB%85%BC%EB%AC%B8-%EB%A6%AC%EB%B7%B0-Why-Should-I-Trust-You-Explaining-the-Predictions-of-Any-Classifier

Lime Advantages

- Model agnostic
  - Work with any model
  - Model internals are hidden
- Work with many data types
  - Text, images, tabulated data, etc.
- Expert knowledge can validate LIME results
  - Accurate explanations create trust

Lime Disadvantages

- No proper definition of local neighborhood
- Needs access to the data
- Only faithful local explanations
- Sparse/high dimensional data could break the technique
  - Unstable explanations
  - Potential manipulation of explanations

https://www.youtube.com/watch?v=d6j6bofhj2M&list=PLV8yxwGOxvvovp-j6ztxhF3QcKXT6vORU&index=3

LIME Tutorial

Pop Quiz

Which of the following statements is TRUE about the LIME (Local Interpretable Model-Agnostic Explanations) method in machine learning explainability?

A. LIME calculates feature contributions by considering all possible combinations of features, similar to SHAP.

B. LIME approximates the model's behavior by creating a simplified, interpretable model around the prediction point of interest.

C. LIME plots the effect of a feature on the prediction by averaging over all values of other features, similar to PDP.

D. LIME is primarily used for global explanations and understanding overall feature importance across the entire dataset.

Review

- Explainability techniques
  - SHAP
  - LIME
  - PDP

THE UNIVERSITY OF TENNESSEE KNOXVILLE

Next Lecture

- Unsupervised learning

Helper Slides

https://ttpoll.com/p/744372