Transformations of Input or Output

Johanni Brea

Introduction to Machine Learning

Table of Contents

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.
- impute missing data with unsupervised learning tools, like matrix completion (see later in the course).

Removing Predictors

- Constant predictors should be removed.
- If multiple predictors are perfectly correlated, keep only one of them.
- One can also try to remove almost constant or almost perfectly correlated predictors, but – WATCH OUT – this may introduce errors.
- If the response Y is known to be independent of a predictor X, i.e. $P(Y, X) = P(Y)P(X)$, it should be removed. In praxis, it is usually not known if the response is really independent of a given predictor.

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

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

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 $H_i = f_i(X)$.

Previous Examples: Polynomial Regression & Bag-of-Words

Make a method more flexible by adding features.

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

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

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

Bag-of-Words for the spam dataset can be seen as another example of feature engineering, where H_i = normalized count of word i in email X.

Categorical Predictors: Dummy Variables/One-Hot-Coding

Chicken weight as a function of time and diet.

$H_i = 1$ 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

Why not an integer code $X_1 \in \{1, 2, 3, 4\}$ or a binary code $X_1 \in \{(0, 0), (1, 0), (0, 1), (1, 1)\}$ for diet?

- The pairwise Euclidean distances are not the same (diet 1 is closer to diet 2 than to diet 4), which may be nonsensical for the given data.
- $\hat{Y} = f(X)$ may look more complicated (non-linear) for the integer or binary code than for the one-hot code.

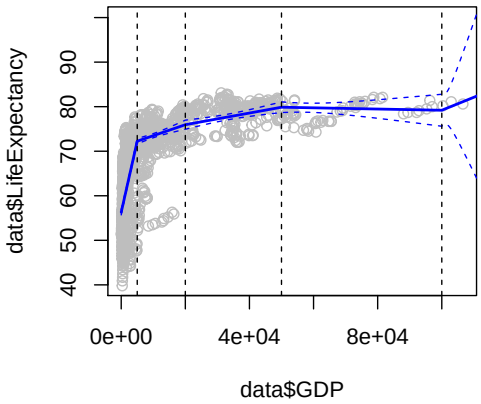
Categorical Predictors: Dummy Variables/One-Hot-Coding

When fitting a linear model with intercept, one level (an arbitrarily selected “standard” level) should be dropped for each predictor; 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

Splines



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

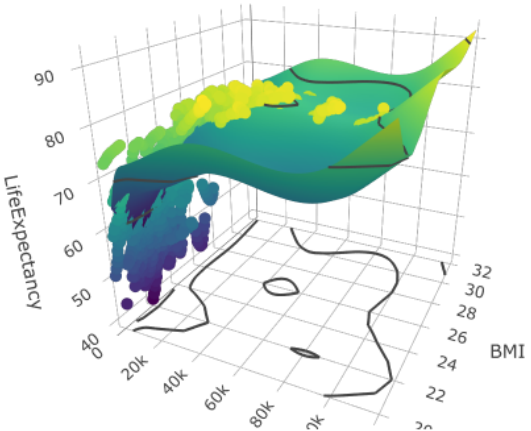
$$H_1 = X, H_2 = X^2, \ldots, H_d = X^d$$
$$H_{1+d} = h(X, c_1), \ldots, H_{K+d} = h(X, c_K)$$

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

Respecting Neighbourhood Relationships

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

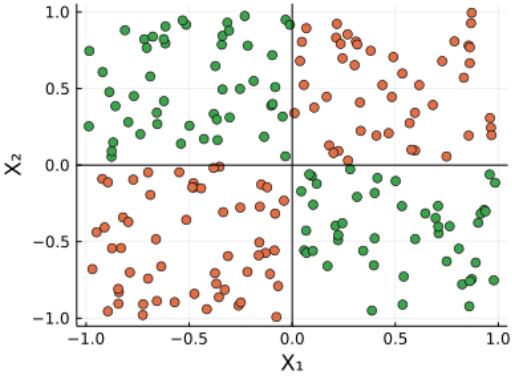
If the values are taken as such, 2° looks more different from 359° than from 90° in the sense that |2 - 359| > |2 - 90|.

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

In this representation 2° is much closer to 359° than to 90° in the sense that $\|(\sin(2), \cos(2)) - (\sin(359), \cos(359))\| < \|(\sin(2), \cos(2)) - (\sin(90), \cos(90))\|$.

Vector Features

XOR-Problem
Training Data

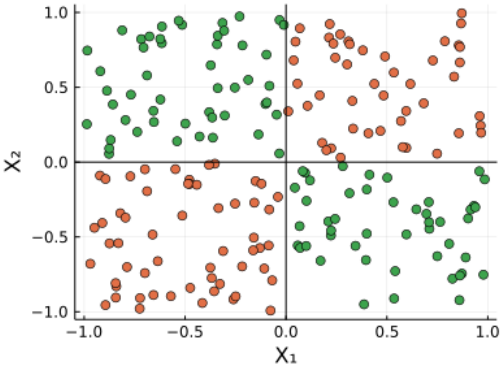


Logistic Regression fails:

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.

Quiz

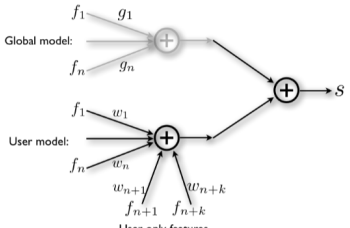
1. For every classification problem with (non-linear) decision boundary and zero irreducible noise there exists a feature representation such that logistic regression on the features solves the classification problem without errors.
2. To fit a degree-d spline with K knots we use a feature representation and linear regression with $d + K + 1$ parameters.
3. We want to predict the number of rented bicycles based on weather condition (sunny, cloudy, foggy, rainy), wind speed, week day (Monday, Tuesday, etc.). After one-hot coding relative to a standard level there are

   A 3 **B** 10 **C** 12 **D** 13 predictors

Gmail Priority Inbox in 2010

The Learning Behind Gmail Priority Inbox

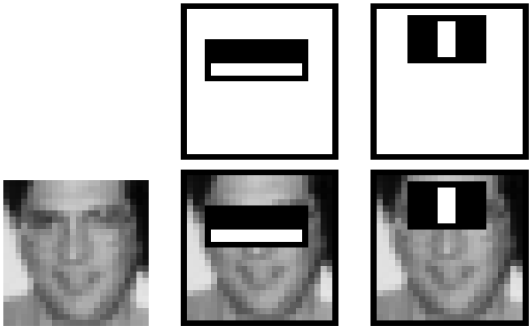
Douglas Aberdeen **Ondrej Pacovsky** **Andrew Slater**
Google Inc.
Zurich, Switzerland



2.1 Features

There are many hundred features falling into a few categories. *Social features* are based on the degree of interaction between sender and recipient, e.g. the percentage of a sender's mail that is read by the recipient. *Content features* attempt to identify headers and recent terms that are highly correlated with the recipient acting (or not) on the mail, e.g. the presence of a recent term in the subject. Recent user terms are discovered as a pre-processing step prior to learning. *Thread features* note the user's interaction with the thread so far, e.g. if a user began a thread. *Label features* examine the labels that the user applies to mail using filters. We calculate feature values during ranking and we temporarily store those values for later learning. Continuous features are automatically partitioned into binary features using a simple ID3 style algorithm on the histogram of the feature values.

Face Detection with Rectangle Features



The two most important features for face detection are shown. The first one is a 2-rectangles feature, the second one a 3-rectangles feature. The sum of the pixels which lie within the white rectangles are subtracted from the sum of pixels in the black rectangles.

Rapid object detection using a boosted cascade of simple features

http://dx.doi.org/10.1109/CVPR.2001.990517

Table of Contents

Transformations of the Output: Changing the Noise Model

Applying linear regression to g-transformed outputs is equivalent to assuming a "g-normal" distribution for the conditional data generator $Y|X$, i.e.

$$p(Y = y|X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(g(y)-f(x))^2}{2\sigma^2}} \tag{1}$$

For example: $g(y) = log(y) = Y$ is log-normally distributed.
Instead of thinking about suitable transformations of the output,
it may be preferable to think about which distribution is most reasonable
for the conditional data generator $Y|X$/

Suggested Reading

- 3.3.1 Qualitative Predictors