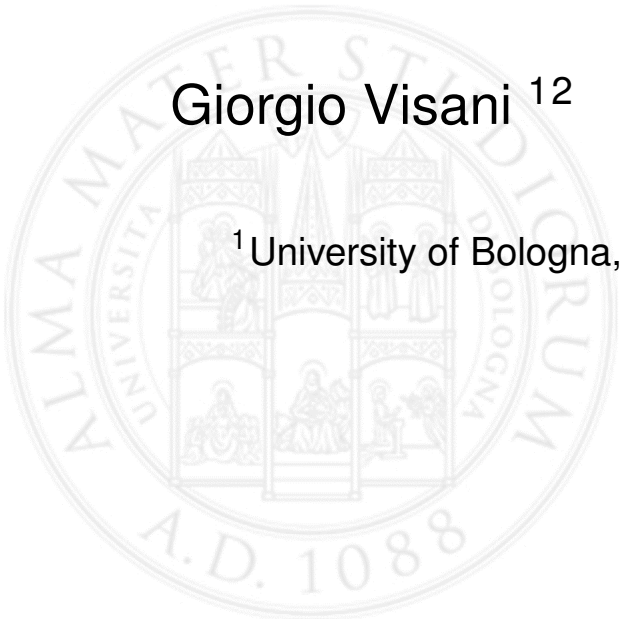
OptiLIME: Optimized LIME Explanations for Diagnostic Computer Algorithms



Giorgio Visani [1 2] Enrico Bagli [2] Federico Chesani [1]

[1] University of Bologna, Department of Computer Science & Engineering

[2] Crif S.p.A.

19th October 2020

LIME

Model agnostic, Local technique, developed in 2016 [5]

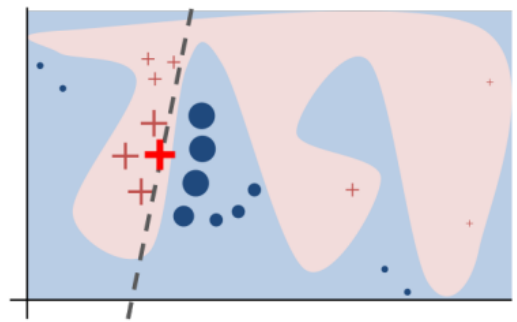
Objective: find the tangent plane to the Machine Learning (ML) function $f(\mathbf{x})$, in the point $(y_i, \mathbf{x}_i)$ we want to explain.

The tangent formula is human-understandable and it should be a good approximation for the ML function in the neighbourhood of $(y_i, \mathbf{x}_i)$ (Taylor Theorem)

Analytically unfeasible

- *don't have a parametric formulation of the ML function*
- *the ML surface may have a huge number of discontinuity points $\rightarrow$ non differentiable*

Solution: sample points on the ML surface, approximate the tangent with a linear model through the sampled points (Ridge Regression), in the neighbourhood of the reference individual.



LIME Issues I

Instability: Repeated explanations, with equal settings, may have different outcomes. Due to the generation step (*different sampled points each time*).

Tackled in *Visani et al., Statistical Stability Indices for LIME: Obtaining Reliable Explanations for Machine Learning Models, 2020*

Idea: Produce *n* LIME explanations (with the same parameters) and measure the difference between them. Two indices for that:

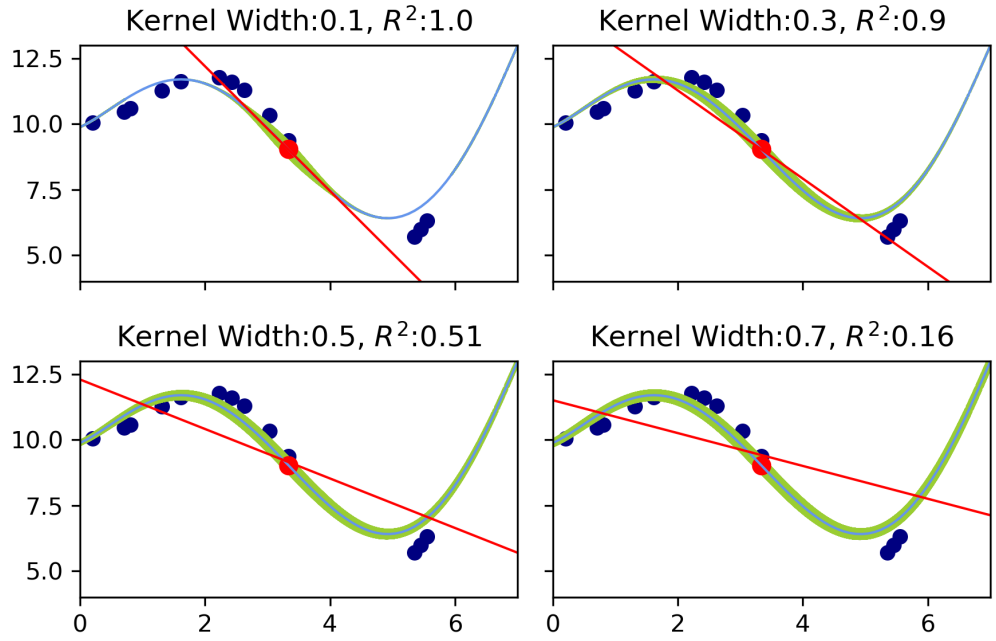
- **Variables Stability Index (VSI)**: measures whether the variables in the repeated LIME explanations are the same
- **Coefficients Stability Index (CSI)**: checks whether the coefficients are the same in the different LIMEs

LIME Issues II

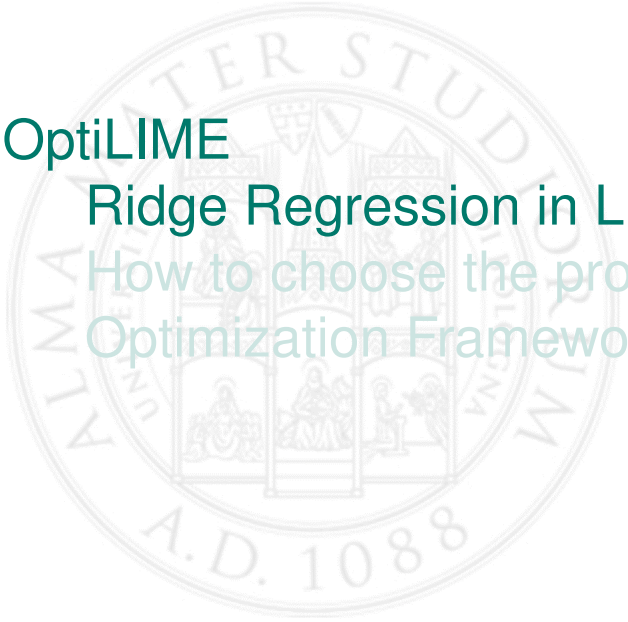
– Paper's focus –

Select the proper Kernel Width

- LIME tangent-like explanation is valid only for a small neighbourhood
- The neighbourhood width is application dependent
- LIME requires to set this beforehand, using the kernel width parameter



LIME explanations (red lines) for different kernel widths



A brief digression on Ridge Regression

Ridge Regression is just a linear model: $\mathbf{E}(Y) = \alpha + \sum_{j=1}^{d} \beta_j X_j$
But the coefficients are estimated using a penalty based on their norm:

$$\hat{\beta}_R = \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

The functions retrieved by Ridge and Linear Regression are different.

Ridge is useful with **noisy data**: the regularization helps finding a more stable model, but causes distortion in the coefficients.

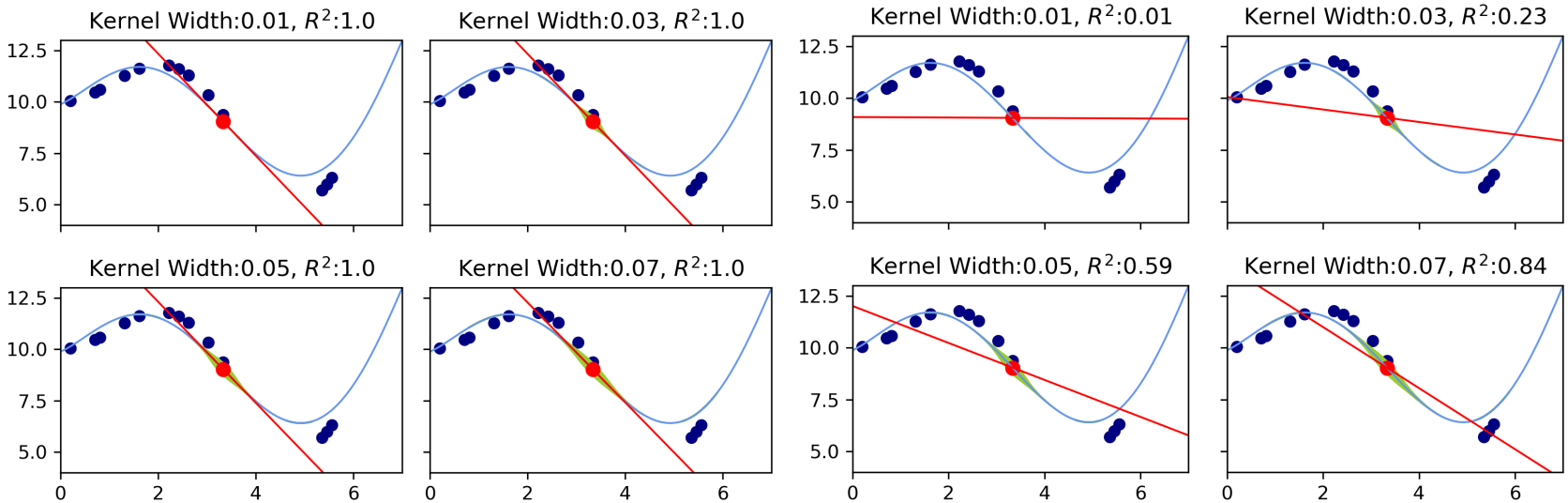
Intuition: Ridge Regression is harmful in LIME

Because we have no noise in the generated dataset (all points lie on the ML surface by definition: we generate the $\mathbf{x}$ *values, while the* y *ones are predicted by the ML function* $f(\mathbf{x})$*)*

Idea: Linear Regression is more suitable

Ridge Regression in LIME

Evident distortion for small kernel width and Ridge penalty,
Linear Regression is not affected



(a) Ridge Penalty $= 0$
(*Linear Regression*)

(b) Ridge Penalty $= 1$

In further analysis, Linear Regression as LIME explainable model

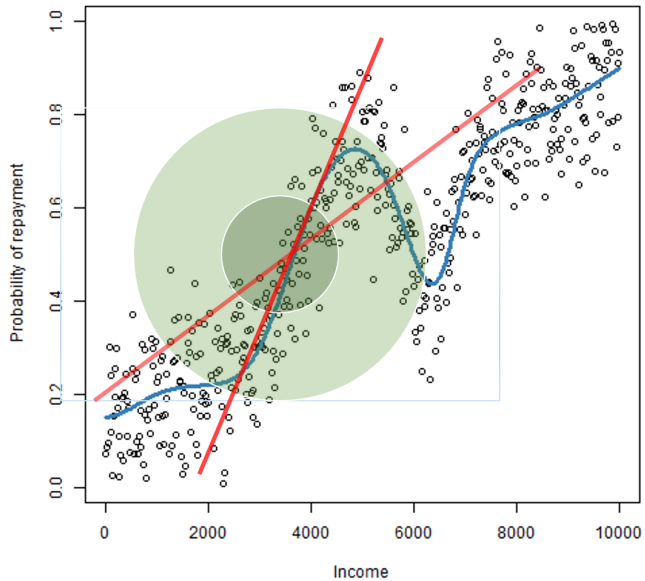
Select the Kernel Width

It is difficult to establish a fair distance measure, thus it is hard to define a meaningful neighbourhood around x.

LIME defines locality through a Gaussian (RBF) kernel: each unit has a weight in the range $[0, 1]$, the higher the closer to the reference point.

RBF Kernel Formula:

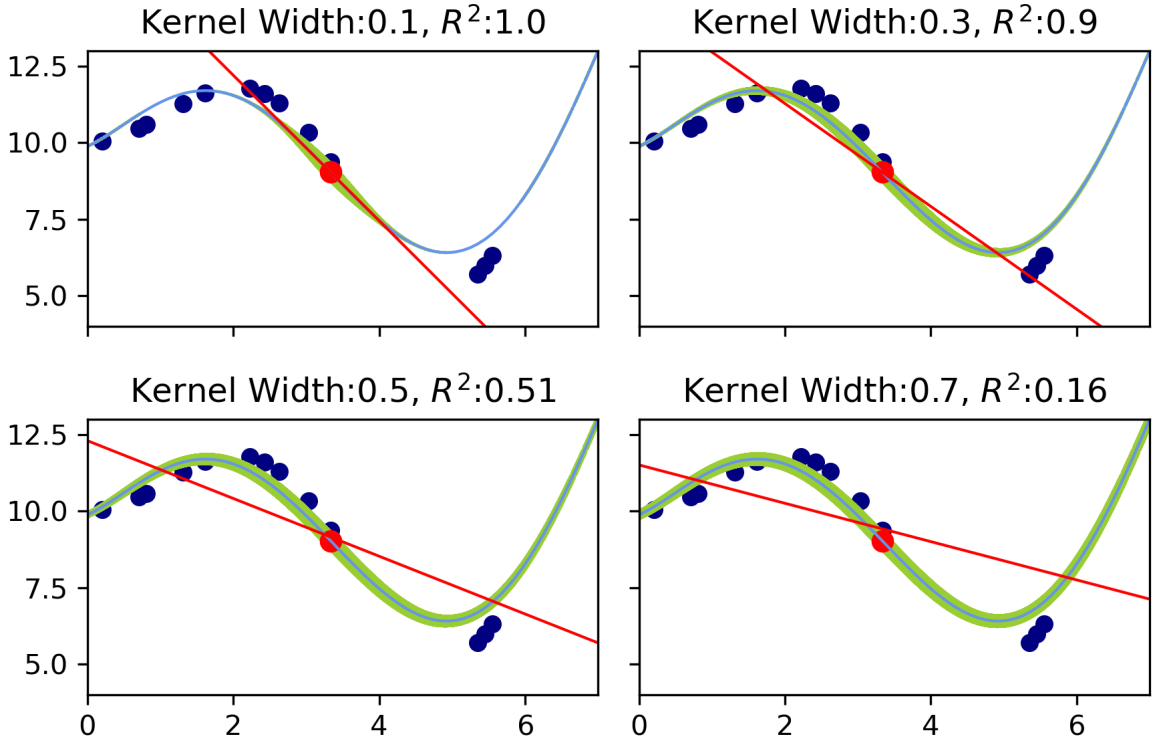
$$RBF(x^{(i)}) = \exp\left(\frac{||x^{(i)} - x^{(ref)}||^2}{kw}\right)$$



Kernel Width (kw) is the only free parameter, defines weights' radius

small Kernel Width $\rightarrow$ *high weights only to very close units,*

the weights tend to 0 very fast moving away from the point



Blue Dots: LIME generated points (size proportional to RBF weight)

Red Lines: LIME linear explanations

- *Small kw → significant weights only to the closest points, further ones do not contribute to the local linear model*
- *Large kw → far away regions of $f(\mathbf{x})$ are given meaningful weight, they contribute to LIME linear model*

- Don't want to consider regions in which $f(\mathbf{x})$ is wiggly.
- Linear approximation valid only in the proximity of the point
- Neighbourhood width depends on $f(\mathbf{x})$ shape

LIME Desiderata

When LIME explanations can be considered reliable?

Adherence: How much LIME linear model is close to the ML model, in the neighbourhood of the reference point.
We aim for a high adherence, namely the explanation makes sense (it represents its tangent). Measured by R^2

Stability: How much repeated LIME explanations are similar?
We hope to retrieve always very similar explanations, so to be able to trust them. Measured by CSI & VSI

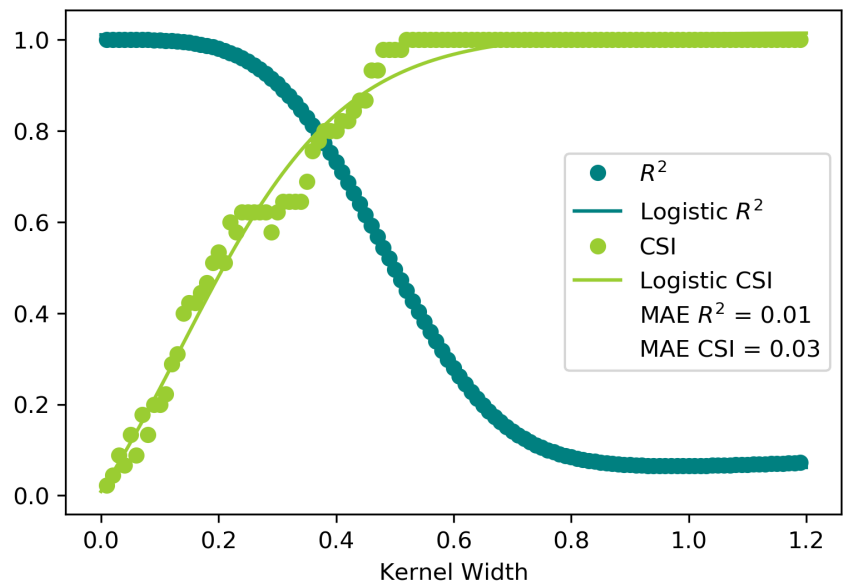
How to achieve both properties in a single LIME explanation?

Desiderata wrt Kernel Width

Trade-off between Stability and Adherence, controlled by the Kernel Width:

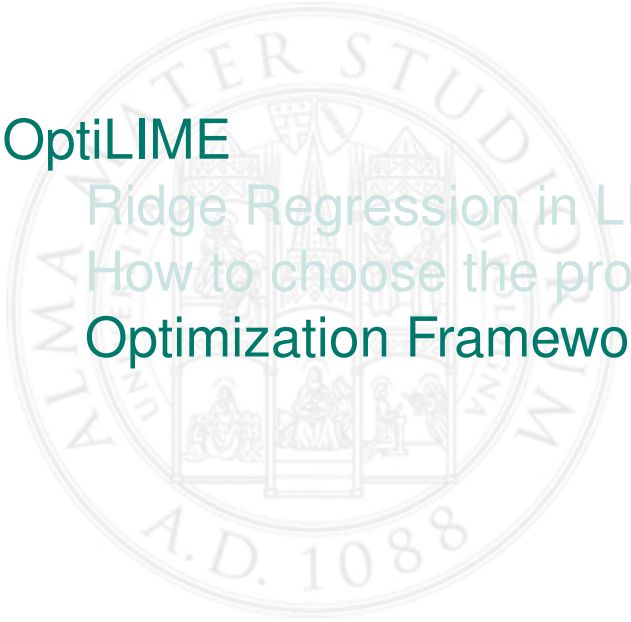
Small neighbourhood: ↑ adherence, ↓ stability
Large neighbourhood: ↑ stability, ↓ adherence

Sustained by Taylor Theorem [1] (adherence), variance of the linear coefficients [4] (stability)



Adherence and Stability curves are Monotonous noisy functions

Noisy: each point derives from a different LIME, hence different generated points



OptiLIME

Objective: Find the best kernel width, which ensures reasonable values for both Adherence and Stability

Key considerations:

- Joint maximisation of two noisy functions
- The monotonicity guarantees the Adherence and Stability to be proportional to the Kernel Width
- The trade-off makes impossible to achieve the highest values for both properties

Method

We consider Adherence is the most important property

Ensures LIME model to be very close to the real unknown tangent of the ML surface

Intuition: Choose the Kernel Width, obtaining satisfactory values for the Adherence. The Stability value will be already maximised given the Adherence constraint *because of the trade-off*

How to find the Kernel Width,
ensuring a given value of Adherence?

OptiLIME Framework

- Choose a predefined level of Adherence $\tilde{R}^2$
- Consider the function

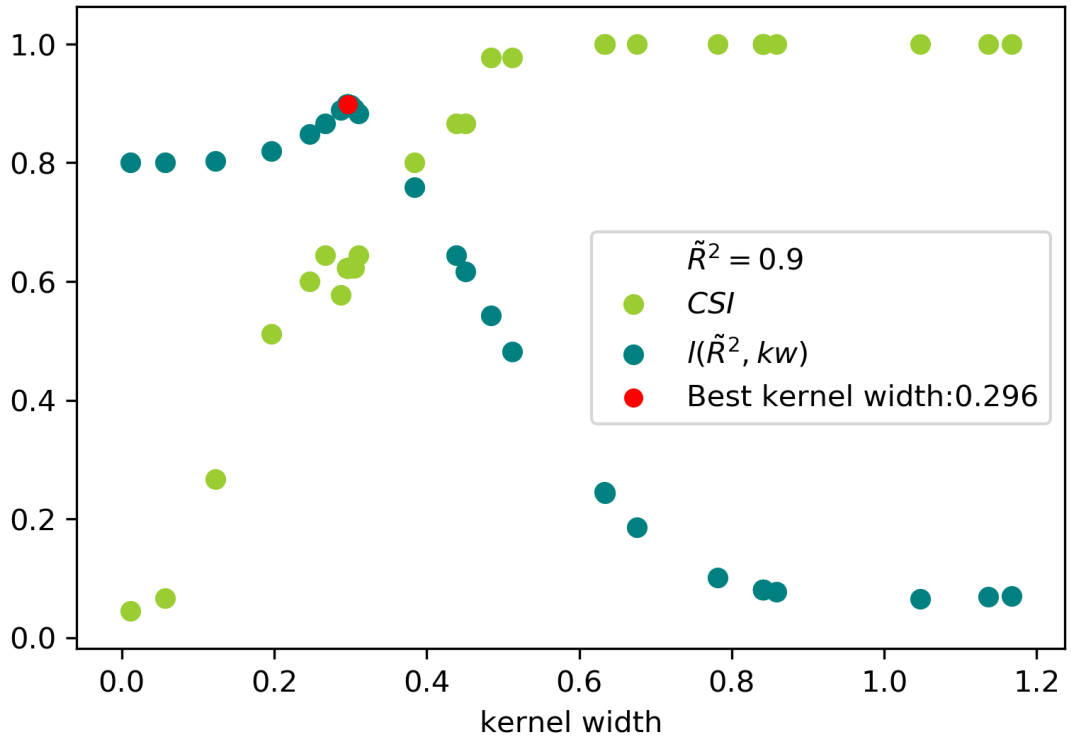
$$I(kw, \tilde{R}^2) = \begin{cases} R^2(kw), & \text{if } R^2(kw) \leq \tilde{R}^2 \\ 2\tilde{R}^2 - R^2(kw) & \text{if } R^2(kw) > \tilde{R}^2 \end{cases}$$

- Use Bayesian Optimisation to obtain $\widehat{kw} = \arg max_{kw} I(kw, \tilde{R}^2)$
- Build the final LIME explanation with the proper kernel width $\widehat{kw}$

Tips:

- Using $I(kw, \tilde{R}^2)$ the problem translates into a maximum search
- The maximum of $I(kw, \tilde{R}^2)$ has value $\tilde{R}^2$
- Noisy functions are easily optimised via Bayesian Optimization

OptiLIME in practice



Points are the distinct evaluations performed by the Bayesian Search. Parameters: $p = 20, m = 40$

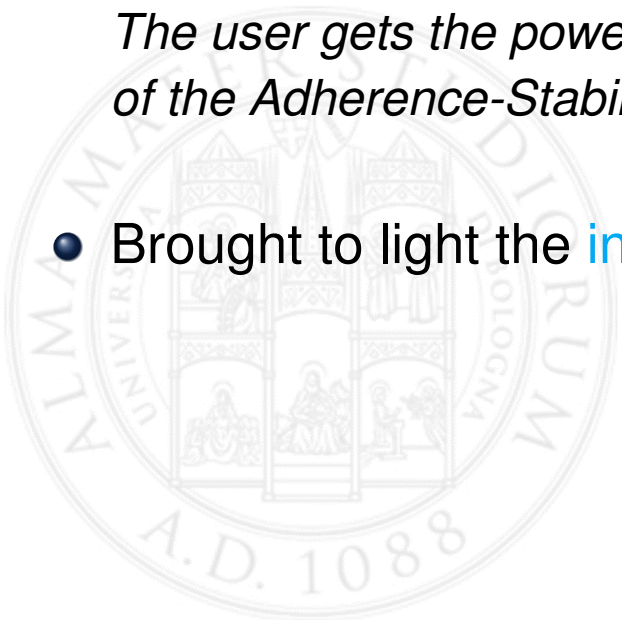
Bayesian Search parameters:

- p: preliminary calls with random kw values
- m: search refinement strategy iterations

We may tweak the parameters to achieve faster but less accurate convergence to the maximum

Conclusions: OptiLIME Achievements

- Sound method to select the proper kernel width
  The user gets the power to choose in which point of the Adherence-Stability trade-off to sit
- Brought to light the inadequacy of Ridge Regression in LIME



References I

[1] Patrick Billingsley. *Probability and Measure*. John Wiley & Sons, 2008. ISBN: 81-265-1771-9.

[2] Roberto Calandra et al. ≪Bayesian Gait Optimization for Bipedal Locomotion≫. In: *International Conference on Learning and Intelligent Optimization*. Springer, 2014, pp. 274–290.

[3] Christine S. Cox. *Plan and Operation of the NHANES I Epidemiologic Followup Study, 1987*. 27. US Department of Health and Human Services, Public Health Service, Centers . . ., 1992.

[4] William H Greene. *Econometric Analysis*. Pearson Education India, 2003. ISBN: 81-7758-684-X.

[5] Marco Tulio Ribeiro, Sameer Singh e Carlos Guestrin. ≪Why Should i Trust You?: Explaining the Predictions of Any Classifier≫. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016, pp. 1135–1144. ISBN: 1-4503-4232-9.

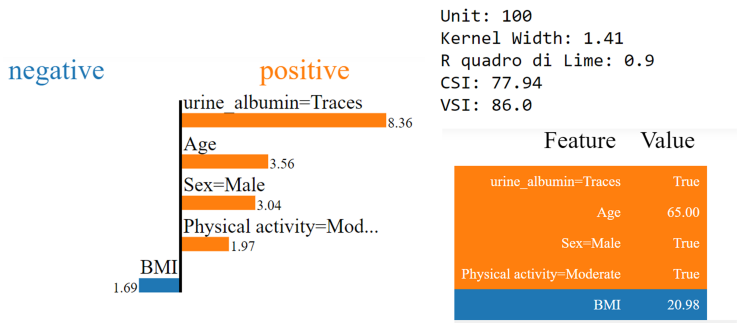
Appendix

OptiLIME Application to Medical Data

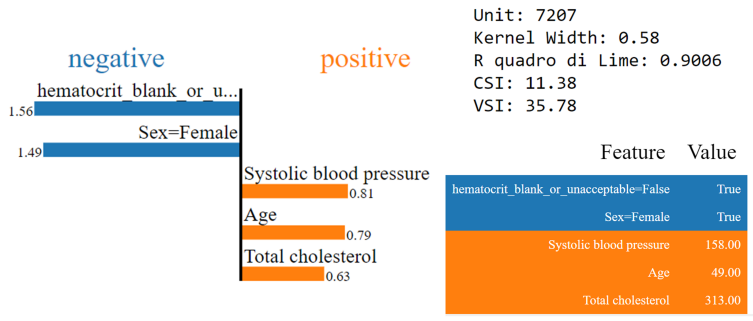
Data: NHANES I dataset [3]

Model: Survival XGBoost

estimate the risk of death over 20 years



(c) Unit 100

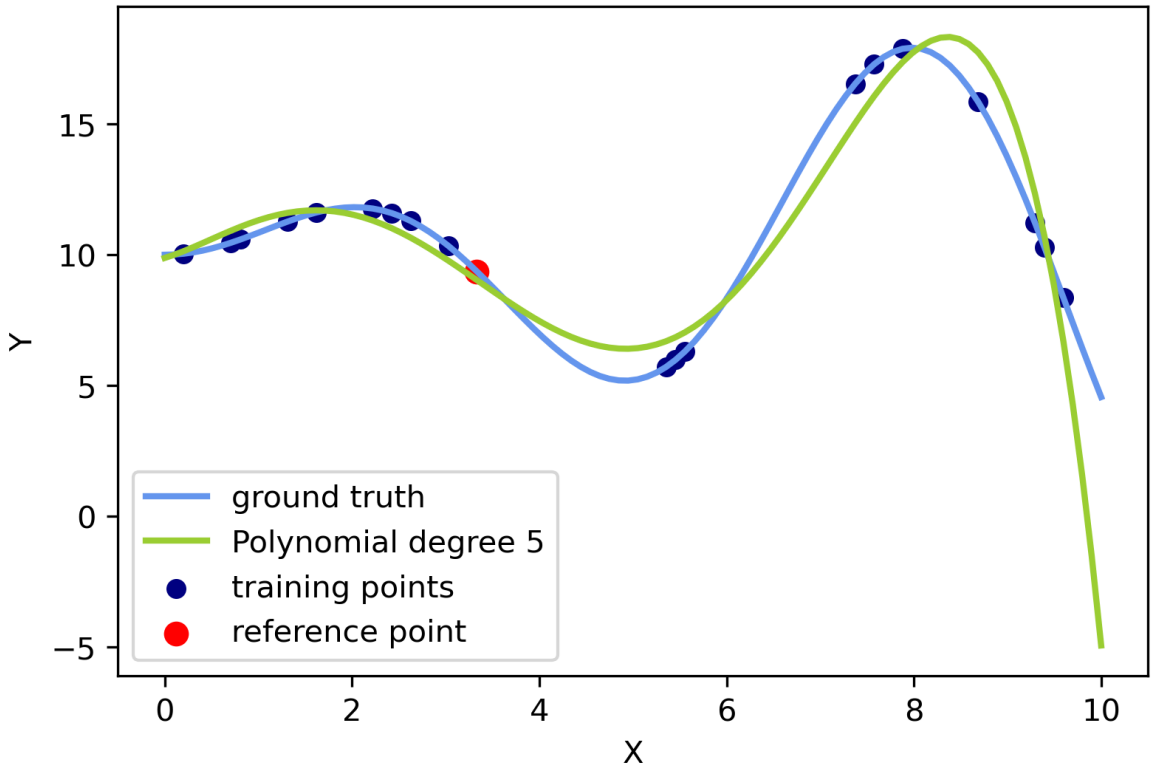


(d) Unit 7207

NHANES individual Explanations using OptiLIME

Toy Dataset

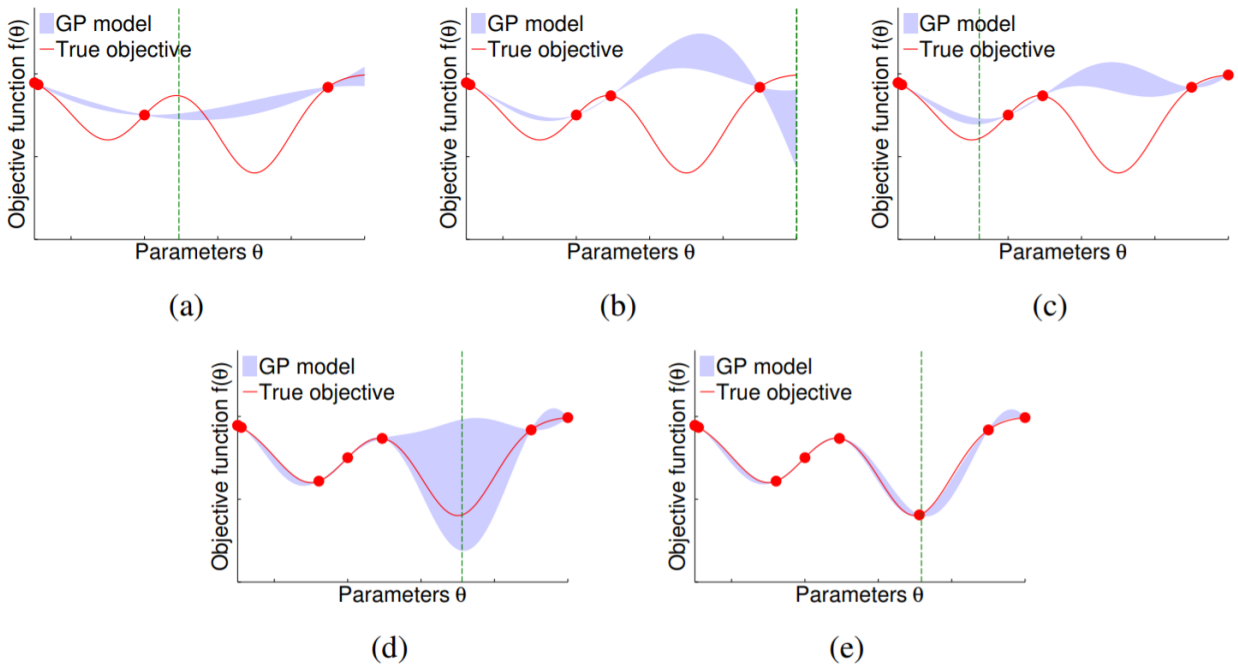
Dataset, with just a single X variable
To graphically show the LIME behaviour



Blue line is the True DGP function, blue points are the dataset, green line is the model to be explained (Polynomial Regression), red dot is the reference point

Bayesian Optimization

Bayesian Optimization approximates the function to be optimized, using a Gaussian Process (GP) model



Source: [2]

Firstly, it performs some function evaluations for random parameter values. GP model tries to interpolate these points, approximating the function. At each iteration, a new evaluation is done. The value is chosen in order to: minimize the GP variance or explore relatively unknown areas.