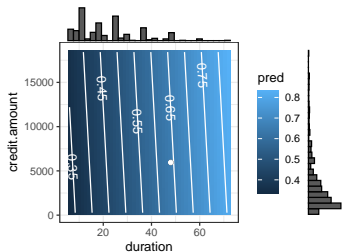
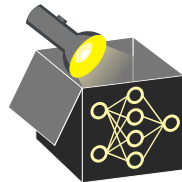


# Interpretable Machine Learning

## Local Explanations: Lime Examples



### Learning goals

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

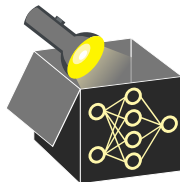
# EXAMPLE: CREDIT SCORING (TABULAR DATA)

- **Black-box model**  $\hat{f}_{bad}$ : SVM with RBF kernel (predicts probability of bad credit risk)

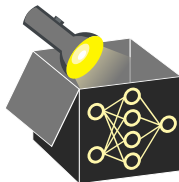
- **Instance to explain  $\mathbf{x}$** : First row in the dataset, with  $\hat{f}_{bad}(\mathbf{x}) = 0.658$

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

- **Surrogate model**: LASSO, restricted to 5 non-0 feats (via regularization)
- **Training data for surrogate**: Samples  $\mathbf{z}$ , weighted by Gower dist. to  $\mathbf{x}$



# EXAMPLE: CREDIT SCORING (TABULAR DATA)



- **Black-box model**  $\hat{f}_{bad}$ : SVM with RBF kernel (predicts probability of bad credit risk)

- **Instance to explain  $\mathbf{x}$** : First row in the dataset, with  $\hat{f}_{bad}(\mathbf{x}) = 0.658$

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

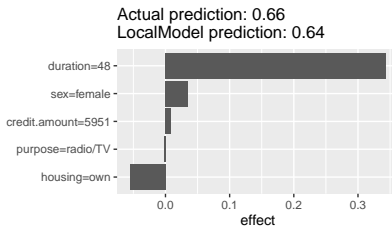
- **Surrogate model**: LASSO, restricted to 5 non-0 feats (via regularization)
- **Training data for surrogate**: Samples  $\mathbf{z}$ , weighted by Gower dist. to  $\mathbf{x}$

- **Prediction:**

$$\hat{g}(\mathbf{x}) = 0.640 \text{ vs. } \hat{f}_{bad}(\mathbf{x}) = 0.658$$

⇒  $\hat{g}$  provides good local approx. of  $\hat{f}_{bad}$ , but omits several features

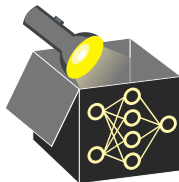
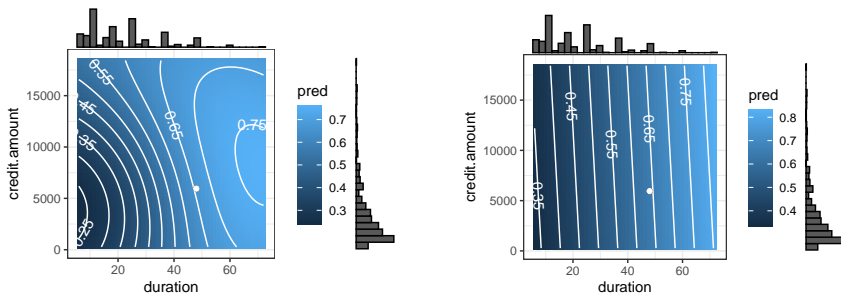
⇒ Small mismatch reflects trade-off: **interpretability vs. fidelity**



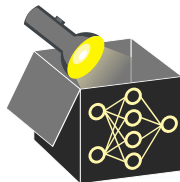
**Interpretation:** Prediction is mainly driven by loan duration, with small positive effect from sex and credit.amount, and negative contributions from housing and purpose.

# EXAMPLE ON CREDIT DATASET (CONT'D)

- 2D ICE plots (pred. surface plots) for duration and credit.amount
- Illustration how  $\hat{g}$  linearly approximates nonlinear decision surface of  $\hat{f}_{bad}$



- **Left:** 2D ICE plot of  $\hat{f}_{bad}$  showing decision surface
- **Right:** Linear approximation by surrogate model  $\hat{g}$ .
  - ~> White dot indicates input  $\mathbf{x}$  to be explained
  - ~> Histograms show marginal distribution of features in training data

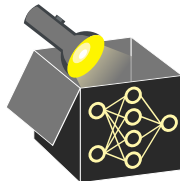


LIME can also be applied to text data:

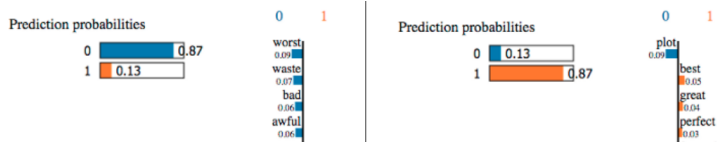
- Raw text representations:
  - Binary vector indicating the presence or absence of a word
  - A vector of word counts
- Examples for *“This text is the first text.”* and *“Finally, this is the last one.”*:

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



- 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

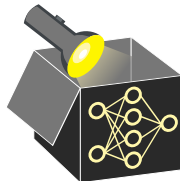
# LIME FOR IMAGE DATA

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



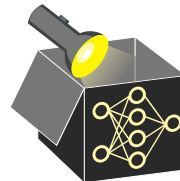
Example for  
superpixels of  
different sizes



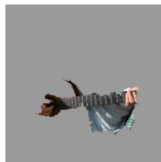
# LIME FOR IMAGE DATA (CONT'D)

► "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