Transformations of Input or Output

Johanni Brea

Introduction à l’apprentissage automatique

GYMINF 2021

Table of Contents

Transformations of the Input: Feature Representations

With input X, Polynomial Regression could also be written as

$$\hat{Y} = \theta_0 + \theta_1 H_1 + \theta_2 H_2 + \cdots + \theta_p H_p$$

where $H_i = X^i$

We call $H_1, \ldots, H_p$ the feature representation of X.

In the following we will see some other feature representations.

Categorical Predictors: Dummy Variables/One-Hot-Coding

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

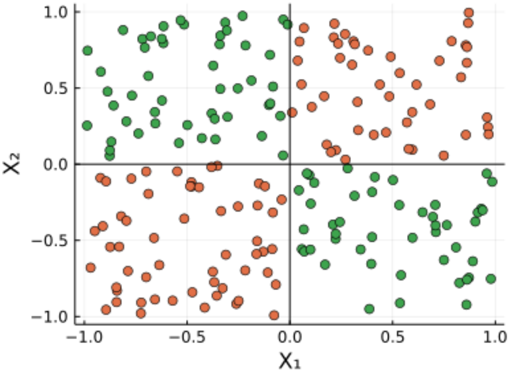
For example, if $x_{11} = 2$

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

	Time	Diet1	Diet2	Diet3	Diet4	Weight
1	0	1	0	0	0	134
2	2	1	0	0	0	145
3	4	1	0	0	0	160
4	0	0	1	0	0	124
5	2	0	1	0	0	139

Vector-Features

XOR-Problem
Training Data

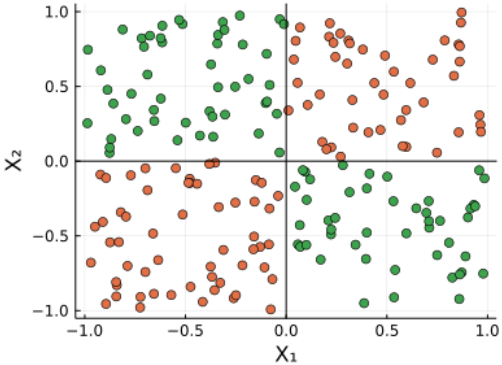


Logistic Regression fails:

There is no linear decision boundary.

Vector-Features

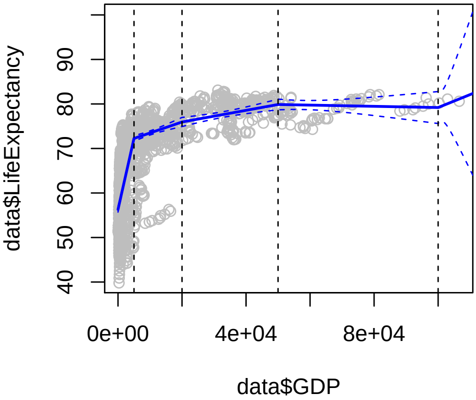
Project data to a higher dimensional space by computing the scalar products between feature vectors $w_1, \dots, w_q$ and input vectors x_i and thresholding.



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

Logistic Regression on the features works.

Splines



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

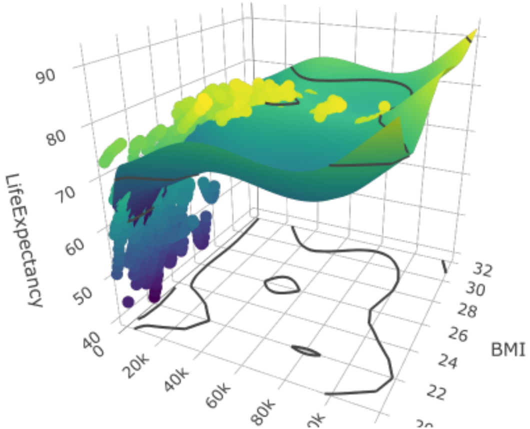
$$H_1 = X, H_2 = X^2, \ldots, H_d = X^d$$
$$H_{1+d} = h(X, c_1), \ldots, H_{K+d} = h(X, c_K)$$

with knots $c_1, \ldots, c_K$ 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 $s_i(X_i) = \sum_j \beta_{ij} H_{ij}$.

Pros and Cons of GAMs

- ▲ GAMs allow to fit non-linear s_j to each s_j.
- ▲ The non-linear fit can potentially make more accurate predictions.
- ▲ GAMs are useful for inference: because of additivity one can still examine each effect, holding all the other variables fixed.
- ▲ The smoothness of each function s_j can be summarized via degrees of freedom.
- ▼ Important interactions can be missed. It is possible to add interactions of the form $s_{ij}(X_i, X_j)$ but this becomes costly very soon.

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 $Y|X$.