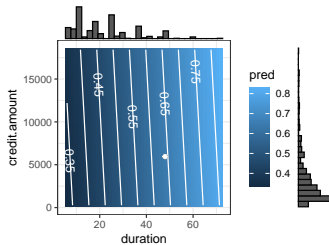


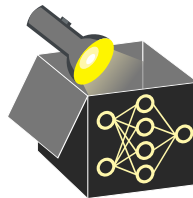
# Interpretable Machine Learning

## LIME Examples



### Learning goals

- See real-world data examples
- See application to image and text data



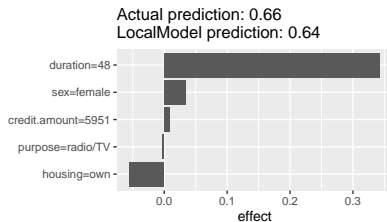
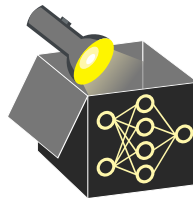
# EXAMPLE ON CREDIT DATASET (TABULAR)



- Model: SVM with RBF kernel
- $\mathbf{x}$ : first data point of the dataset with  $\hat{f}_{bad}(\mathbf{x}) = 0.658$
- $\mathbf{z}$ : training data  $\rightsquigarrow$  weighted by the Gower proximity
- Surrogate model  $\hat{g}$ :  $L_1$ -regularized linear model with 5 features

age	sex	job	housing	saving	checking	credit.amount	duration	purpose
22	female	2	own	little	moderate	5951	48	radio/TV

# EXAMPLE ON CREDIT DATASET (CONT'D)

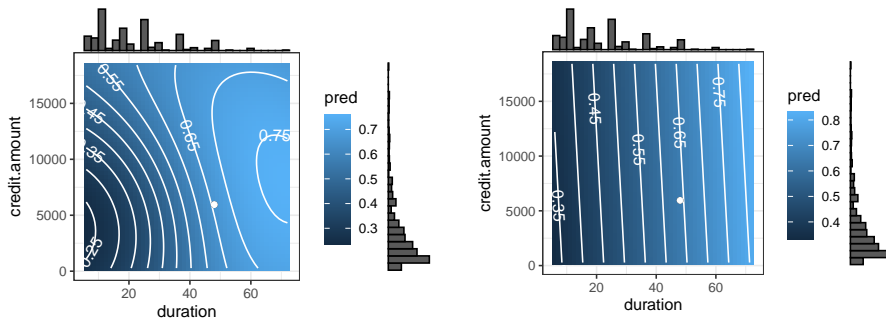


Effects of surrogate model, i.e.  $\hat{\theta}^T \mathbf{x}$

- The local model prediction for  $\mathbf{x}$  is  $\hat{g}(\mathbf{x}) = 0.64$  vs.  $\hat{f}(\mathbf{x}) = 0.658$
- $\hat{g}$  has a local fidelity of  $L(\hat{f}, \hat{g}, \phi_{\mathbf{x}}) = 4.82$  with  $\phi_{\mathbf{x}}(\mathbf{z})$  as the Gower proximity and  $L(\hat{f}_{bad}(\mathbf{z}), g(\mathbf{z}))$  as the euclidean distance

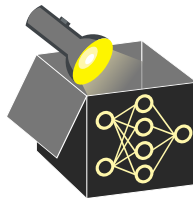
# EXAMPLE ON CREDIT DATASET (CONT'D)

- 2-dim ICE plots (aka. prediction surface plot) of credit amount and duration show how the surrogate model  $g$  linearly approximates the previously nonlinear prediction surface of  $\hat{f}_{bad}$



2-dim ICE plot of  $\hat{f}_{bad}$  (left) and surrogate  $g$  (right) for features duration and credit amount.

The white dot is  $\mathbf{x}$ . The histograms display the marginal distribution of the training data  $\mathbf{X}$ .



LIME can also be applied to text data:

- Raw text representations:
  - Binary vector indicating the presence or absence of a word
  - A vector of word counts
- Examples for *"This text is the first text."* and *"Finally, this is the last one."*:

this	text	is	the	first	finally	last	one
1	2	1	1	1	0	0	0
1	0	1	1	0	1	1	1

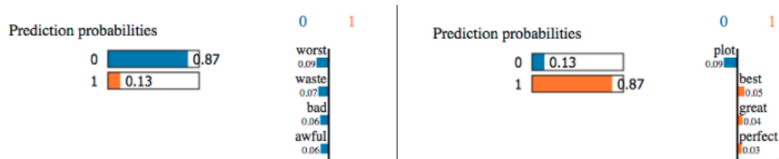
- **Sampling:** Randomly set the entry of individual words to 0; equal to removing all occurrences of this word in the text.
- **Proximity:** Exponential kernel with cosine distance.
  - Neglects words that do not occur in both texts
  - Measures the distance irrespective of the text size

# LIME FOR TEXT DATA (CONT'D)

► Shen, Ian, (2019)



- Random forest classifier labeling movie reviews from IMDB
  - 0: negative
  - 1: positive
- Surrogate model is a sparse linear model



Words like “worst” or “waste” indicate negative review while words like “best” or “great” indicate positive review

# LIME FOR IMAGE DATA

LIME also works for image data:

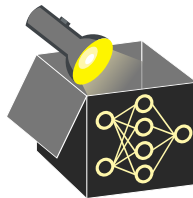
- **Idea:** Each obs. is represented by a binary vector indicating the presence or absence of superpixels

► Achanta et al. 2012

- Superpixels are interconnected pixels with similar colors (absence of a single pixel might not have a (strong) effect on the prediction)
- **Warning:** Size of superpixels needs to be determined before the segmentation takes place
- **Sampling:** Randomly switching some of the superpixels “off”, i.e., by coloring some superpixels uniformly



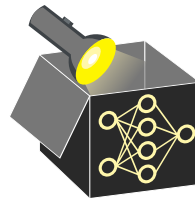
Example for  
superpixels of  
different sizes



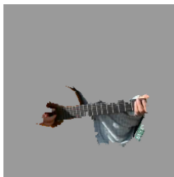
# LIME FOR IMAGE DATA (CONT'D)

► Ribeiro. 2016

- Explaining prediction of pre-trained inception neural network classifier
- **Sampling**: Graying out all superpixels besides 10 superpixels
- **Surrogate**: Locally weighted sparse linear models
- **Proximity**: Exponential kernel with euclidean distance



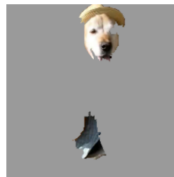
(a) Original Image



(b) Explaining *Electric guitar*



(c) Explaining *Acoustic guitar*



(d) Explaining *Labrador*

Top 3 classes predicted