Friedrich-Alexander-Universität
Technische Fakultät

Machine Learning
Data Analytics

Machine Learning for Time Series

(MLTS or MLTS-Deluxe Lectures)

Dr. Dario Zanca
Machine Learning and Data Analytics (MaD) Lab
Friedrich-Alexander-Universität Erlangen-Nürnberg
16.11.2023

Topics overview

- Time series fundamentals and definitions (2 lectures)
- Bayesian Inference (1 lecture)
- Gussian processes (2 lectures) ←
- State space models (2 lectures)
- Autoregressive models (1 lecture)
- Data mining on time series (1 lecture)
- Deep learning on time series (4 lectures)
- Domain adaptation (1 lecture)

In this lecture…

1. Gaussian process classification (GPC) formulation

   Classification formulation

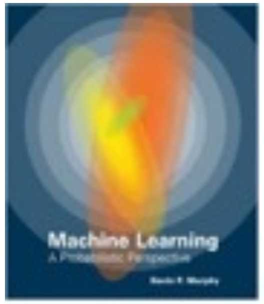
2. Gaussian process classification (GPC) prediction

   Prediction

References

Machine learning: A Probabilistic Perspective,

by Kevin Murphy (2012)



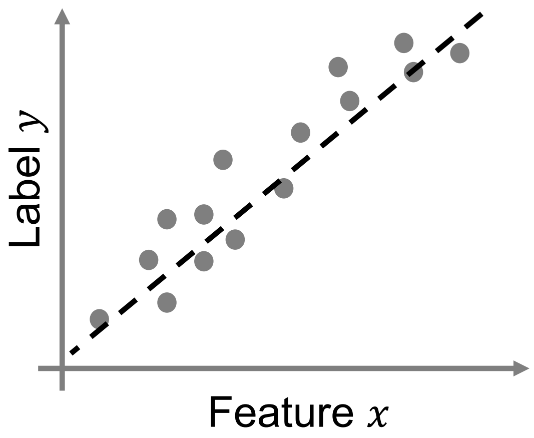
Gaussian process classification (GPC)

GPC formulation

Gaussian Process Classification (GPC)

Regression vs. Classification

Regression



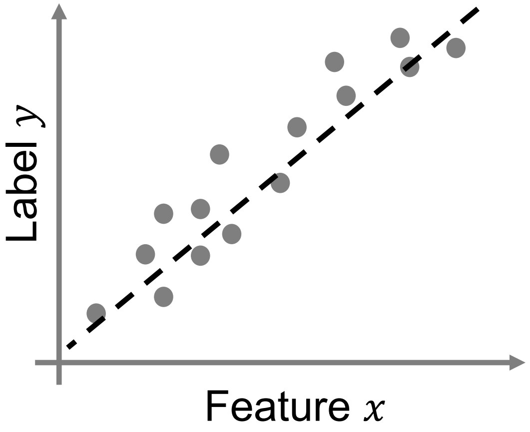
For regression we typically have:

- $x \in \mathbb{R}^d$
- $y_R \in \mathbb{R}$
- $y_R = f(x)$

Gaussian Process Classification (GPC)

Regression vs. Classification

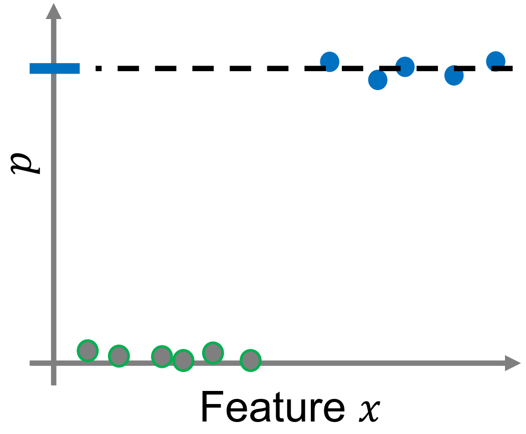
Regression



For regression we typically have:

- $x \in \mathbb{R}^d$
- $y_R \in \mathbb{R}$
- $y_R = f(x)$

Classification



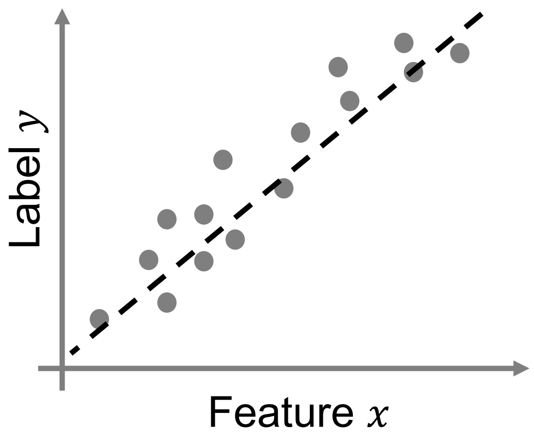
For (binary) classification, instead:

- $x \in \mathbb{R}^d$
- Task: $y_C \in \{-1, +1\}$
- $p \in [0, 1]$

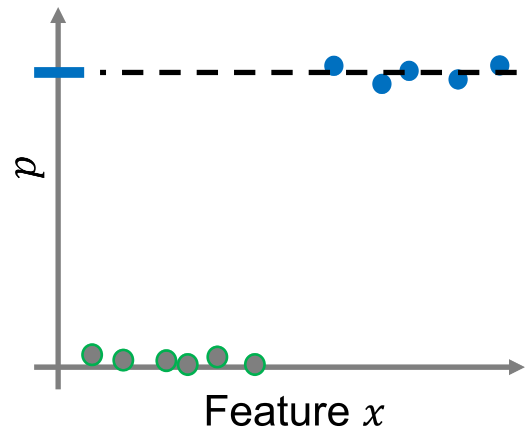
Gaussian Process Classification (GPC)

Regression vs. Classification

Regression



Classification



For regression we typically have:

- $x \in \mathbb{R}^d$
- $y_R \in \mathbb{R}$
- $y_R = f(x)$

For (binary) classification, instead:

- $x \in \mathbb{R}^d$
- Task: $y_C \in \{-1, +1\}$
- $p \in [0, 1]$

Gaussian Process Classification (GPC)

Gaussian linear model

We use a Gaussian linear model in order to obtain the likelihood:

$$p(y = \pm 1 \,|x, w) = \sigma(x^T w)$$

where $\sigma(x^T w)$ is called sigmoid function.

Gaussian Process Classification (GPC)

Gaussian linear model

We use a Gaussian linear model in order to obtain the likelihood:

$$p(y = \pm 1\,|x, w) = \sigma(x^T w)$$

where $\sigma(x^T w)$ is called sigmoid function.

sigmoid

Notice:

- $p(y = \pm 1\,|x, w)$ is the likelihood.

likelihood.

- Generally, we denote $\pi(x) := \sigma(x^T w)$

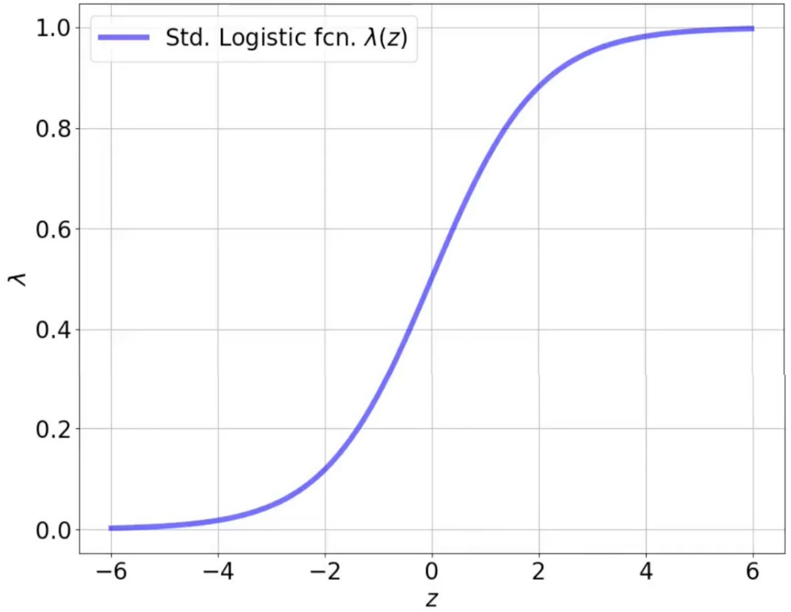
$\pi(x) = \sigma(x^T w)$

notation.

Gaussian Process Classification (GPC)

The sigmoid function

Common options for the sigmoid functions:



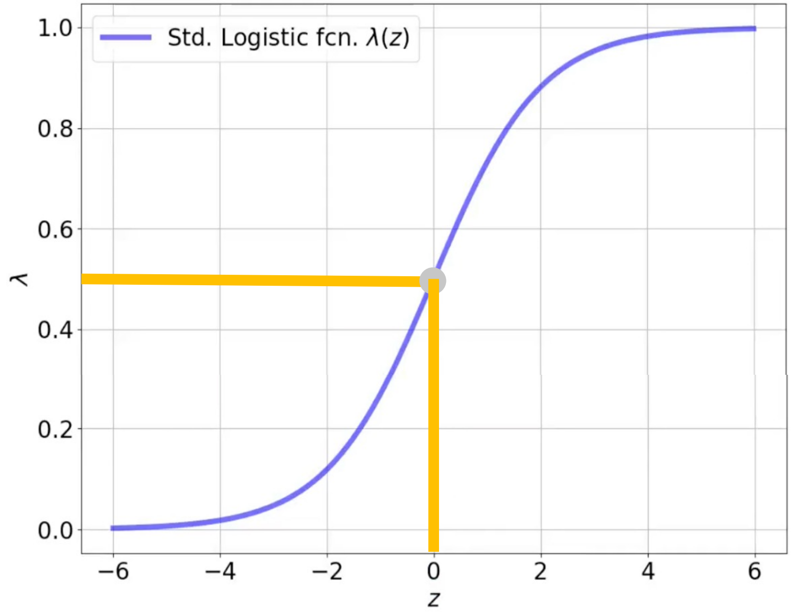
$$\lambda(z) = \frac{1}{1+e^{-z}}$$

(Logistic function)

Gaussian Process Classification (GPC)

The sigmoid function

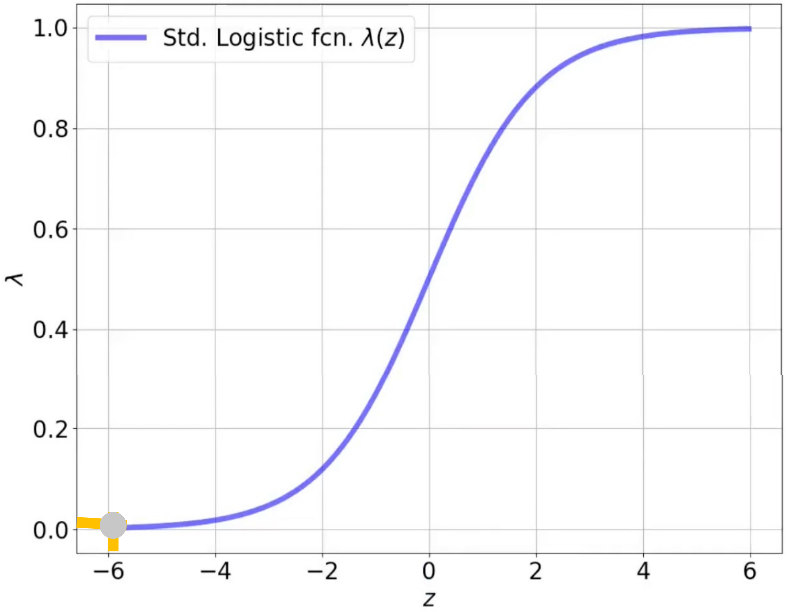
Common options for the sigmoid functions:



$$\lambda(z) = \frac{1}{1+e^{-z}}$$

(Logistic function)

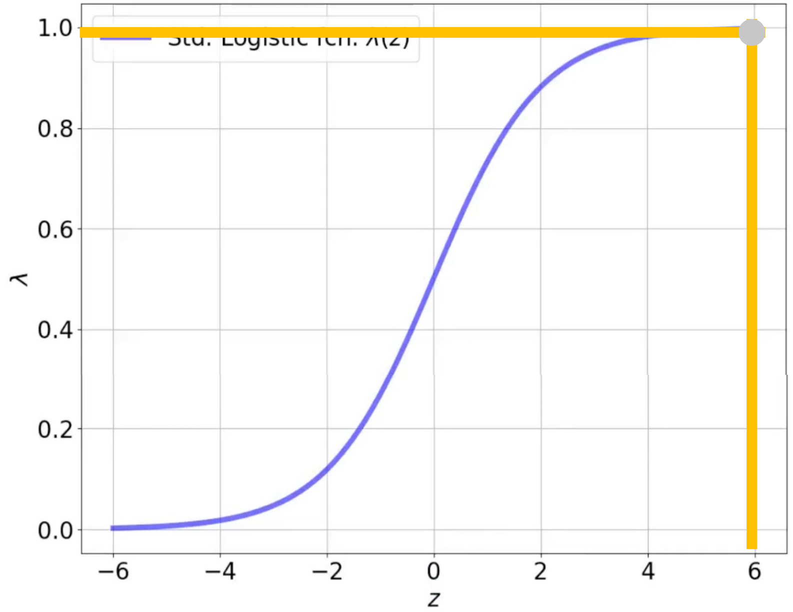
Common options for the sigmoid functions:



$$\lambda(z) = \frac{1}{1+e^{-z}}$$

(Logistic function)

Common options for the sigmoid functions:



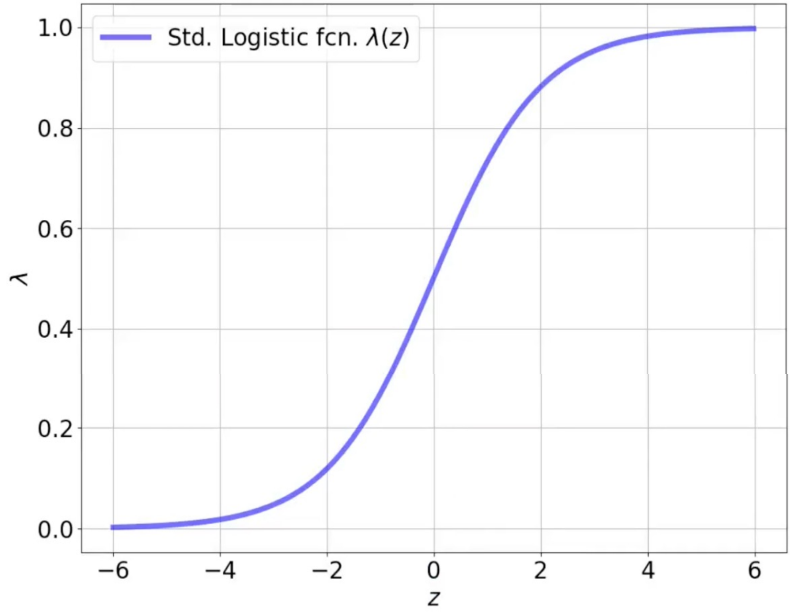
$$\lambda(z) = \frac{1}{1+e^{-z}}$$

(Logistic function)

Gaussian Process Classification (GPC)

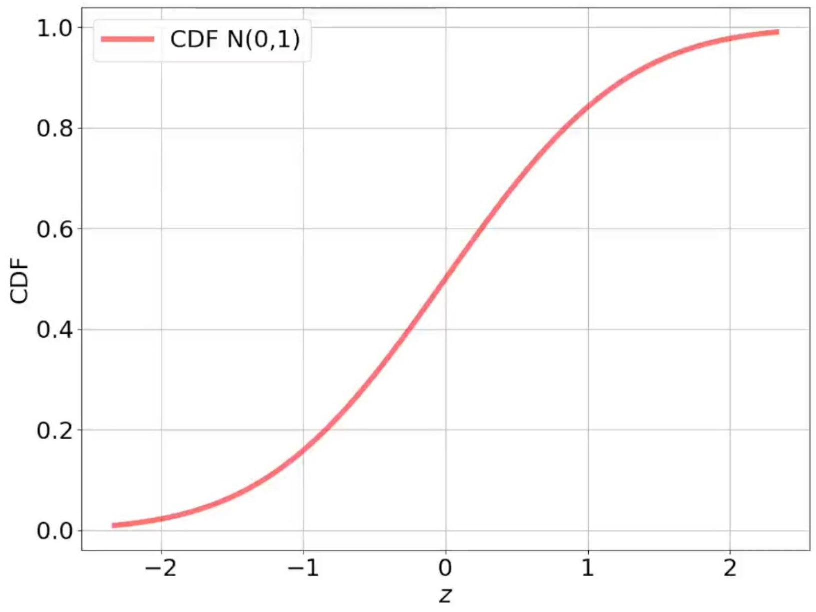
The sigmoid function

Common options for the sigmoid functions:



$$\lambda(z) = \frac{1}{1+e^{-z}}$$

(Logistic function)

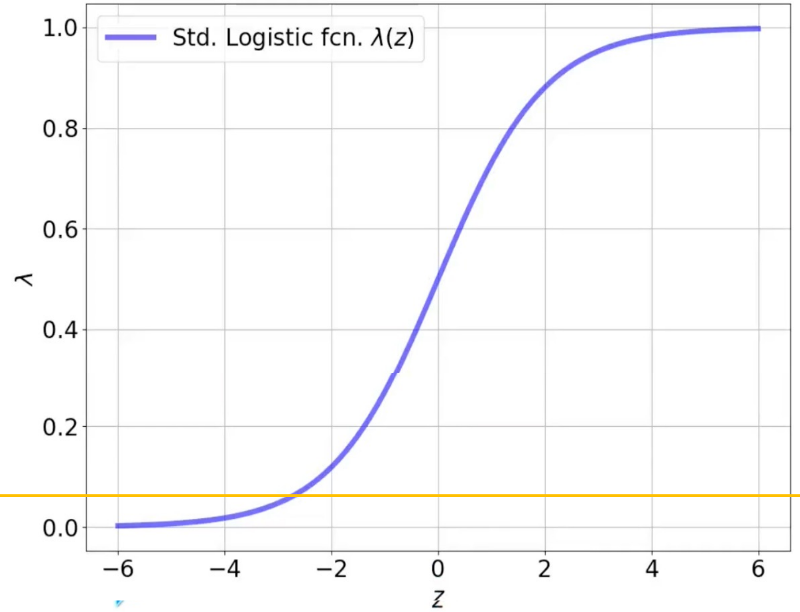


$$\phi(z) = \int_{-\infty}^{z} \mathcal{N}(x|0,1)\, dz$$

(Cumulative distribution function - CDF)

Common options for the sigmoid functions:

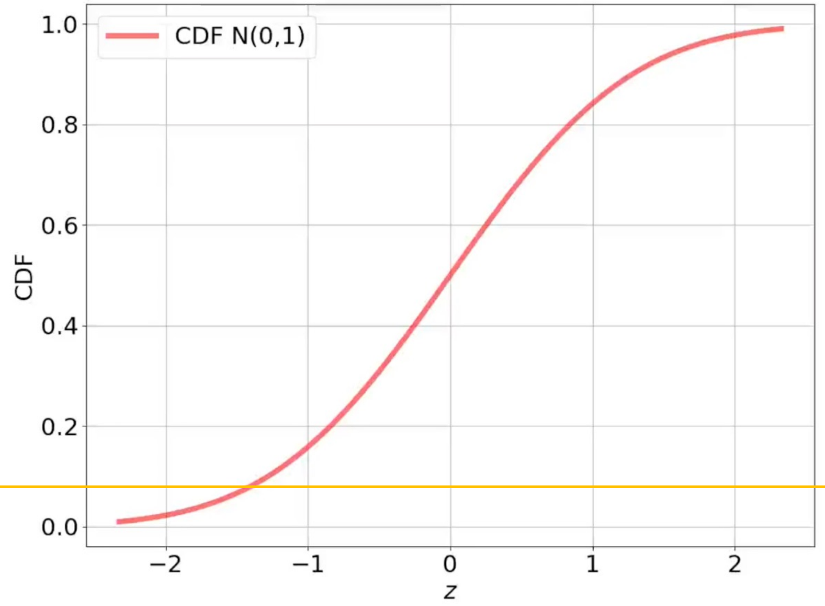
Cumulative distribution function.



$$\lambda(z) = \frac{1}{1+e^{-z}}$$

(-4,4)

(Logistic function)



0 at -2,2

$$\phi(z) = \int_{-\infty}^{z} \mathcal{N}(x|0,1)\, dz$$

(Cumulative distribution function - CDF)

Gaussian Process Classification (GPC)

Likelihood

For a 2-class problem, we can write the likelihood of the value pair (x_i, y_i):

$$\rightarrow \sigma\left(x_i^T w\right) \quad \text{if } y_i = +1$$

$$\rightarrow 1 - \sigma\left(x_i^T w\right) \quad \text{if } y_i = -1$$

Gaussian Process Classification (GPC)

Likelihood

For a 2-class problem, we can write the likelihood of the value pair (x_i, y_i):

$$\rightarrow \sigma(x_i^T w) \qquad \text{if } y_i = +1$$

$$\rightarrow 1 - \sigma(x_i^T w) \quad \text{if } y_i = -1$$

For symmetric sigmoid functions: $\sigma(-z) = 1 - \sigma(z)$

Symmetric

$\sigma(-z) = 1 - \sigma(z)$

Thus: $p(y_i | x_i^T w) = \sigma(y_i f_i)$

- ➢ $y_i = \pm 1$ **(Sign)**
- ➢ $f_i = f(x_i) = x_i^T w$ **(Gaussian Process)**

Gaussian Process Classification (GPC)

Posterior

Let's assume the prior on w:

Enforcing prior.

$$w \sim \mathcal{N}(0, \sigma_p) \quad \text{or} \quad w \sim \mathcal{N}(0, \Sigma_p)$$

Then, we can write the posterior over weights:

$$p(w|y, X) = \frac{p(y|X, w)\ p(w)}{p(y|X)}$$

The marginal likelihood can be written as:

$$p(y|X) = \int p(y|X, w)\ p(w)\, dw$$

Gaussian Process Classification (GPC)

A two-steps approach

Step 1: Gaussian Process (GP) over latent funtion $f(x)$

Step 2: Filter f through a sigmoid function to obtain

$$\pi(x) = p(y = +1|x) = \sigma(f(x))$$

Gaussian process classification

GPC prediction

GPC prediction

Predict a new point x^*:

$$p(y^* = +1|x^*, D) = \int p(y *= +1|w, x^*)\ p(w|D)\ dw$$

GPC prediction

Predict a new point $\boldsymbol{x}^*$:

$$p(y^* = +1|x^*, D) = \int \boxed{p(y^* = +1|w, x^*)}\ p(w|D)\ dw$$

Prediction = $\sigma(x^T w)$

GPC prediction

Predict a new point x^*:

$$\boxed{p(y^* = +1|x^*, D)} = \int \boxed{p(y^* = +1|w, x^*)}\; p(w|D)\; dw$$

Unimodal Non-Gaussian

Prediction = $\sigma(x^T w)$

Prediction

Implementing prediction

Step 1: Compute the distribution of f^* at case x^*.

$$p(f^*|X, y, x^*) = \int p(f^*|X, x^*, f)\ p(f|X, y)\ df$$

The posterior on $f(x)$ can be written as:

$$p(f|X, y) = \frac{p(y|f)\ p(f|X)}{p(y|X)}$$

Implementing prediction

Step 1: Compute the distribution of f^* at case x^*.

$$p(f^*|X, y, x^*) = \int p(f^*|X, x^*, f)\; p(f|X, y)\; df$$

The posterior on $f(x)$ can be written as:

$$p(f|X, y) = \frac{\boxed{p(y|f)}\, p(f|X)}{p(y|X)}$$

Implementing prediction

Compute distribution

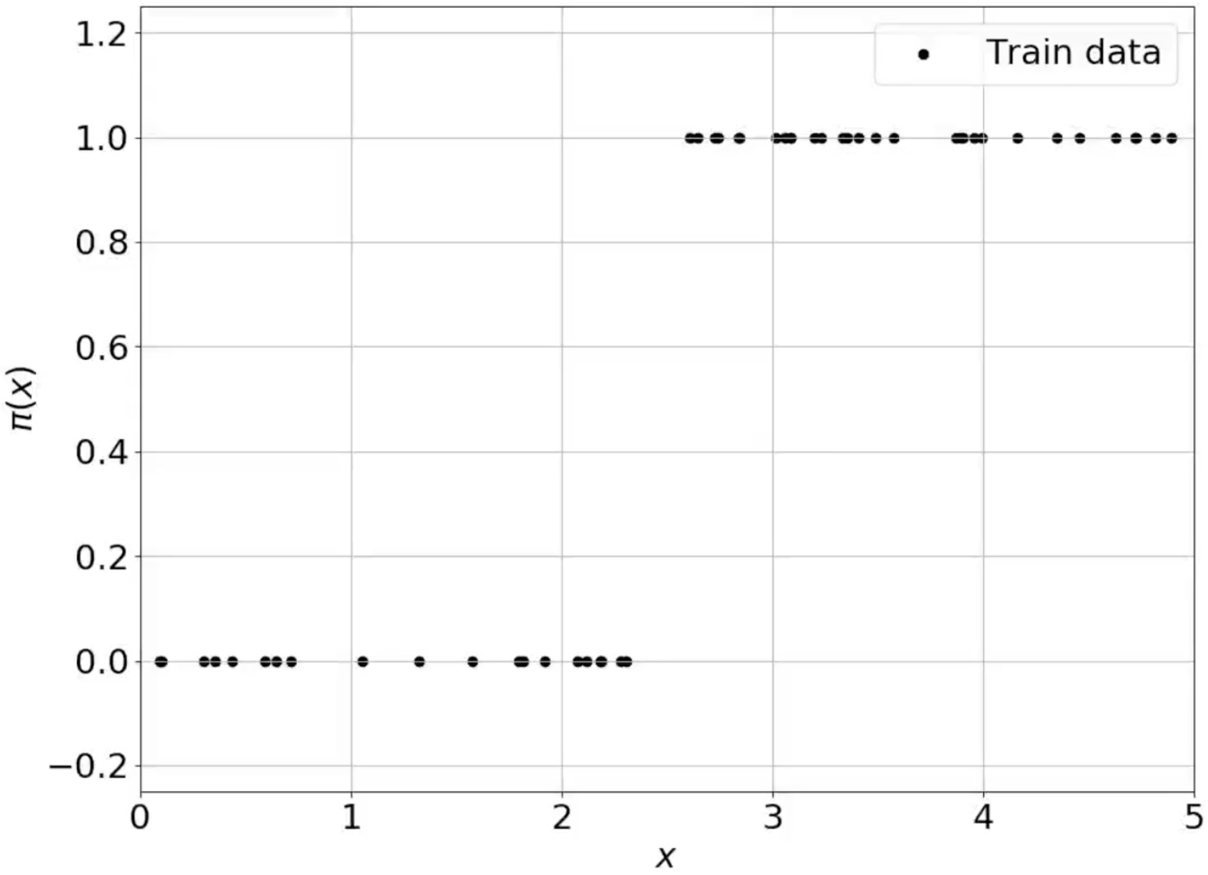
produce probabilistic prediction π^*

Step 2: Produce a probabilistic prediction π^*.

$$\pi^* \triangleq p(y^* = +1 \mid X, y, x^*) = \int \sigma(f^*) \; p(f^*|X, y, x^*) \; df^*$$

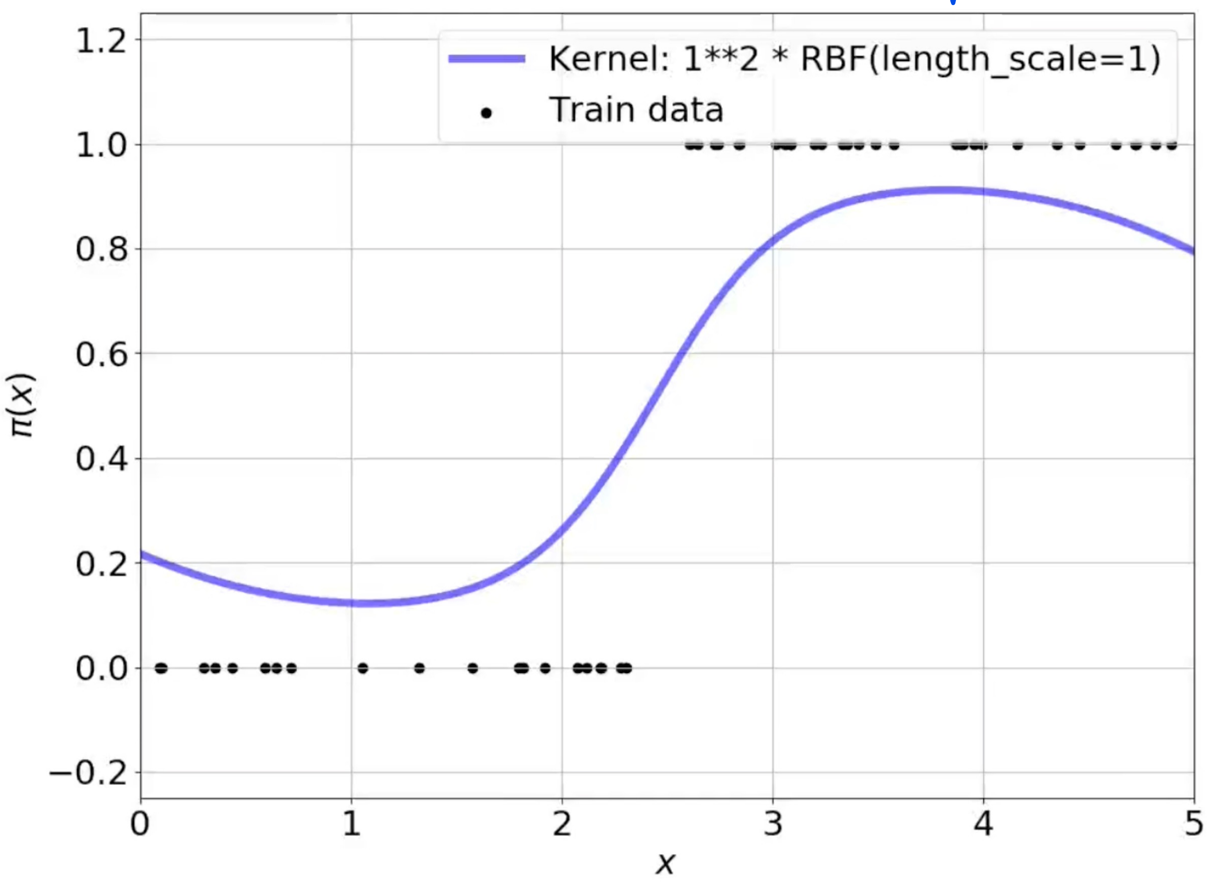
- $\pi^* = \pi(x^*)$ expresses the probability of the class
- The latent f has the role of nuissance function (we do not observe it)

GPC Example

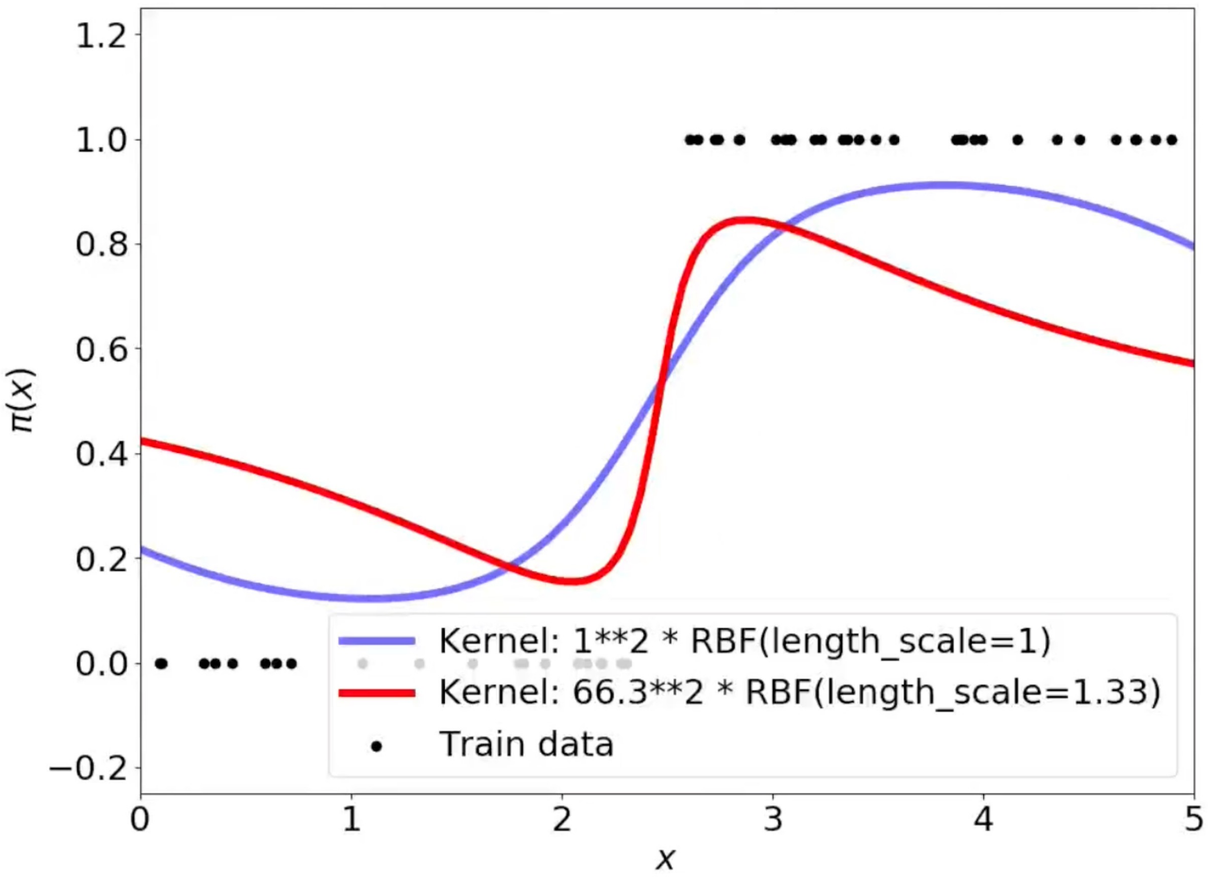


GPC Example

Radial bass fnetton.

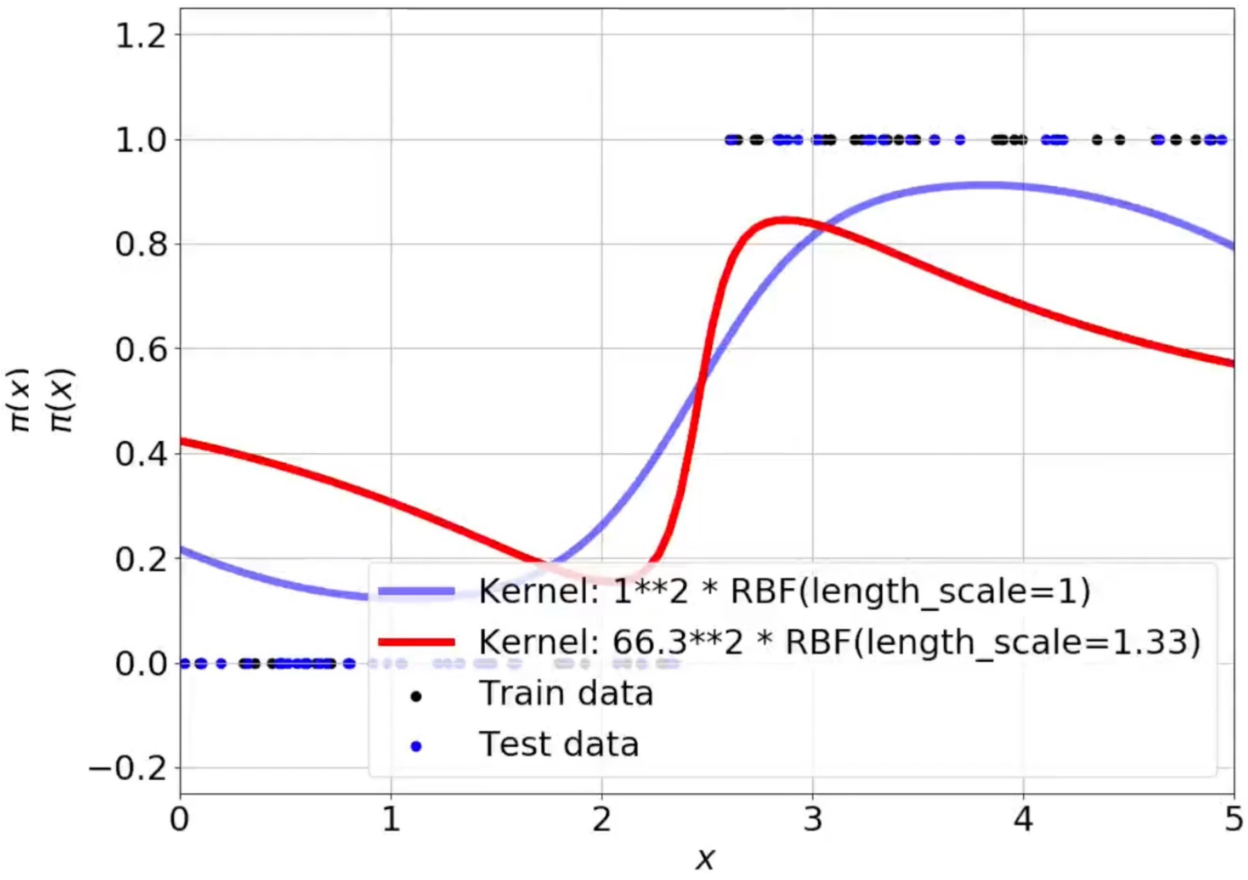


GPC Example



1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
π(x)
0
1
2
3
4
5
x
Kernel: 1**2 * RBF(length_scale=1)
Kernel: 66.3**2 * RBF(length_scale=1.33)
Train data

GPC Example



1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
$\pi(x)$
$\pi(x)$
0
1
2
3
4
5
x
Kernel: 1**2 * RBF(length_scale=1)
Kernel: 66.3**2 * RBF(length_scale=1.33)
Train data
Test data

Gaussian process classification (GPC)

Recap

Recap

- Gaussian process classification (GPC) formulation
  - GLM
  - Sigmoid
- Gaussian process classification (GPC) prediction
  - Two step process