EEP 596: AI and Health Care || Lecture 14

Dr. Karthik Mohan

Univ. of Washington, Seattle

June 3, 2022

Last Lecture

1. Explainable and Interpretable models

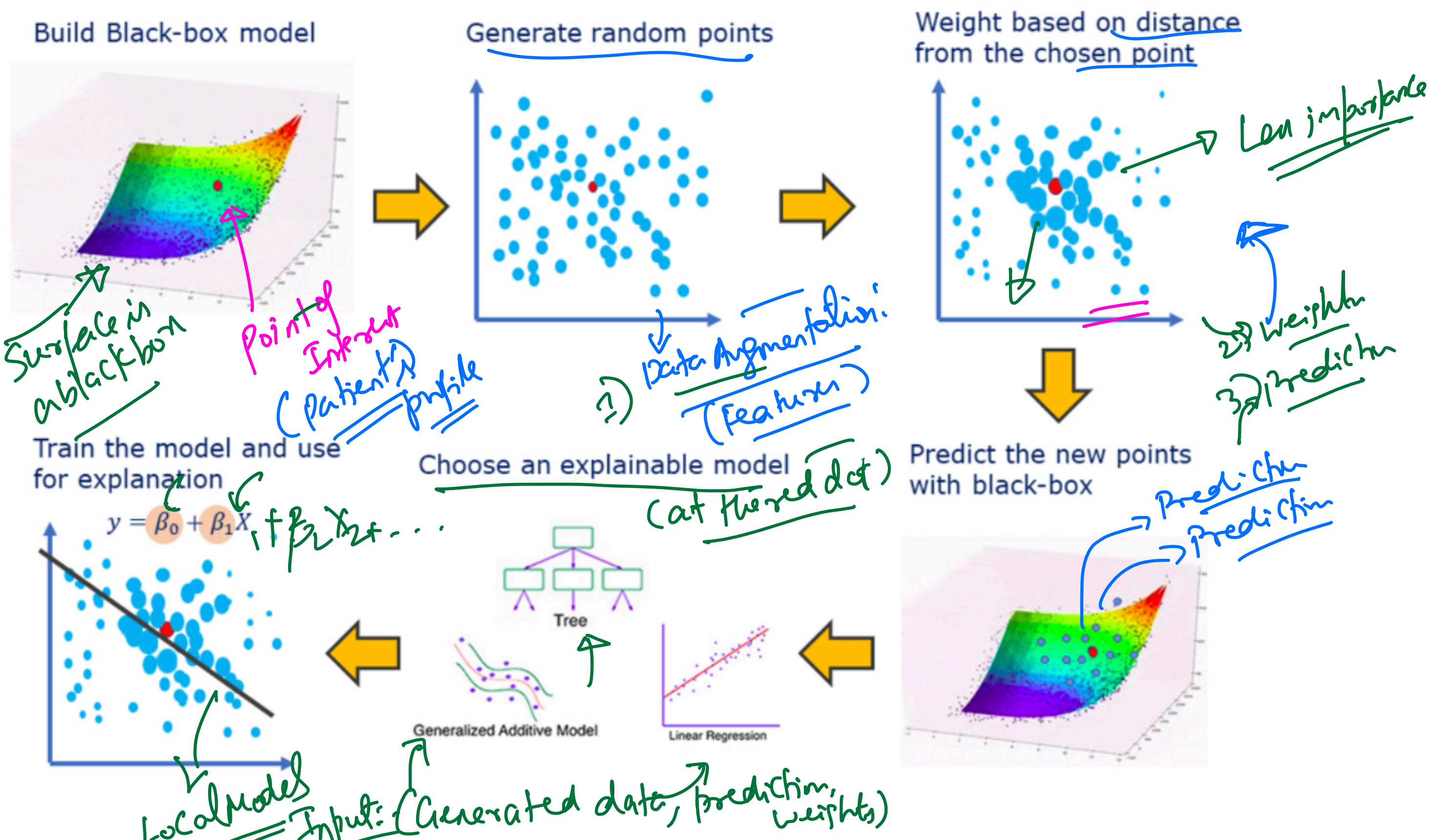
2) used to explain a more complicated black box model

1) Interpretable! easy to explain

Today's Lecture

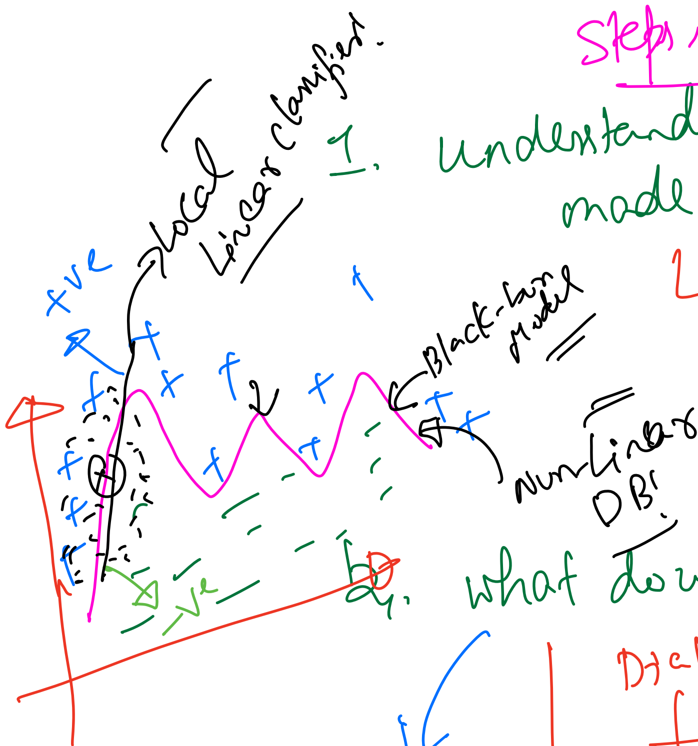
1. More on Explainable models
2. Visualization tools for deep learning
3. Wrap up

LIME - Model Agnostic Explainable Framework



Steps of the LIME algorithm. Picture by Giorgio Visani

Steps in Explainability



1. understanding why a black-box made a prediction
   ↳ Subset of features/inputs caused this prediction
   — LIME, DT

2. what do we do now? LIME →

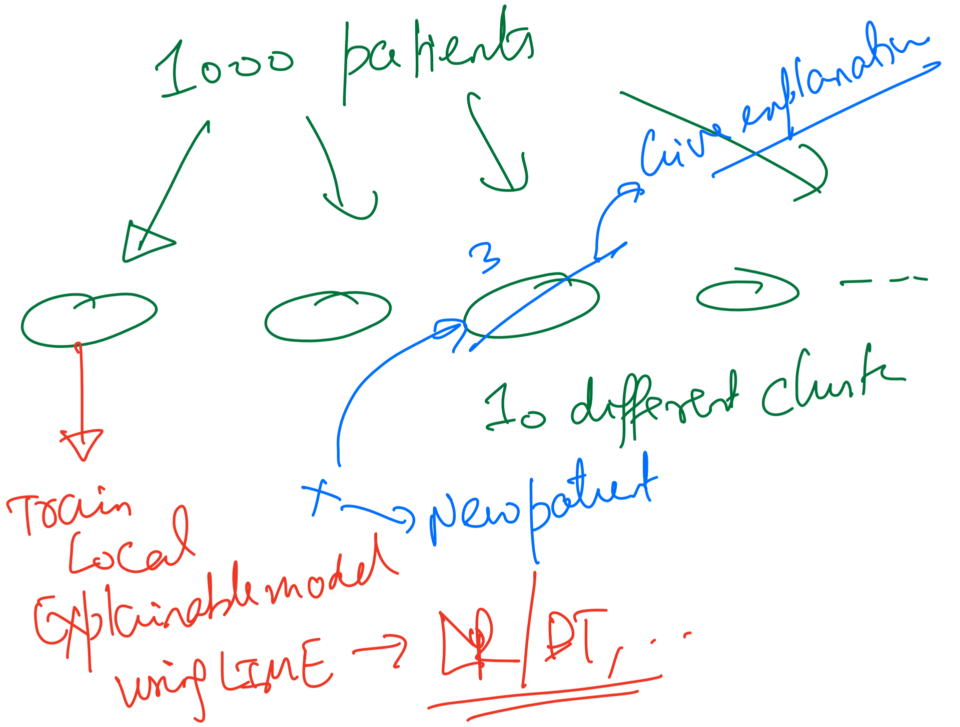
{ Counter-factual Analysis
↳ what if?

Diabetes
— Sugar
— Factors — Thickness
— BP

Assumptions

{ { { 1. Black-box model is accurate & accurate to interpolation (F-score or AUC or ...)

2. Local Linear models are good approx to black-box


1000 patients
Give explanation
3
10 different clusters
Train
Local
Explainable model
using LIME → LR / DT, . . .
x → New patient

LIME

L I M E

Local Interpretable Model-Agnostic Explanations

1. **Local**

LIME

Local Interpretable Model-Agnostic Explanations

1. **Local**
2. **Model Agnostic**

Local Interpretable Model-Agnostic Explanations

1. Choose the ML model and a reference point to be explained

Blackbox / DL

patient profile

LIME

Local Interpretable Model-Agnostic Explanations

1. Choose the ML model and a reference point to be explained
2. Generate points all over the space (sample X values from a Normal distribution inferred from the training set)

LIME

Local Interpretable Model-Agnostic Explanations

1. Choose the ML model and a reference point to be explained
2. Generate points all over the space (sample X values from a Normal distribution inferred from the training set) (weighted)

   (X)
3. Predict the Y coordinate of the sampled points, using the ML model (the generated points are guaranteed to perfectly lie on the ML surface)

   (y) = $\hat{y}$ → blackbox prediction

   target for Local model!

LIME

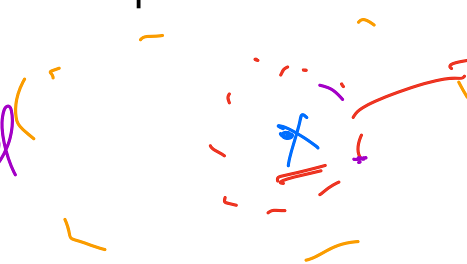
Local Interpretable Model-Agnostic Explanations

1. Choose the ML model and a reference point to be explained
2. Generate points all over the space (sample X values from a Normal distribution inferred from the training set)
3. Predict the Y coordinate of the sampled points, using the ML model (the generated points are guaranteed to perfectly lie on the ML surface)

   wx in loss fn.

4. Assign weights based on the closeness to the chosen point (use RBF Kernel, it assigns higher weights to points closer to the reference)

Weights → Soft thresholding of data for Local model



→ Most importance for training Local

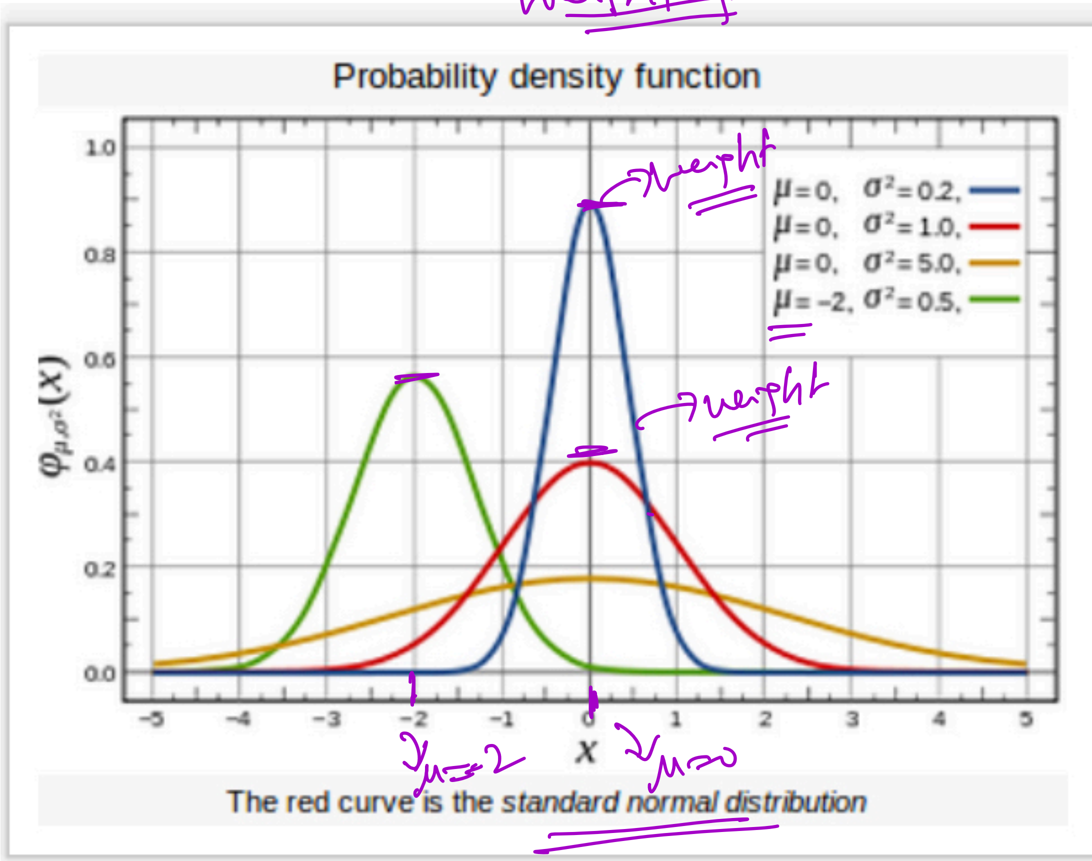
LIME

Local Interpretable Model-Agnostic Explanations

1. Choose the ML model and a reference point to be explained
2. Generate points all over the space (sample X values from a Normal distribution inferred from the training set)
3. Predict the Y coordinate of the sampled points, using the ML model (the generated points are guaranteed to perfectly lie on the ML surface)
4. Assign weights based on the closeness to the chosen point (use RBF Kernel, it assigns higher weights to points closer to the reference)
5. Train Regression or Classifier model on the weighted data set.

RBF kernel

Weight fn



The red curve is the *standard normal distribution*

σ → widening the net

μ → patient profile that you want to explore

$$f(\underline{x}) = e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

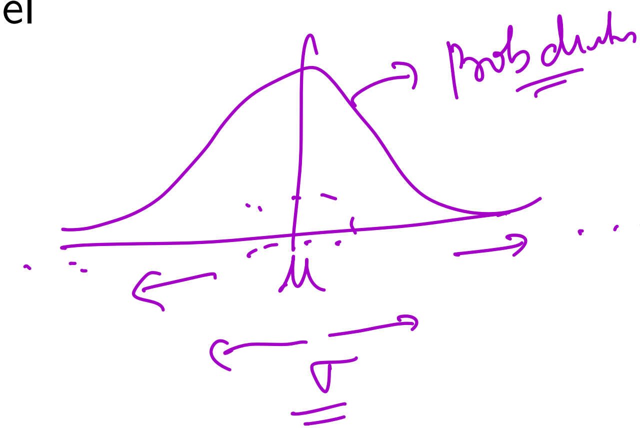
mean

std/variance

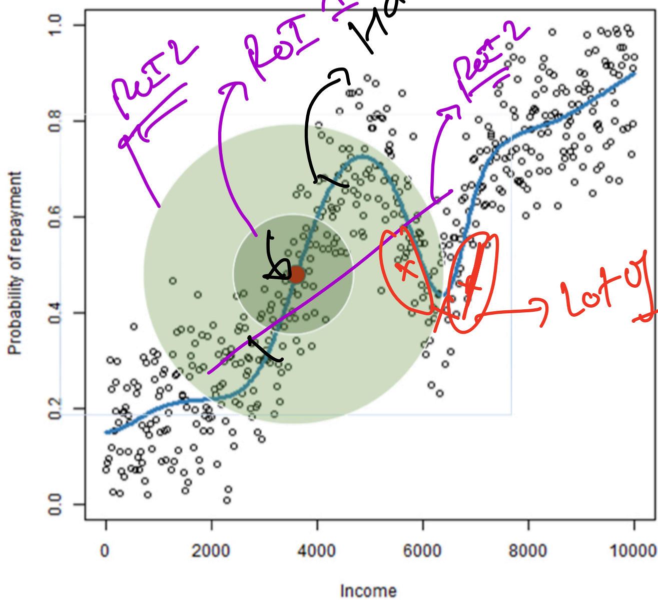
ICE #1

What would be an equivalent to weighting data in LIME, so the model is based on local data points?

a) Regularize the objective
b) Cut off data outside a bounding box of the point of interest
c) Sample according to the RBF kernel
d) None of above



LIME



White dots are the original dataset points, red dot is the reference point and the blue line is the prediction function of the ML model. Green circles show how the kernel weights are assigned, based on the kernel width parameter: the inner circle gives meaningful weights only to very close units because kw is low, the outer circle employs a larger kw. Picture by the author

RoI 2

RoI 1

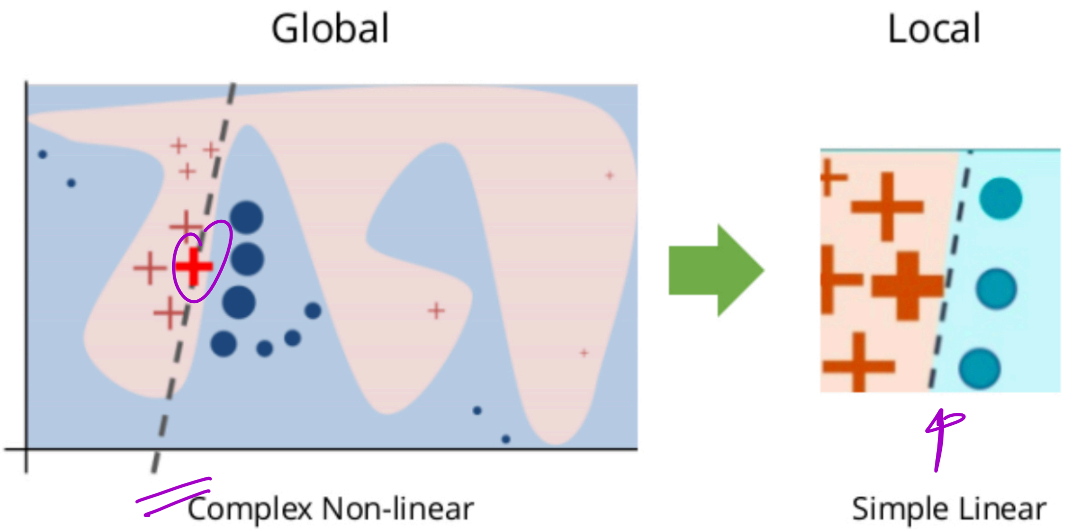
Have a larger σ

RoI 2

Lot of nonlinearity (smaller σ?)

Sensitivity analysis to understand / pick σ

LIME



LIME Idea: approximate the tangent to the curve. To understand the shape of the ML function LIME generates points around the red cross (we have an idea of the boundary f(**x**) thanks to the colors of the generated points).
Credits to Joseph

ICE #2

What would be a good way to pick σ for a point of interest x?

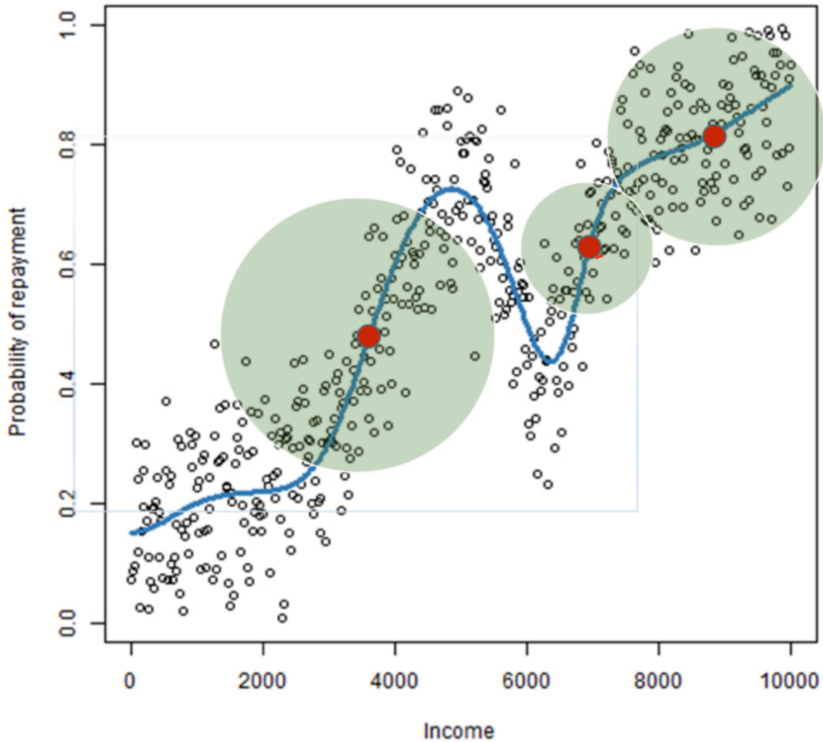
a) Fix it apriori e.g. $\sigma = 2$ as things get weighted down anyway ✗

b) Pick based on sensitivity of objective function to the feature space locally

c) Pick it based on data density in training set ✗

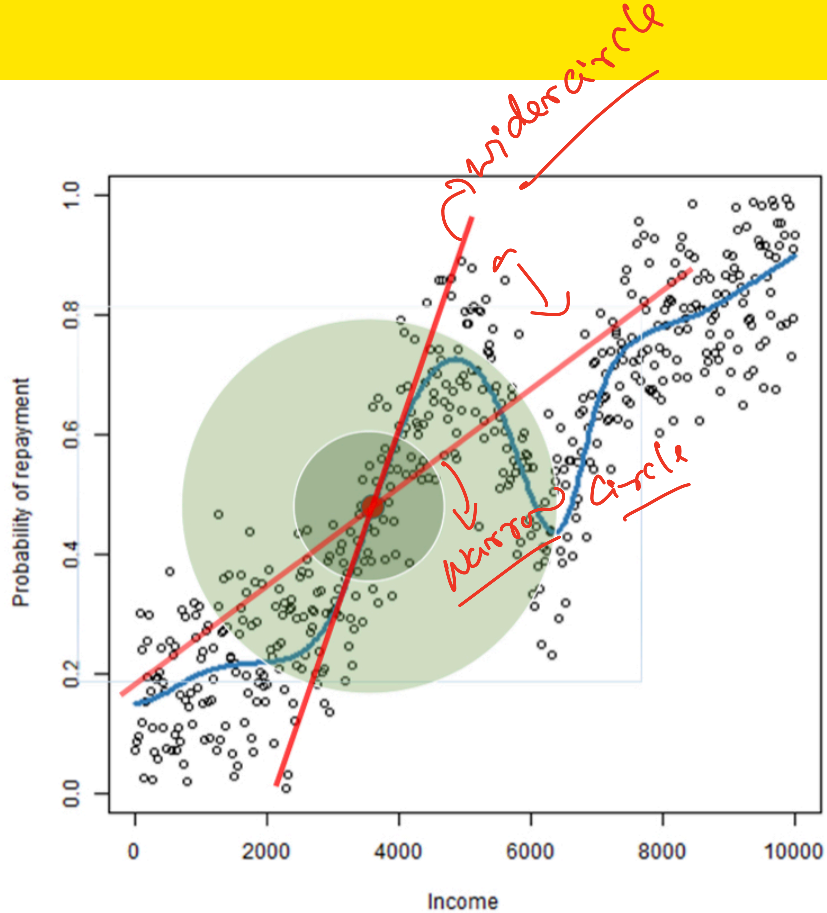
d) Doesn't really matter

LIME



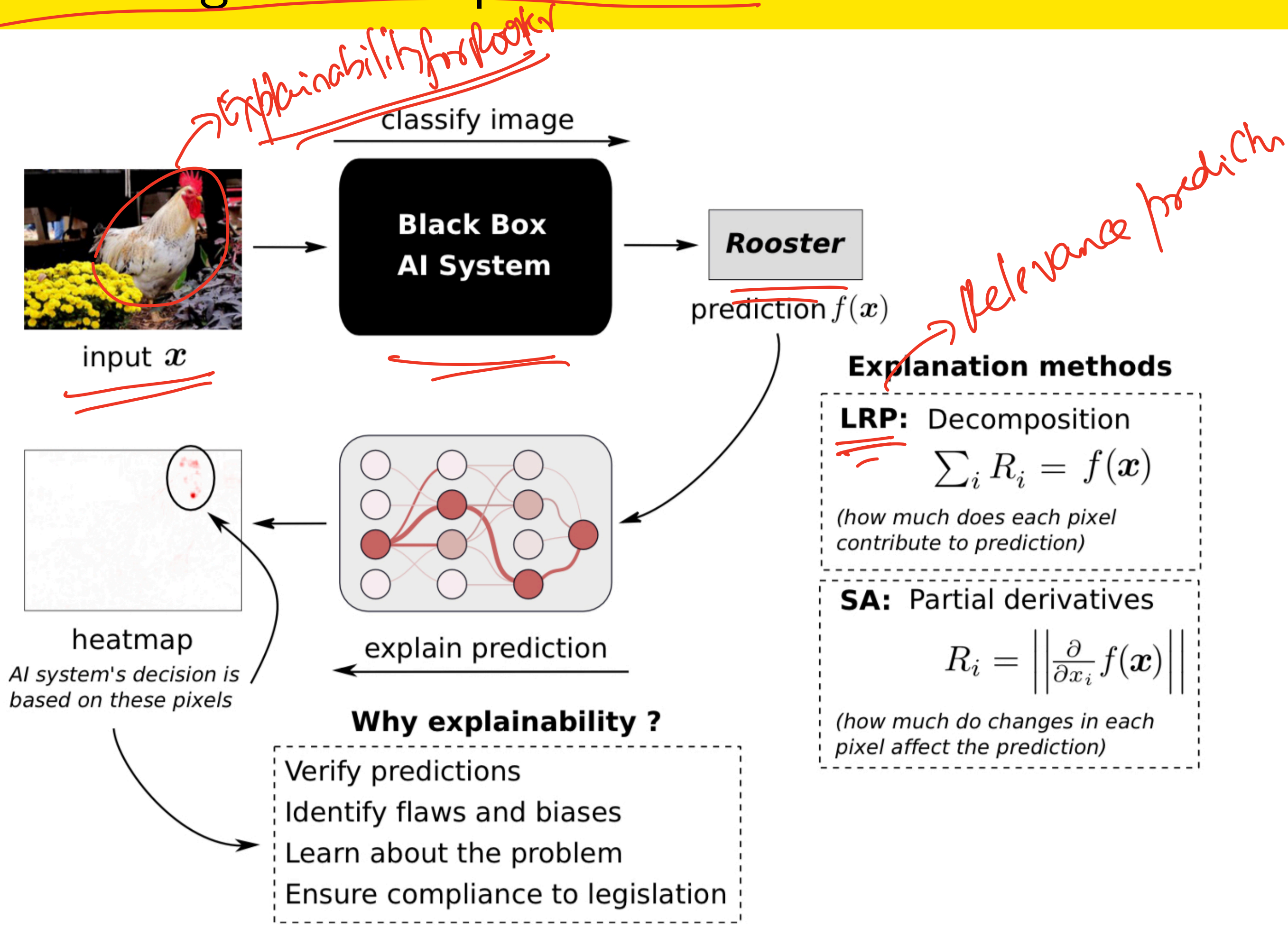
The best neighborhood size depends on the reference point and the curvature of the ML function around it.
Picture by the author

LIME

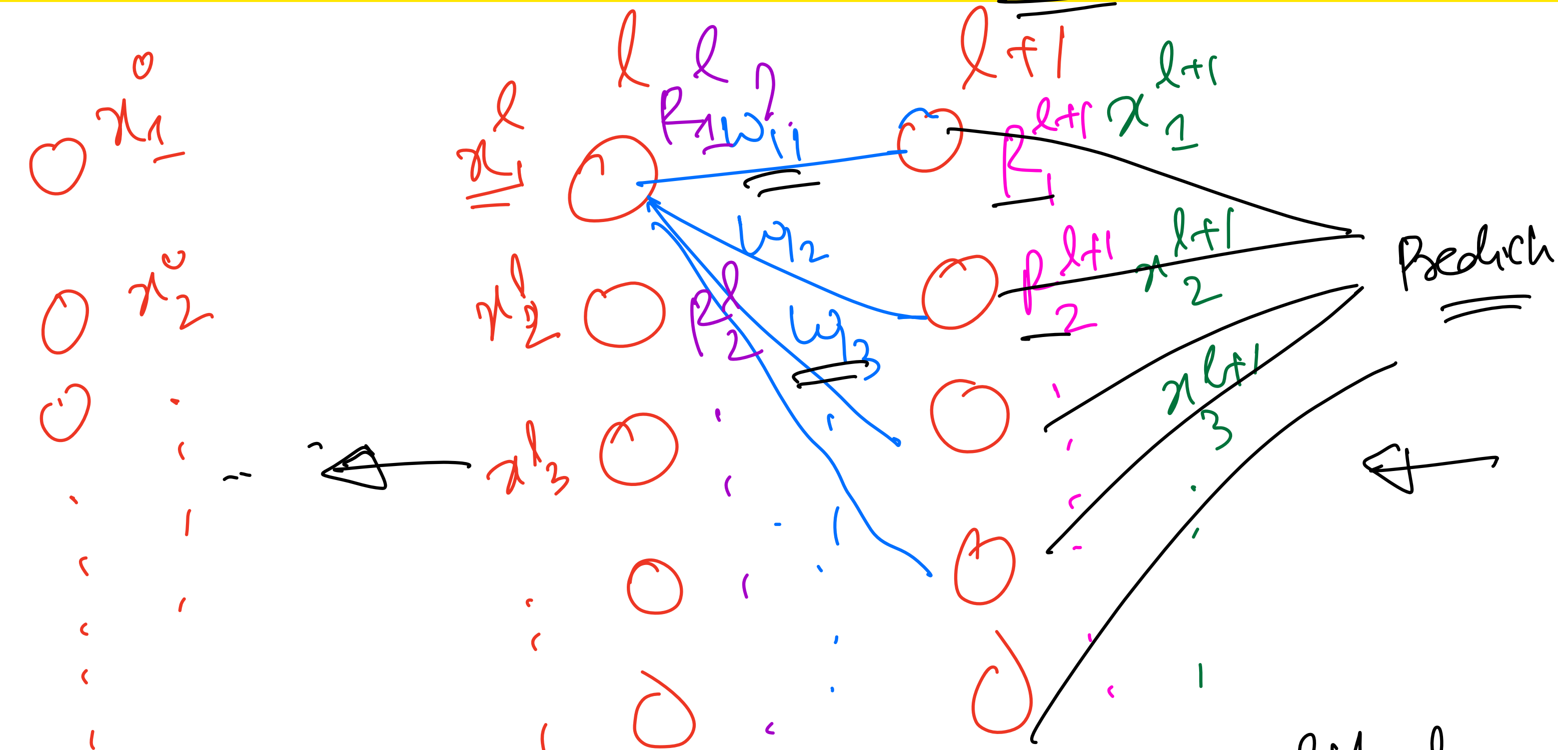


LIME explanations using different kernel width values. Picture by the author

Deep Learning based Explainable Models



Layer based Relevance Propagation (LRP)



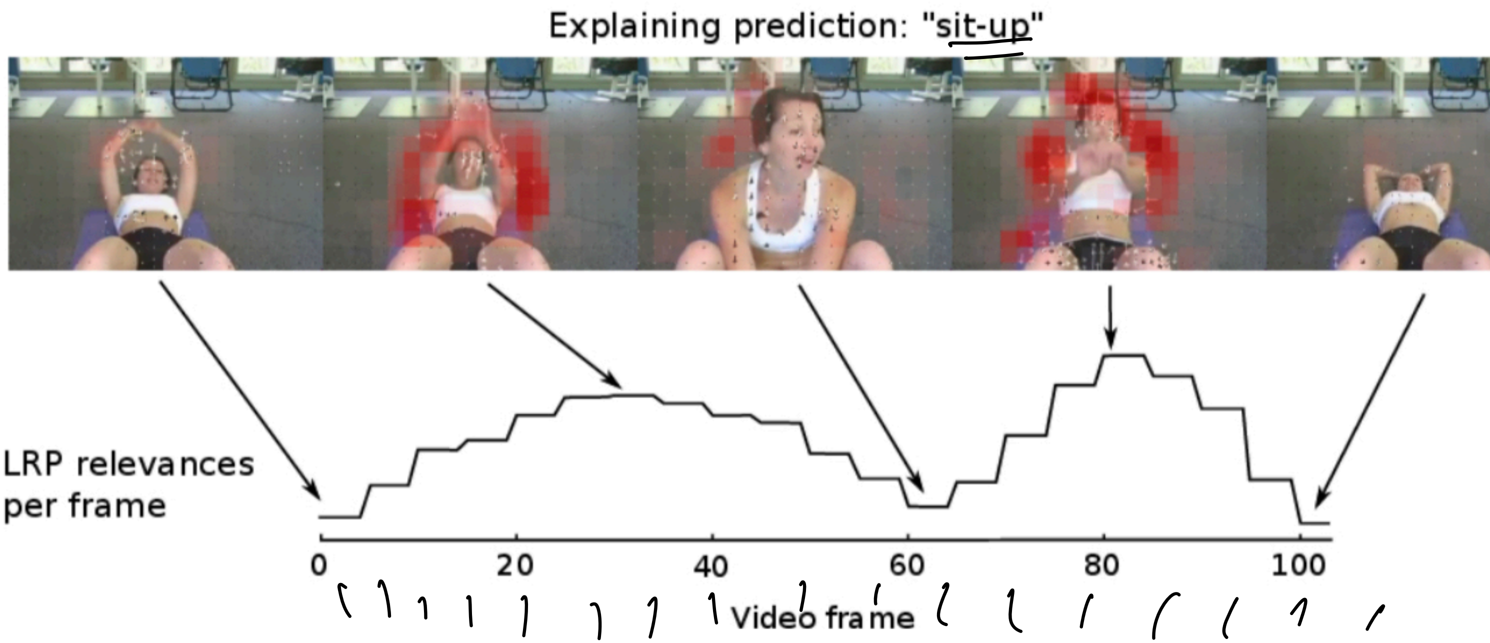
Input Layer (Image)

Pixels

$$R^l_1 = \frac{\sum_k w_{1k} R^{l+1}_k x^l_1}{\sum_k w_{1k} x^l_1}$$

Deep Learning based Explainable Models

Video → Prediction



Assignment 3

Imaging! | DTree for d[illegible]

1. Black box model will be a DL method on a health data set (e.g. cancer prediction, etc) as specified in the assignment
2. Choices to try different types of explainable models including feature importance, DTs and LIME
3. Layer based relevance propagation method for understanding DL feature importance (LRP) (Imaging Dataset)

Wrap up

1. Health care issues: Misdiagnosis, early diagnosis issues, mortality rates, lowering costs, lowering time to diagnosis, digital scribe, explaining diagnostics and treatments, preventative health care (wearables)

Wrap up

1. Health care issues: Misdiagnosis, early diagnosis issues, mortality rates, lowering costs, lowering time to diagnosis, digital scribe, explaining diagnostics and treatments, preventative health care (wearables)
2. Models for health care - Classification, Auto Encoders, Deep Learning for Health care, Data augmentation methods, Data Summarization methods and Topic modeling, Explainable models, etc

Wrap up

1. Health care issues: Misdiagnosis, early diagnosis issues, mortality rates, lowering costs, lowering time to diagnosis, digital scribe, explaining diagnostics and treatments, preventative health care (wearables)
2. Models for health care - Classification, Auto Encoders, Deep Learning for Health care, Data augmentation methods, Data Summarization methods and Topic modeling, Explainable models, etc
3. Data sets: wearables, disease prediction, imaging data, document summarization, etc

Wrap up

1. Health care issues: Misdiagnosis, early diagnosis issues, mortality rates, lowering costs, lowering time to diagnosis, digital scribe, explaining diagnostics and treatments, preventative health care (wearables)
2. Models for health care - Classification, Auto Encoders, Deep Learning for Health care, Data augmentation methods, Data Summarization methods and Topic modeling, Explainable models, etc
3. Data sets: wearables, disease prediction, imaging data, document summarization, etc (data augmentation)
4. Model biases, Data biases, data collection issues, interpretability vs explainability, adoption of models in health care (FDA approvals, etc)

References

1. Blog on LIME Explainable Model
2. Explainability for artificial intelligence in health care: a multidisciplinary perspective
3. Explainable AI model to predict acute illness from EHR
4. Readmission Risk assessment