Transformations of Input or Output

Johanni Brea

Introduction to Machine Learning

EPFL BIO322 2021

Table of Contents

Feature Representation

Idea: Instead of fitting linear regression on p predictors,
fit linear regression on q features of the original predicators.

$$\hat{Y} = \theta_0 + \theta_1 H_1 + \theta_2 H_2 + \cdots + \theta_q H_q$$

with $H_i = f_i(X)$.

Polynomial Regression

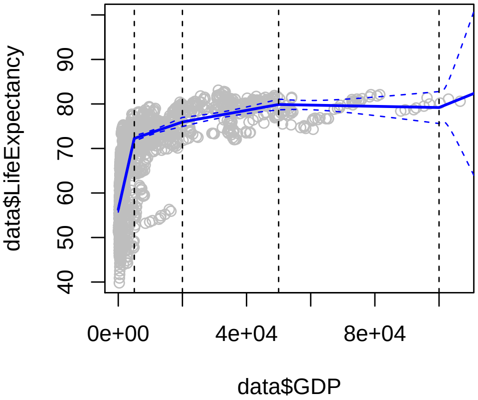
Make a method more flexible by adding features.

With one-dimensional input X ($p = 1$), Polynomial Regression can be written as

$$\hat{Y} = \theta_0 + \theta_1 H_1 + \theta_2 H_2 + \cdots + \theta_q H_q$$

$$\text{where } H_i = f_i(X) = X^i$$

Splines



A **degree-d spline** is a piecewise degree-d polynomial, with continuity in derivatives up to degree $d-1$.

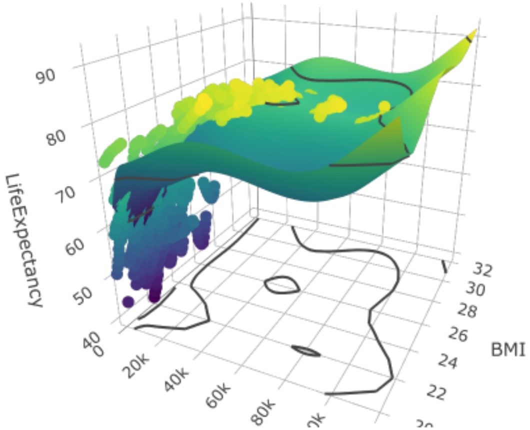
$$H_1 = X, H_2 = X^2, \dots, H_d = X^d$$
$$H_{1+d} = h(X, c_1), \dots, H_{K+d} = h(X, c_K)$$

with knots $c_1, \dots, c_K$ and truncated power basis function:

$$h(x, c) = \begin{cases} (x-c)^d & x > c \\ 0 & \text{otherwise} \end{cases}$$

There are also other possibilities for the basis of a degree-d spline. E.g. the B-spline basis (not discussed here) has better numerical properties.

Generalized Additive Model (GAM)



$$\hat{Y} = s_1(X_1) + s_2(X_2) + \ldots + s_p(X_p)$$

with splines $s_i(X_i) = \sum_j \beta_{ij} H_{ij}$.

Categorical Predictors: Dummy Variables/One-Hot-Coding

Chicken weight as a function of time and diet.
Encode diet as $X_1 \in \{1, 2, 3, 4\}$? No.
$H_i = 1$ if diet $X_1 = i$, otherwise $H_i = 0$.

For example, if $x_{11} = 2$

$$(h_{11}, h_{12}, h_{13}, h_{14}) = (0, 1, 0, 0)$$

Time	Diet1	Diet2	Diet3	Diet4	Weight
0	1	0	0	0	134
2	1	0	0	0	145
4	1	0	0	0	160
0	0	1	0	0	124
2	0	1	0	0	139

When fitting with an intercept, one level (an arbitrarily selected "standard" level) can be dropped; the coefficients are interpreted as change relative to the standard level.

E.g. gender (female or male), treatment (1, 2 or 3)

Intercept	Female	Treat1	Treat2
1	1	0	0
1	1	0	1
1	1	0	0
1	0	1	0
1	0	0	0

Respecting Neighbourhood Relationships

Suppose some predictor X_1 is an angle between 0° and 360°.

If the values are taken as such, 2° looks more different from 259° than from 90° in the sense that |2 - 259| > |2 - 90|.

Alternative: $H_1 = \sin(X_1)$, $H_2 = \cos(X_1)$

In this representation 2° is much closer to 259° than to 90° in the sense that $\|(\sin(2), \cos(2)) - (\sin(259), \cos(259))\| < \|(\sin(2), \cos(2)) - (\sin(90), \cos(90))\|$.

Dealing with Missing Data

We can either

- drop all data points that contain missing data.
  Disadvantage: fewer data points.
- impute missing data with e.g. the mean or the median of that predictor.
  Disadvantage: “wrong” data points.

Standardization

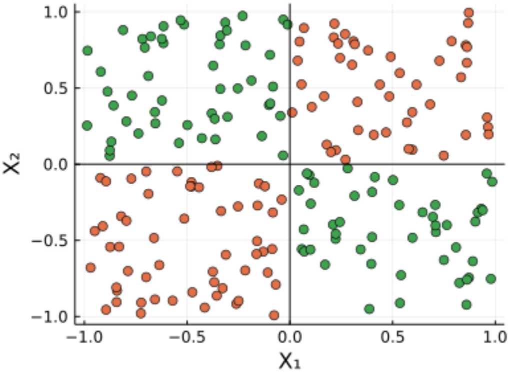
Standardization is a transformation that shifts the data such that its mean is 0 and scales it such that its standard deviation is 1.

Formally: for data $x_1, \ldots, x_n$ with mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and standard deviation $\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ the standardized data is given by

$$\tilde{x}_i = \frac{x_i - \bar{x}}{\sigma}$$

Vector-Features

XOR-Problem
Training Data

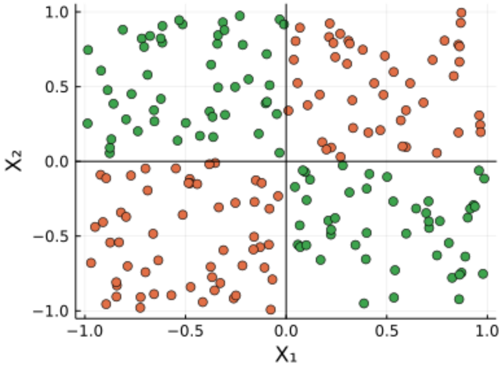


Logistic Regression fails:

There is no linear decision boundary.

Vector-Features

Project data to a higher dimensional space by computing the scalar products between feature vectors $w_1, \ldots, w_q$ and input vectors x_i and thresholding.



For example $h_{21} = \max(0, w_1^T x_2)$.

Logistic Regression on the features works.

Table of Contents

Transformations of the Output: Changing the Noise Model

Applying linear regression to log-transformed outputs is equivalent to assuming a log-normal distribution for the conditional data generator $Y|X$.

Instead of thinking about suitable transformations of the output, it is preferable to think about which distribution is most reasonable for the conditional data generator $Y|X$.