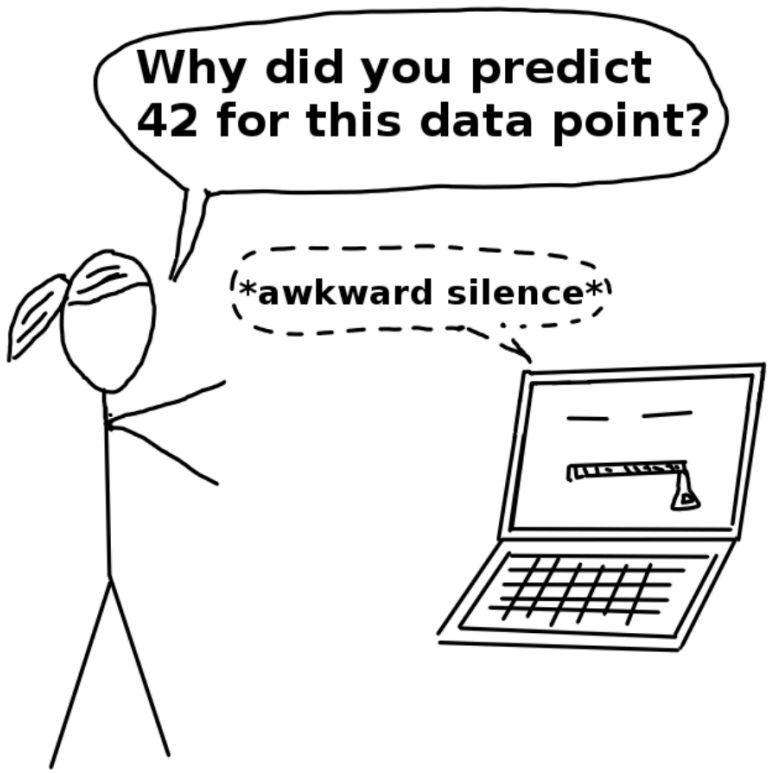
Interpretable Machine Learning

Alex Wang 2020.08.26

- Introduction
- Interpretable Models
- Model-Agnostic Methods
- Example-Based Explanations

Introduction

Black Box Model: a system that does not reveal its internal mechanisms. In machine learning, "black box" describes models that cannot be understood by looking at their parameters (e.g.: a neural network).



- **Interpretability**
  - the degree to which a human can understand the cause of a decision
- **Importance of Interpretability**
  - The more a machine's decision affects a person's life, the more important it is for the machine to explain its behavior.
    - Algorithmic product recommendation
    - Loan application
  - Debugged and Audited
  - Fairness: Ensuring that predictions are unbiased and do not implicitly or explicitly discriminate against protected groups.
  - Reliability or Robustness: Ensuring that small changes in the input do not lead to large changes in the prediction.

Interpretable Models

- Linear Regression
- Logistic Regression
- Decision Trees

Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

- Easy to interpret
- Numeric Features: size of a house, temperature, price
  - Increasing the numerical feature by one unit changes the estimated outcome by its weight
- Categorical Features: Season, Gender
  - Changing the feature from the reference category to the other category changes the estimated outcome by the feature's weight.

Linear Regression $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$

	Weight	SE	t
(Intercept)	2399.4	238.3	10.1
seasonSUMMER	899.3	122.3	7.4
seasonFALL	138.2	161.7	0.9
seasonWINTER	425.6	110.8	3.8
holidayHOLIDAY	-686.1	203.3	3.4
workingdayWORKING DAY	124.9	73.3	1.7
weathersitMISTY	-379.4	87.6	4.3
weathersitRAIN/SNOW/STORM	-1901.5	223.6	8.5
temp	110.7	7.0	15.7
hum	-17.4	3.2	5.5
windspeed	-42.5	6.9	6.2
days_since_2011	4.9	0.2	28.5

Prediction of the number of rented bikes on a particular day

Linear Regression

Disadvantages

- Nonlinearity or interaction has to be hand-crafted
- The interpretation of a weight can be unintuitive because it depends on all other features
  - **House size** and **number of rooms** are highly correlated:
  - the bigger a house is, the more rooms it has.
  - Take both features into a linear model, the **size of the house** might get a large positive weight while **the number of rooms** might get a negative weight
  - Because, given that a house has the same size, increasing the number of rooms could make it less valuable

Logistic Regression

$$P(y^{(i)} = 1) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_p x_p^{(i)}))}$$

Logistic Regression

$$P(y^{(i)} = 1) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_p x_p^{(i)}))}$$

$$log\left(\frac{P(y=1)}{P(y=0)}\right) = log\left(\frac{P(y=1)}{1 - P(y=1)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

Logistic Regression

$$P(y^{(i)} = 1) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_p x_p^{(i)}))}$$

$$log\left(\frac{P(y=1)}{P(y=0)}\right) = log\left(\frac{P(y=1)}{1 - P(y=1)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

$$odds = \frac{P(y=1)}{1 - P(y=1)} = exp\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p\right)$$

Logistic Regression

$$P(y^{(i)} = 1) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_p x_p^{(i)}))}$$

$$log\left(\frac{P(y=1)}{P(y=0)}\right) = log\left(\frac{P(y=1)}{1 - P(y=1)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

$$odds = \frac{P(y=1)}{1 - P(y=1)} = exp\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p\right)$$

$$\frac{odds_{x_j+1}}{odds} = \frac{exp\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_j (x_j + 1) + \ldots + \beta_p x_p\right)}{exp\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_j x_j + \ldots + \beta_p x_p\right)}$$

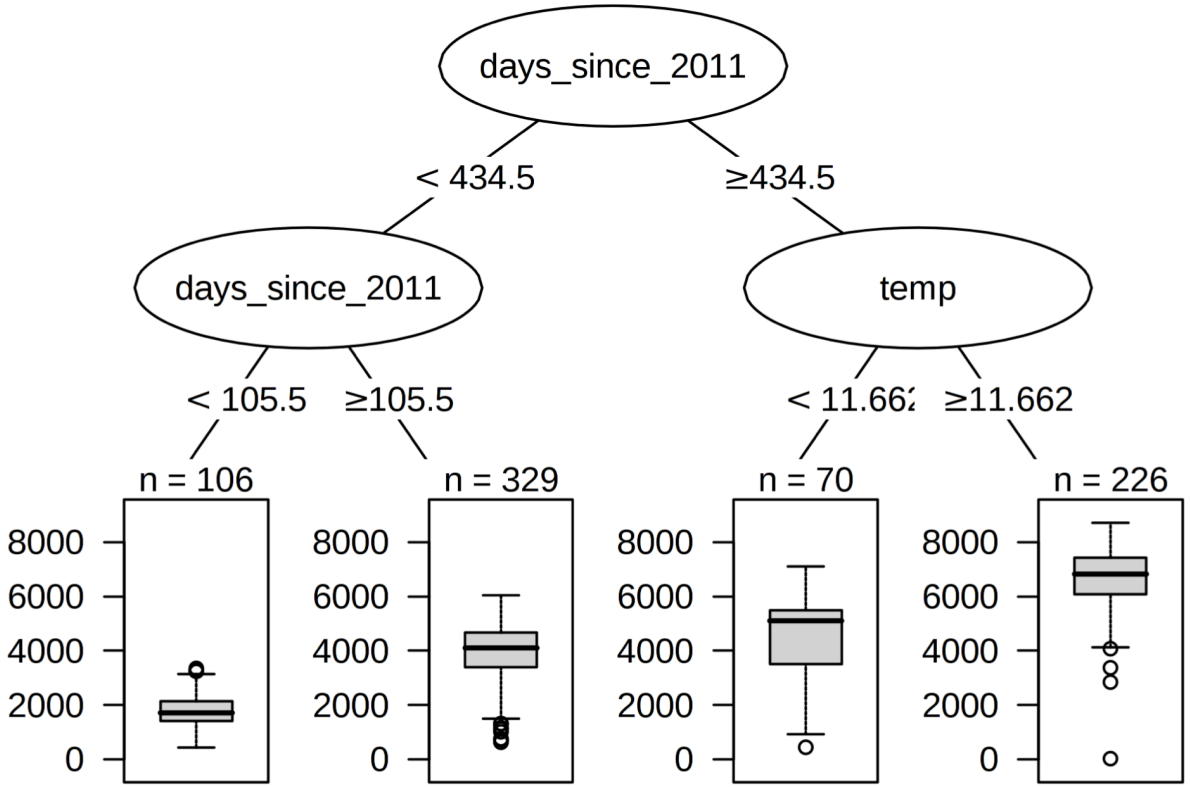
Logistic Regression

$$\frac{odds_{x_j+1}}{odds} = exp\left(\beta_j(x_j+1) - \beta_j x_j\right) = exp\left(\beta_j\right)$$

$$odds = \frac{P(y=1)}{P(y=0)}$$

- A change in a feature by one unit changes the odds ratio (multiplicative) by a factor of $exp\left(\beta_j\right)$
- Disadvantages: the interpretation of the weights is multiplicative makes it more difficult to interpret

Decision Trees



Prediction of the number of rented bikes on a particular day

Decision Trees

- Advantages
  - Trees create good explanations
  - Trees Structure has a natural visualization
- Disadvantages
  - **Lack of smoothness:** Slight changes in the input feature can have a big impact on the predicted outcome. If split by 100, prediction of 99.9 and 100.0 may differ a lot.
  - **Unstable:** A few changes in the training dataset can create a completely different tree
    - A different feature that selected as parent split, the entire structure will change

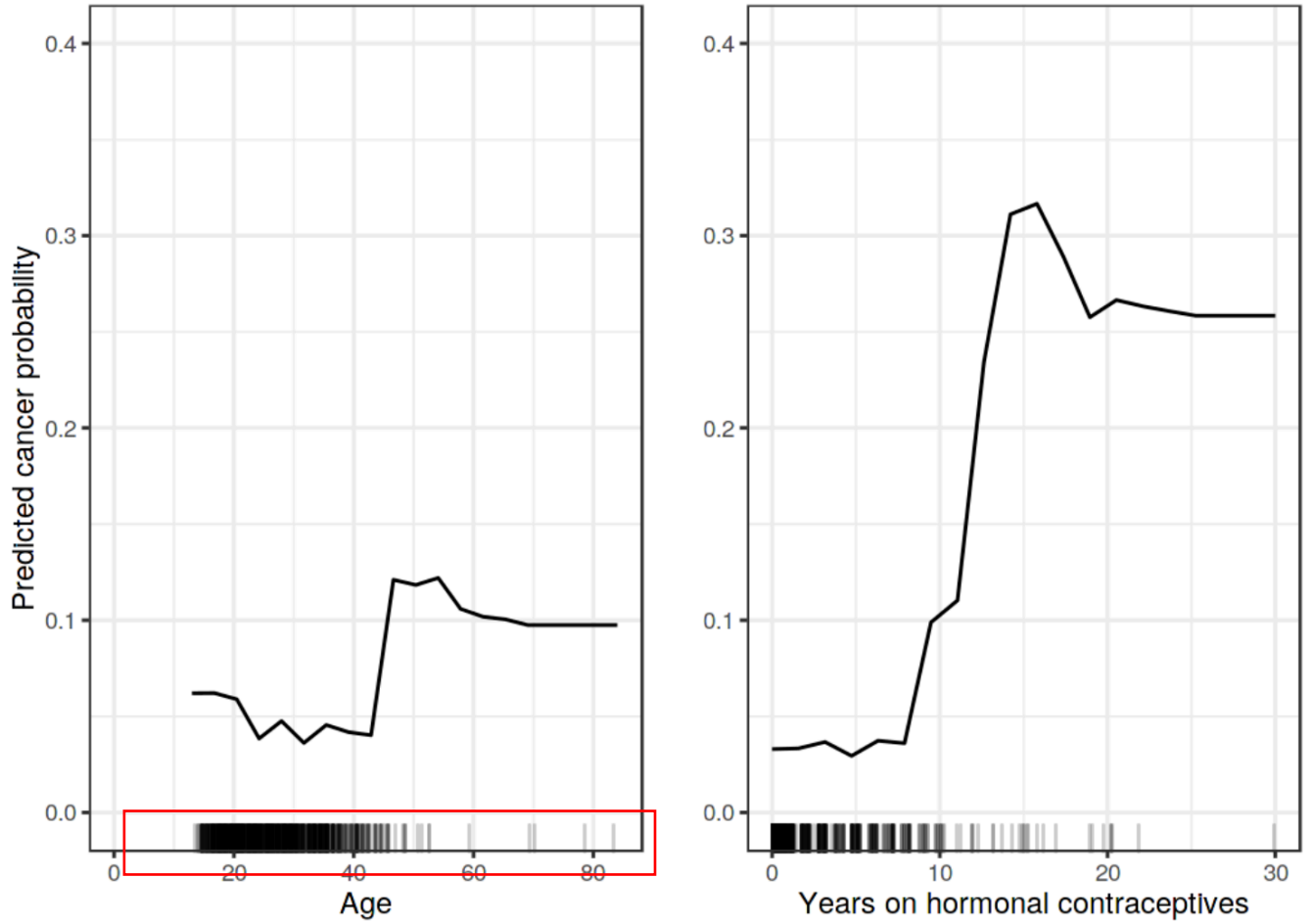
Model-Agnostic Methods

- Partial Dependence Plot (PDP)
- Individual Conditional Expectation (ICE)
- Accumulated Local Effects (ALE)
- Global Surrogate
- Local Surrogate (LIME)
- Shapley Values

Partial Dependence Plot (PDP)

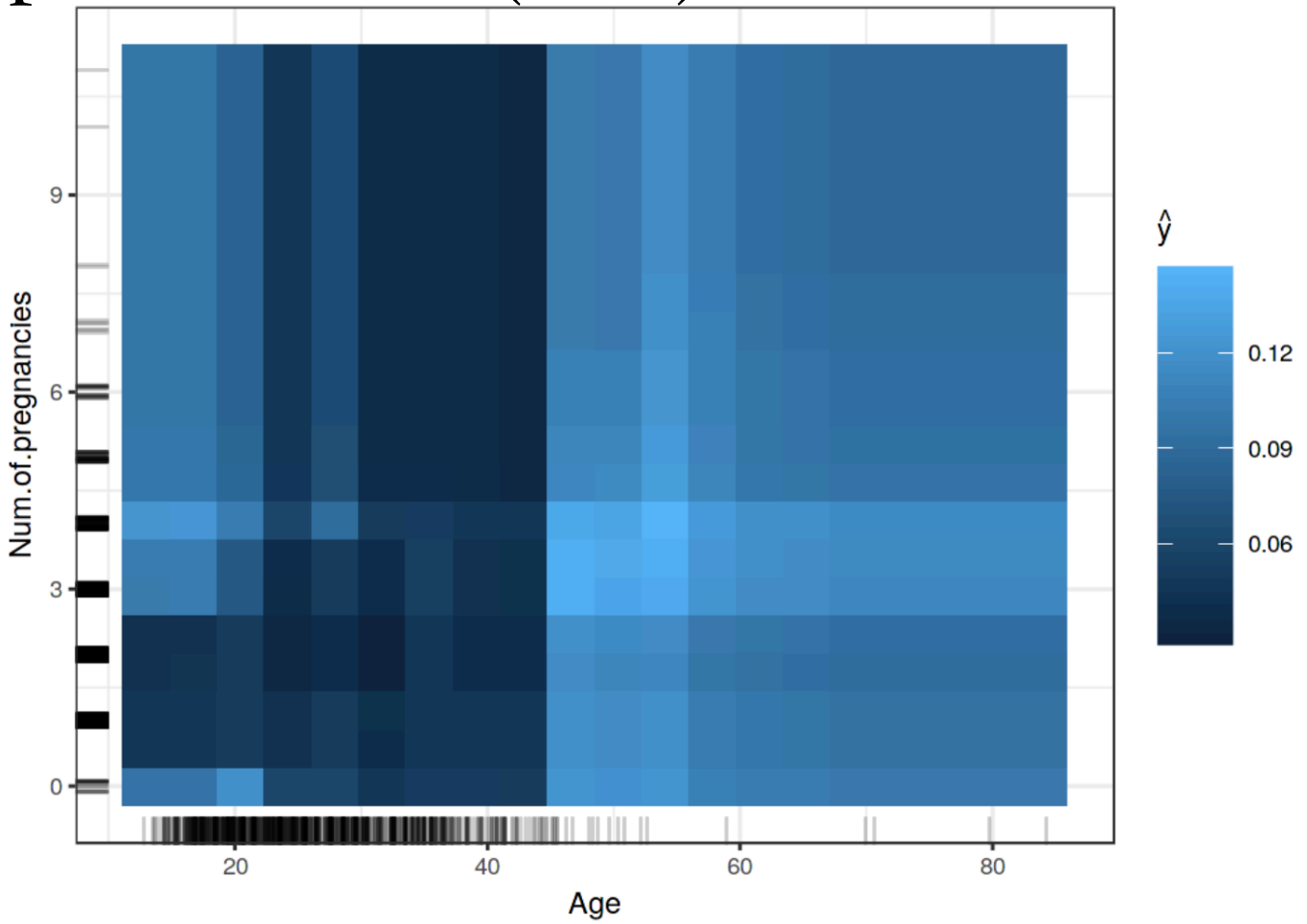
- shows the marginal effect one or two features have on the predicted outcome of a machine learning model
- Calculate PDP when x_i=14
  - Set x_i=14 for all instances
  - Make the prediction
  - PDP = Average of all the instances' predictions
- Calculate PDP for all possible x_i, plot PDP Curve

Partial Dependence Plot (PDP)



PDPs of cancer probability based on age and years with hormonal contraceptives. For age, the PDP shows that the probability is low until 40 and increases after. The more years on hormonal contraceptives the higher the predicted cancer risk, especially after 10 years. For both features not many data points with large values were available, so the PD estimates are less reliable in those regions.

Partial Dependence Plot (PDP)



PDP of cancer probability and the interaction of age and number of pregnancies. The plot shows the increase in cancer probability at 45. For ages below 25, women who had 1 or 2 pregnancies have a lower predicted cancer risk, compared with women who had 0 or more than 2 pregnancies. But be careful when drawing conclusions: This might just be a correlation and not causal!

Partial Dependence Plot (PDP)

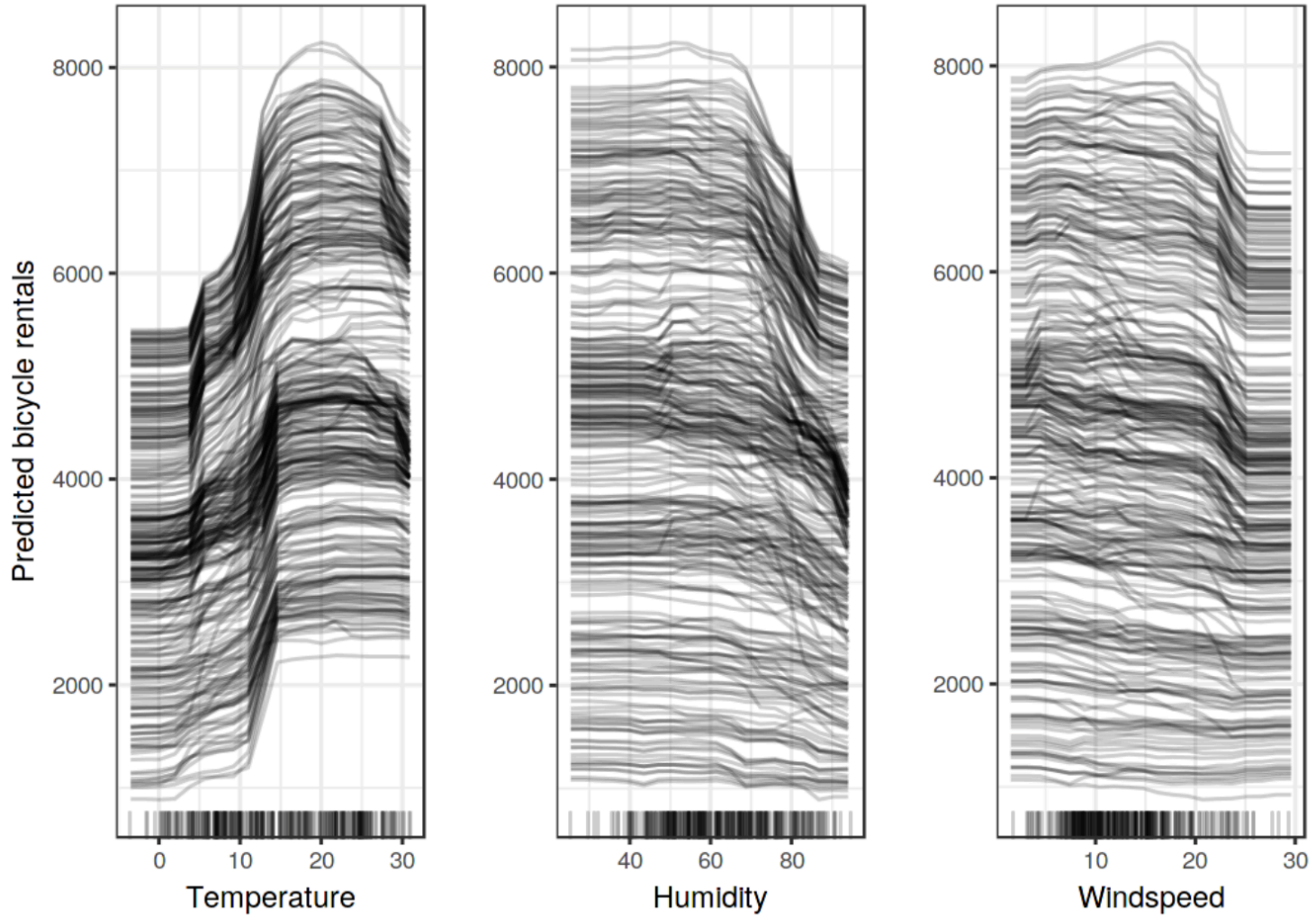
- Advantages
  - Intuitive
  - **In the uncorrelated case**, the interpretation is clear
  - Causal interpretation
- Disadvantages
  - Difficult to plot high dimension representation, maximum number of features are usually limited to 2
  - The assumption of independence
  - Heterogeneous effects might be hidden: For a given x, half of instances have positive predictions while others have negative predictions, PDP may close to zero

Individual Conditional Expectation (ICE)

- Plot visualizes the dependence of the prediction on a feature for each instance separately.

- Disadvantages: When there are too many instances, the plot is Difficult to interpret (too many lines)

Individual Conditional Expectation (ICE)

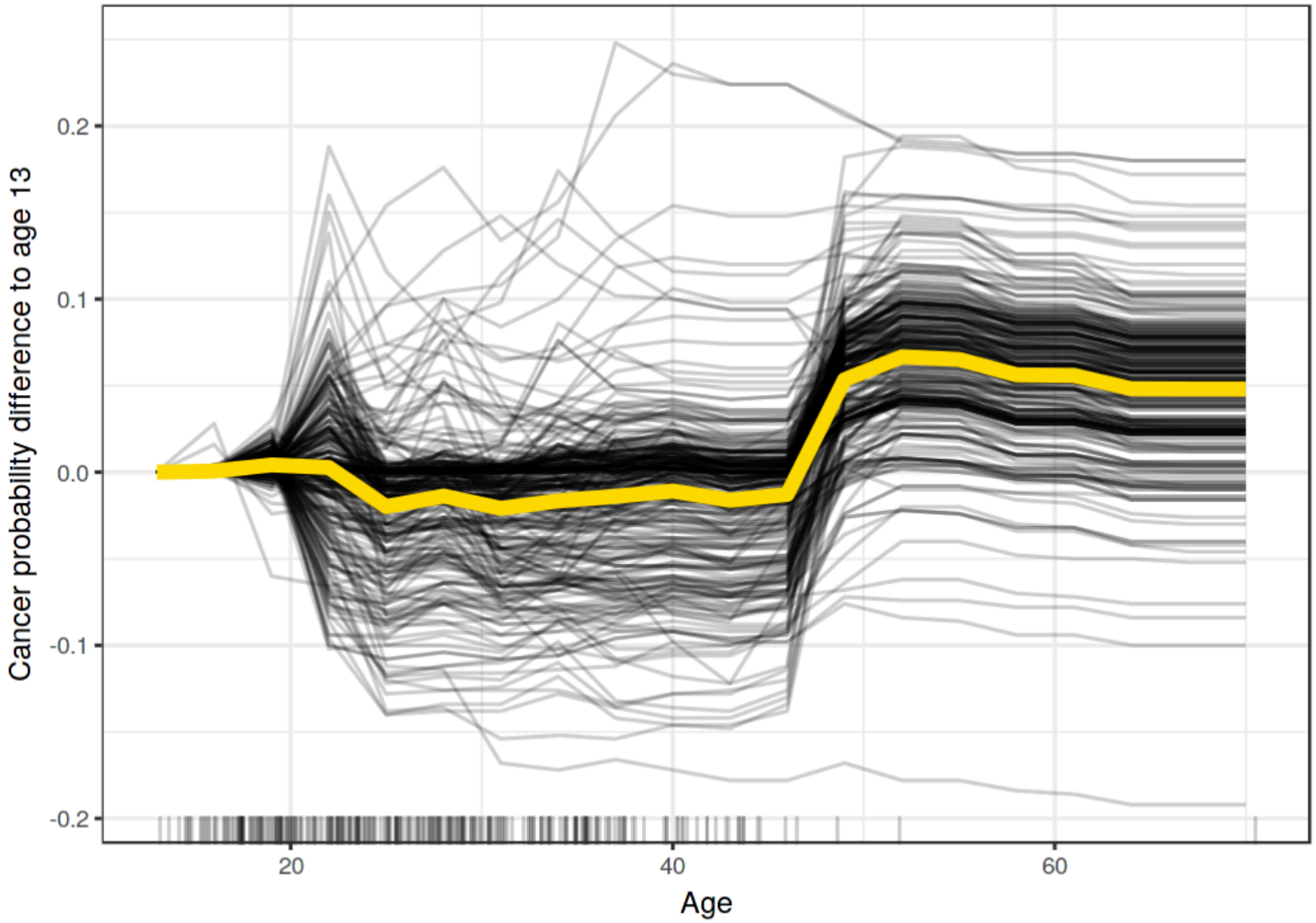


ICE plots of predicted bicycle rentals by weather conditions. The same effects can be observed as in the partial dependence plots.

Centered ICE Plot

- Center the curves at a certain point in the feature
- Display only the difference in the prediction to this point

Centered ICE Plot



Centered ICE plot for predicted cancer probability by age. Lines are fixed to 0 at age 13. Compared to age 13, the predictions for most women remain unchanged until the age of 45 where the predicted probability increases.

ICE Plot

- Advantages
  - More intuitive to understand than PDP
  - Can uncover heterogeneous relationships
- Disadvantages
  - Can only display one feature
  - The assumption of independence
  - Plot can become overcrowded

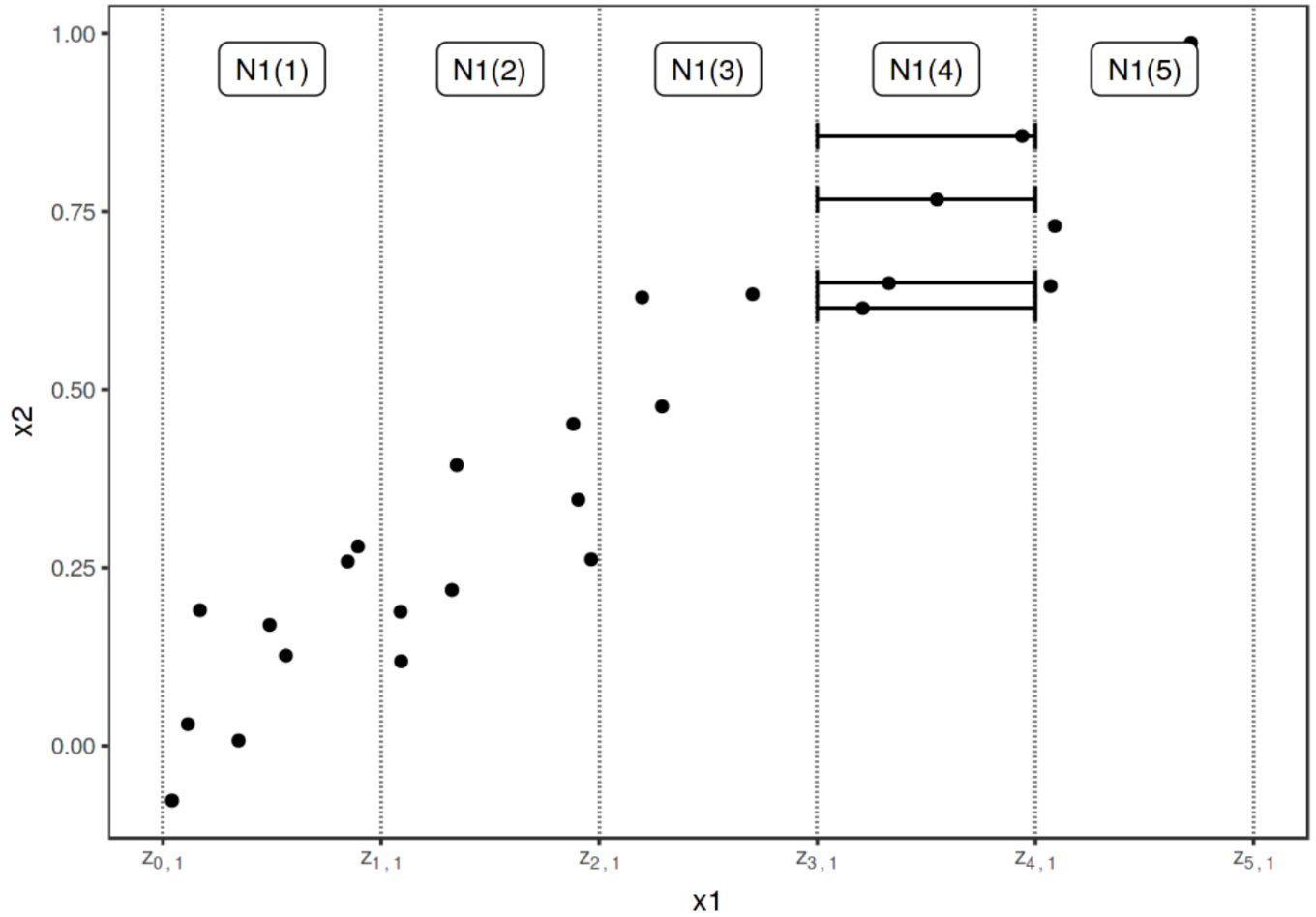
Accumulated Local Effects Plot (ALE)

- A faster and unbiased alternative to PDP
- Describe how a feature affects the prediction on average
- Calculating differences in predictions instead of averages

Accumulated Local Effects Plot (ICE)

- ALE when size of house = 30
  - Uses all houses with about 30
  - Set size of house to 31 and 29 and make prediction respectively
  - The former one minus the latter one is the ALE when size of house = 30

Accumulated Local Effects Plot (ICE)



Calculation of ALE for feature x1, which is correlated with x2. First, we divide the feature into intervals (vertical lines). For the data instances (points) in an interval, we calculate the difference in the prediction when we replace the feature with the upper and lower limit of the interval (horizontal lines). These differences are later accumulated and centered, resulting in the ALE curve.

Accumulated Local Effects Plot (ICE)

- Advantages
  - Unbiased, still work when features are correlated
  - Faster to compute
  - Interpretation of ALE plots is clear
- Disadvantages
  - Shaky, no perfect solution for setting the number of intervals
  - Instances number in each interval is different, so is accuracy
  - Interpretation remains difficult for strongly correlated features

Global Surrogate

- Use an interpretable model that is trained to approximate the predictions of a black box model.
- Methods
  - For the selected dataset X, get the predictions of the black box model
  - Select an interpretable model (linear model, decision tree, ...)
  - Train the interpretable model on the dataset X and its predictions
  - Interpret the surrogate model
  - Use R^2 to Measure how well the surrogate model replicates the predictions of the black box model.

Global Surrogate

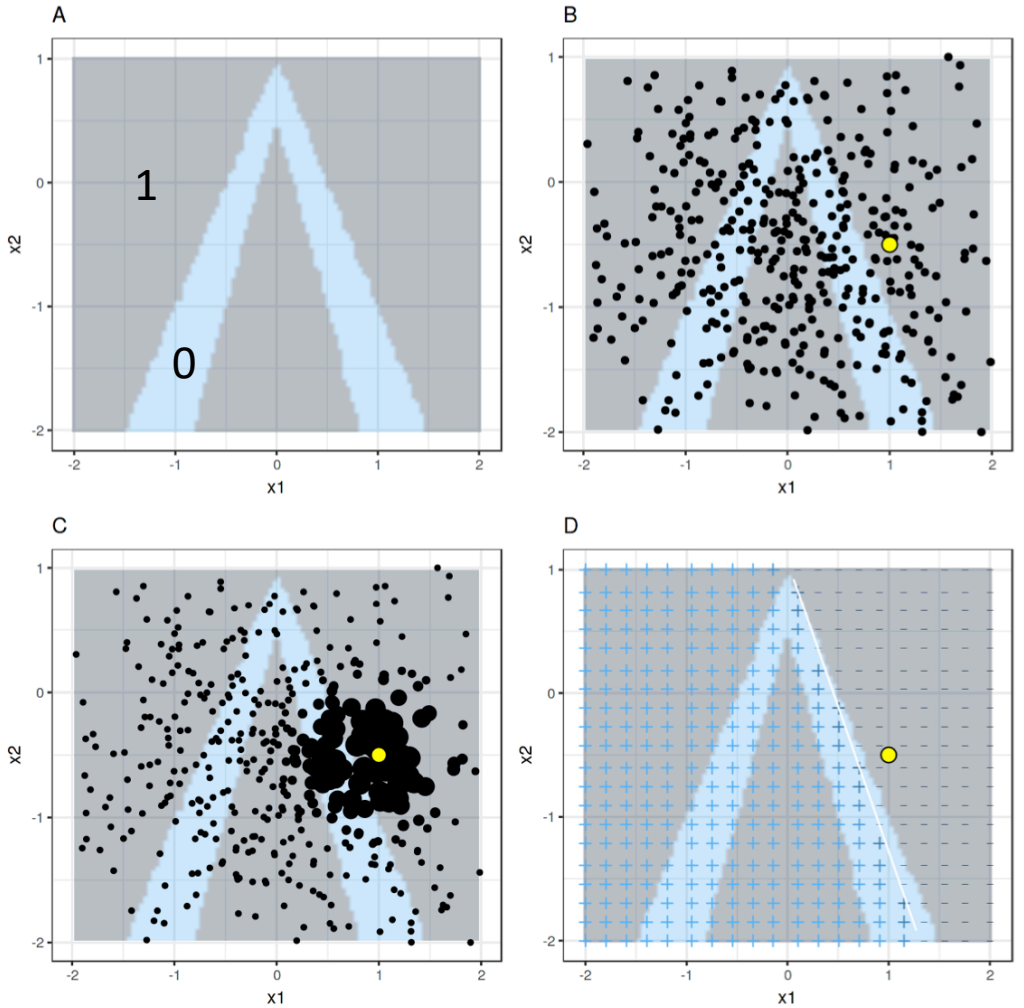
- Advantages
  - Flexible, Intuitive
- Disadvantages
  - Determination of the best cut-off for R^2
  - R^2 may change for same surrogate model and black box model on different dataset

Local Surrogate

- Focuses on training local surrogate models to explain individual predictions
- Methods
  - Select the instance of interest
  - Perturb the dataset and get the black box predictions for these new points
  - Weight the new samples according to their proximity to the instance of interest
  - Train a weighted, interpretable model on the dataset with the variations
  - Explain the prediction by interpreting the local model

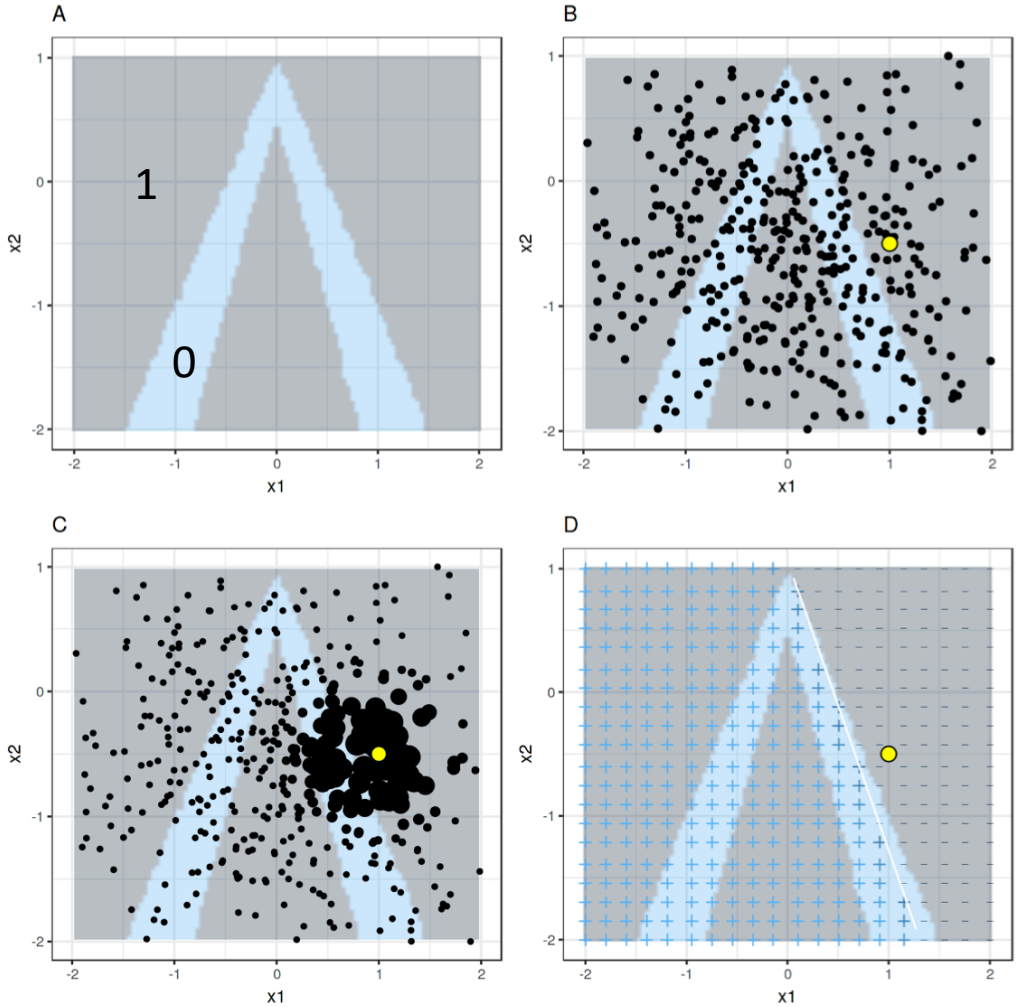
Local Surrogate

Predictions of
Random Forest



LIME algorithm for tabular data. A) Random forest predictions given features x1 and x2. Predicted classes: 1 (dark color) or 0 (light color). B) Instance of interest (yellow dot) and data sampled from a normal distribution (black dots). C) Assign higher weight to points near the instance of interest. D) Colors and signs of the grid show the classifications of the locally learned model from the weighted samples. The white line marks the decision boundary (P(class=1) = 0.5).

Local Surrogate

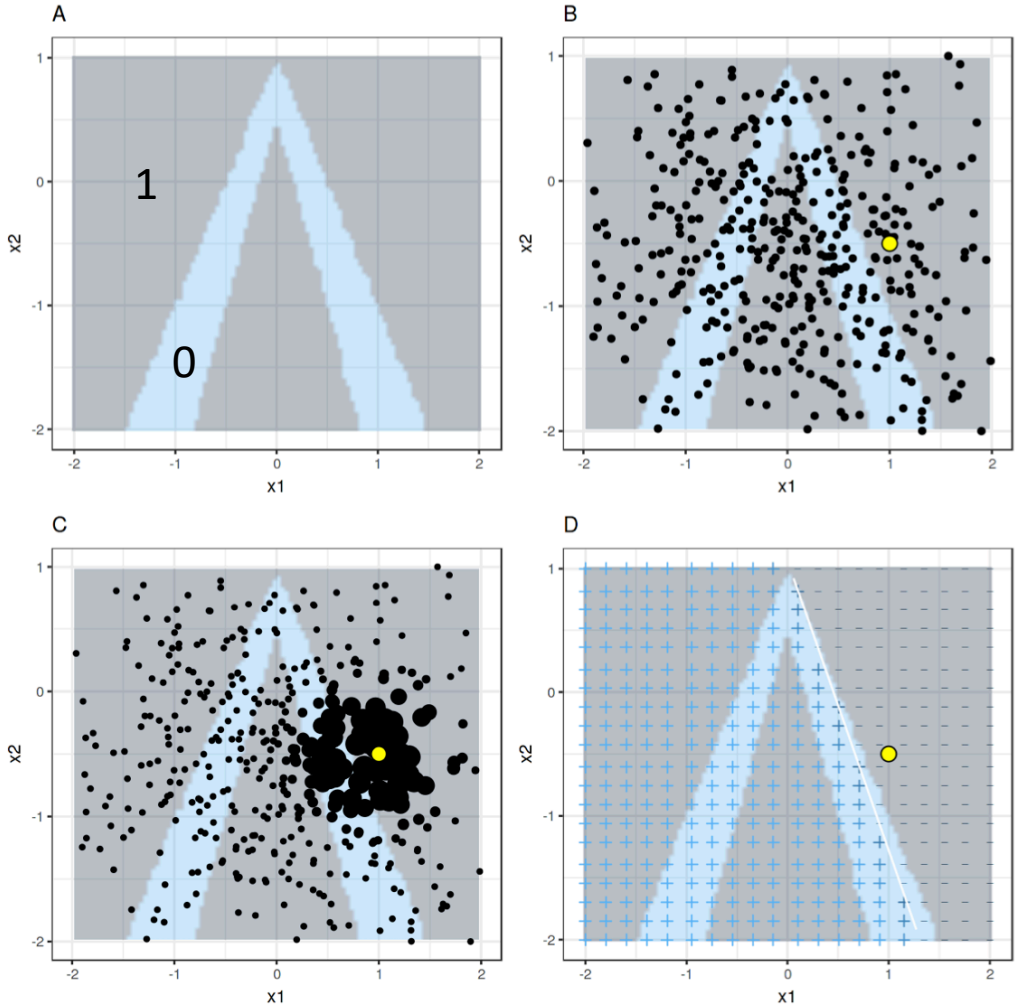


LIME algorithm for tabular data. A) Random forest predictions given features x1 and x2. Predicted classes: 1 (dark color) or 0 (light color). B) Instance of interest (yellow dot) and data sampled from a normal distribution (black dots). C) Assign higher weight to points near the instance of interest. D) Colors and signs of the grid show the classifications of the locally learned model from the weighted samples. The white line marks the decision boundary (P(class=1) = 0.5).

Instance of interest and data sampled from a normal distribution

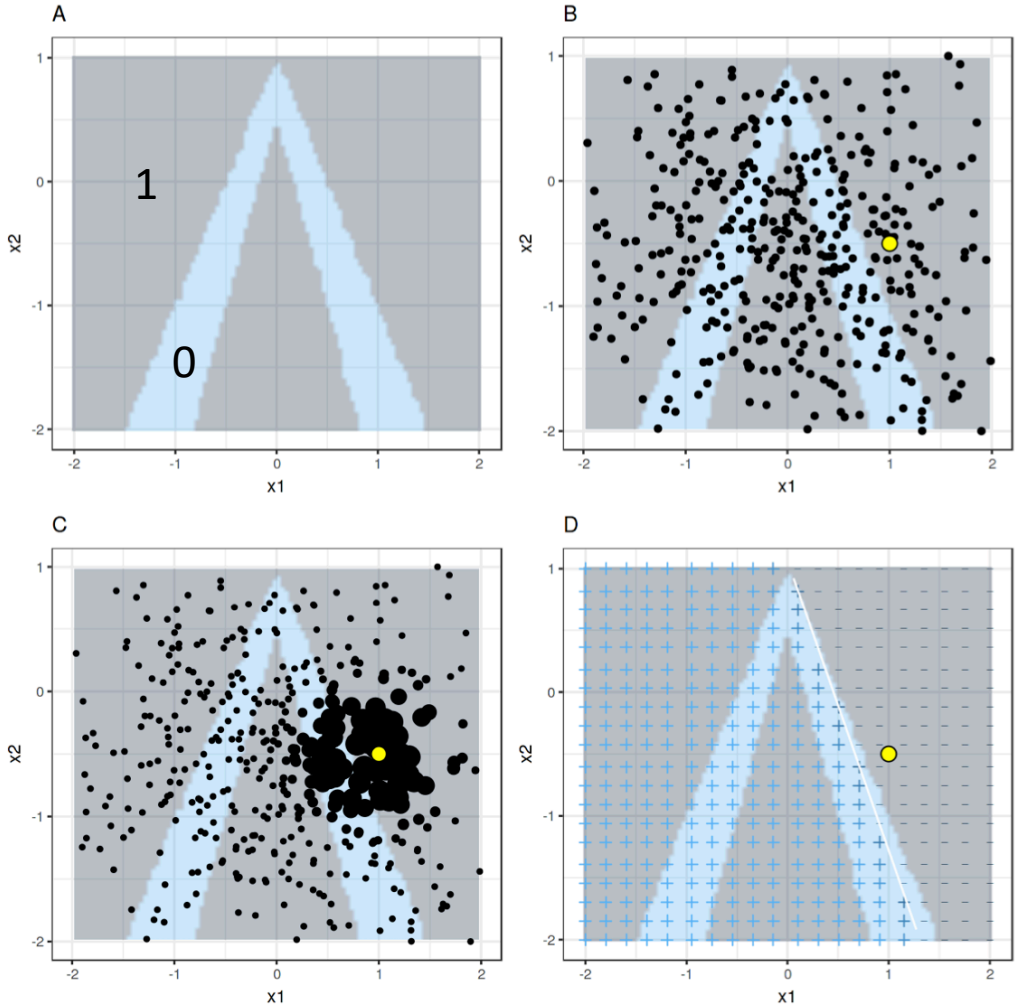
Local Surrogate

Assign higher weight to points near the instance of interest



LIME algorithm for tabular data. A) Random forest predictions given features x1 and x2. Predicted classes: 1 (dark color) or 0 (light color). B) Instance of interest (yellow dot) and data sampled from a normal distribution (black dots). C) Assign higher weight to points near the instance of interest. D) Colors and signs of the grid show the classifications of the locally learned model from the weighted samples. The white line marks the decision boundary (P(class=1) = 0.5).

Local Surrogate



LIME algorithm for tabular data. A) Random forest predictions given features x1 and x2. Predicted classes: 1 (dark color) or 0 (light color). B) Instance of interest (yellow dot) and data sampled from a normal distribution (black dots). C) Assign higher weight to points near the instance of interest. D) Colors and signs of the grid show the classifications of the locally learned model from the weighted samples. The white line marks the decision boundary (P(class=1) = 0.5).

Colors and signs of the grid show the classifications of the locally learned model from the weighted samples.

Local Surrogate

- Advantages
  - Works for tabular data, text and images
  - Can use other features than the original model
- Disadvantages
  - The correct definition of the neighborhood
  - Instability, the explanations of two very close points varied greatly in a simulated setting

Shapley

- The average marginal contribution of a feature value across all possible coalitions

Shapley

- Methods
  - For a dataset X with $\{x_1, x_2, ..., x_j\}$ features
  - Find features other than x_i: $\{x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_j\}$
  - Find all possible coalitions
  - Calculate the marginal contributions of x_i on each coalition
  - Average all marginal contributions

Shapley

- Example
  - Features in rent price prediction: 50 m2, 2nd floor, with a nearby park and cat ban
  - The average prediction for all apartments is €310,000, prediction for this apartment is €300,000
  - Find out each feature value contributed to the difference (€-10,000)

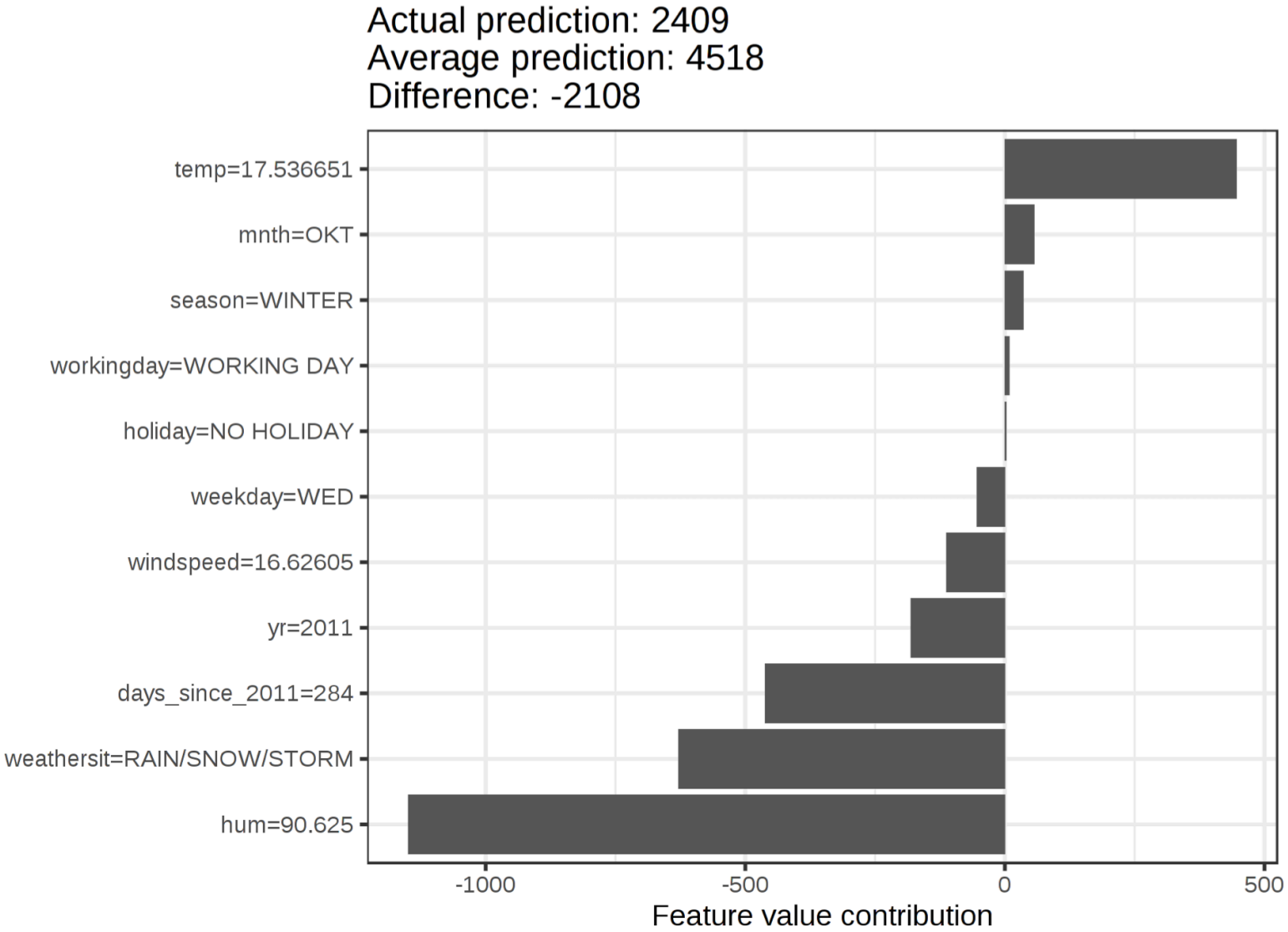
Shapley

- Example
  - Contribution of feature $50m^2$
  - Find all coalitions for: {2nd floor, with a nearby park, cat ban}
    - {No feature values}
    - {2nd floor}, {with a nearby park}, {cat ban}
    - {2nd floor, with a nearby park}, {2nd floor, cat ban}, {with a nearby park, cat ban}
    - {2nd floor, with a nearby park, cat ban}

Shapley

- Example
  - For coalitions {2nd floor}
  - A denotes the predicted apartment price for instances has {2nd floor}
  - B denotes the predicted apartment price for instances has {2nd floor, $50m^2$}
  - B-A is the marginal contribution of $50m^2$ on this coalition
  - Shapley Value is the average of the marginal contribution on all coalitions

Shapley


Actual prediction: 2409
Average prediction: 4518
Difference: -2108
temp=17.536651
mnth=OKT
season=WINTER
workingday=WORKING DAY
holiday=NO HOLIDAY
weekday=WED
windspeed=16.62605
yr=2011
days_since_2011=284
weathersit=RAIN/SNOW/STORM
hum=90.625
-1000
-500
0
500
Feature value contribution

Shapley

- Properties
  - Efficiency: The feature contributions must add up to the difference of prediction for x and the average.

$$\sum_{j=1}^{p} \phi_j = \hat{f}(x) - E_X(\hat{f}(X))$$

  - Symmetry: The contributions of two feature values j and k should be the same if they contribute equally to all possible coalitions
  - Dummy: A feature j that does not change the predicted value – regardless of which coalition of feature values it is added to – should have a Shapley value of 0

Shapley

- Advantages
  - Allows contrastive explanations
  - Might be the only method to deliver a full explanation
  - The only explanation method with a solid theory
- Disadvantages
  - Requires a lot of computing time
  - Always use all the features
  - Cannot explain changes in prediction for changes in the input
  - The assumption of independence

Example-Based Explanations

Counterfactual Explanations

- The smallest change to the feature values that changes the prediction to a predefined output

Score	GPA	LSAT	Race	GPA x'	LSAT x'	Race x'
0.17	3.1	39.0	0	3.1	34.0	0
0.54	3.7	48.0	0	3.7	32.4	0
-0.77	3.3	28.0	1	3.3	33.5	0
-0.83	2.4	28.5	1	2.4	35.8	0
-0.57	2.7	18.3	0	2.7	34.9	0

Changed features, predictions are close to zero

Others

- Adversarial examples
- Prototypes and Criticisms
- Influential Instances

Thanks

Alex Wang 2020.08.26