Machine Learning and Intelligent Systems

Feature Transformations

Maria A. Zuluaga

October 20, 2023

EURECOM - Data Science Department

Table of contents

Higher Order Models

Feature Transformations

Recap

Higher Order Models

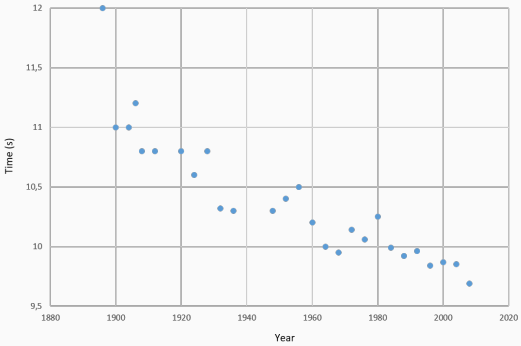
Recap on Men's 100m Olympic Times Problem

Data Assumptions:

$$y_i \in \mathbb{R}$$

Model Assumptions:

$$y = f(x)$$
$$y_i = \mathbf{w}^T \mathbf{x}_i + \epsilon_i$$



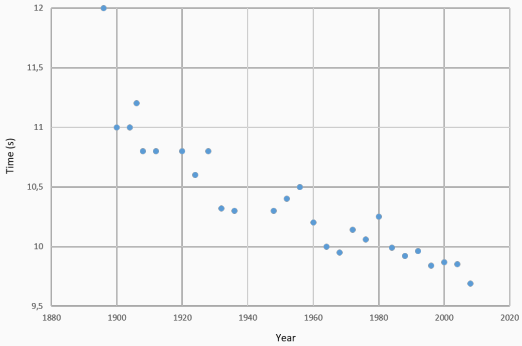
Recap on Men's 100m Olympic Times Problem

Data Assumptions:

$$y_i \in \mathbb{R}$$

Model Assumptions:

$$y = f(x)$$
$$y_i = \mathbf{w}^T \mathbf{x}_i + \epsilon_i$$



What if we consider a higher order model?

Higher Order Model

- One could consider a higher order model by using polynomial features

Higher Order Model

- One could consider a higher order model by using polynomial features
- n^{th} order model:

$$\hat{y} = \hat{w}_0 + \hat{w}_1 x + \hat{w}_2 x^2 + \ldots + \hat{w}_n x^n$$

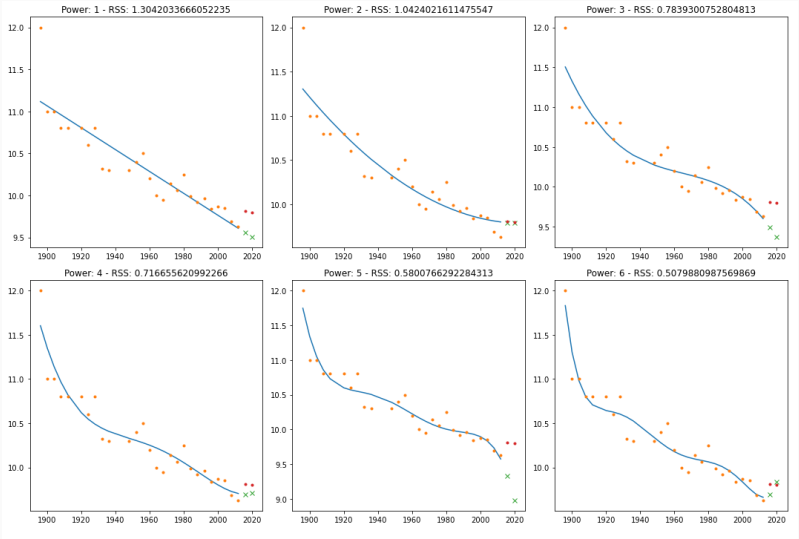
Higher Order Model

- One could consider a higher order model by using polynomial features
- n^{th} order model:

$$\hat{y} = \hat{w}_0 + \hat{w}_1 x + \hat{w}_2 x^2 + \ldots + \hat{w}_n x^n$$

- This is still considered to be a **linear model** as the weights associated with the features are still linear

Polynomial Features: An Example with the Men's 100m Olympic Problem



See: 01_linear_models.ipynb

Questions on the Previous Example

Residual Sum of Squares (RSS): sum of the squares of residuals, where residuals are the deviations predicted from actual empirical values of data

$$RSS = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

Questions: Given that the reported RSS are computed on the training data

- Which one to choose?
  - It seems the more complex the model (higher n) the lower the RSS
  - Is it better to have a more complex model?

Questions on the Previous Example

Residual Sum of Squares (RSS): sum of the squares of residuals, where residuals are the deviations predicted from actual empirical values of data

$$RSS = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

Questions: Given that the reported RSS are computed on the training data

- Which one to choose?
  - It seems the more complex the model (higher n) the lower the RSS
  - Is it better to have a more complex model?
- Based on the RSS, 6 seems the best option. How it performs on 2016 and 2020? What about 2?

Questions on the Previous Example

Residual Sum of Squares (RSS): sum of the squares of residuals, where residuals are the deviations predicted from actual empirical values of data

$$RSS = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

Questions: Given that the reported RSS are computed on the training data

- Which one to choose?
  - It seems the more complex the model (higher n) the lower the RSS
  - Is it better to have a more complex model?
- Based on the RSS, 6 seems the best option. How it performs on 2016 and 2020? What about 2?
- Food for thoughts: If we have to choose a model - how would you proceed?

Feature Transformations

Feature Transformations: Basis Functions

Let us denote $\phi_j(\mathbf{x})$ a *basis function*, $\phi_j(\mathbf{x}) : \mathbb{R}^D \longmapsto \mathbb{R}$

Feature Transformations: Basis Functions

Let us denote $\phi_j(\mathbf{x})$ a *basis function*, $\phi_j(\mathbf{x}) : \mathbb{R}^D \longmapsto \mathbb{R}$
We then model:

$$y(\mathbf{x}, \mathbf{w}) = \mathbf{w}_0 + \sum_{j=1}^{M-1} \mathbf{w}_j \phi_j(\mathbf{x})$$

The total number of parameters of this new model will be M (not necessarily D as before).

Feature Transformations: Basis Functions

Let us denote $\phi_j(\mathbf{x})$ a *basis function*, $\phi_j(\mathbf{x}) : \mathbb{R}^D \longmapsto \mathbb{R}$
We then model:

$$y(\mathbf{x}, \mathbf{w}) = \mathbf{w}_0 + \sum_{j=1}^{M-1} \mathbf{w}_j \phi_j(\mathbf{x})$$

The total number of parameters of this new model will be M (not necessarily D as before).

What has been achieved?

We have extended the class of models by considering linear combinations of fixed nonlinear functions of the input variable.

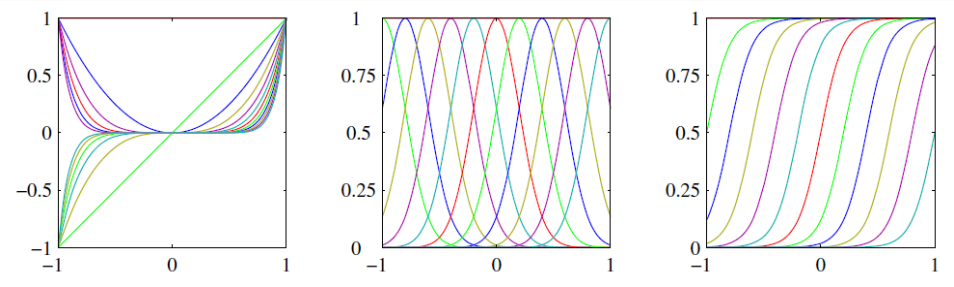
Basis Functions: Examples

Basis Function	Interpretation
$\phi_j(\mathbf{x}) = \mathbf{x}_j,\ j = 1, \ldots, D$	Original linear model
$\phi_j(\mathbf{x}) = \mathbf{x}^j$	Polynomial
$\phi_j(\mathbf{x}) = \exp\left(-\frac{(\mathbf{x} - \mu_j)^2}{2s^2}\right)$	Gaussian
$\phi_j(\mathbf{x}) = \sigma\left(\frac{\mathbf{x} - \mu_j}{s}\right)$	Sigmoid

Table 1: Some commonly used basis functions

* μ_j govern the locations of the basis functions in input space, and the parameter s governs their spatial scale.
* $\sigma(a)$ is denoted the Sigmoid function.

Basis Functions: Examples



Source: Figure 3.1 Bishop, PRML

Examples of basis functions, showing polynomials on the left, Gaussians in the centre, and sigmoidal on the right (from Fig 3.1, Bishop).

Basis Functions: Matrix Representation

After introducing, it is still possible to use a matrix notation for our problems:
Let us define:

$$\phi_0(\mathbf{x}) = 1$$

Then we write Φ the design matrix after transforming the data:

$$\Phi = \begin{bmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \phi_0(\mathbf{x}_3) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_3) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{bmatrix}$$

This means that in the matrix notation $\mathbf{X}$ is transparently replaced by Φ.

Recap

Recap

- We saw how higher order models and how they can lead to more "expressive" models
- We saw a concrete example using polynomial features
- We introduced the concept of basis functions as a way to achieve feature transformation

Later on we will see other ways to transform features

Key Concepts

- Supervised learning
- Input, features, attributes
- Output, target
- Training set
- Basis function

References

Further Reading and Useful Material

Source	Notes
Pattern Recognition and Machine Learning	Ch. 3
The Elements of Statistical Learning	Ch. 5 (deep coverage of basis functions)
Features and Basis Functions (Princeton)	Lecture notes (link)