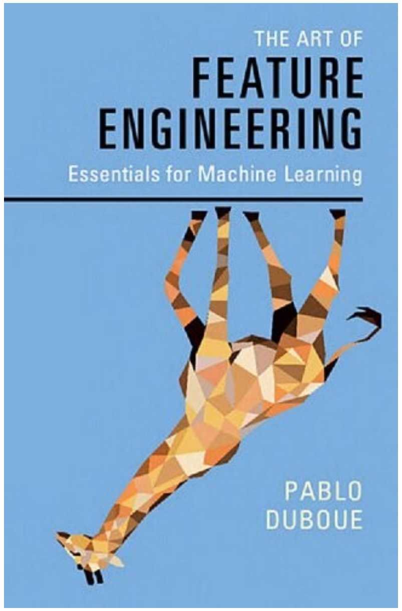
Feature engineering and feature selection

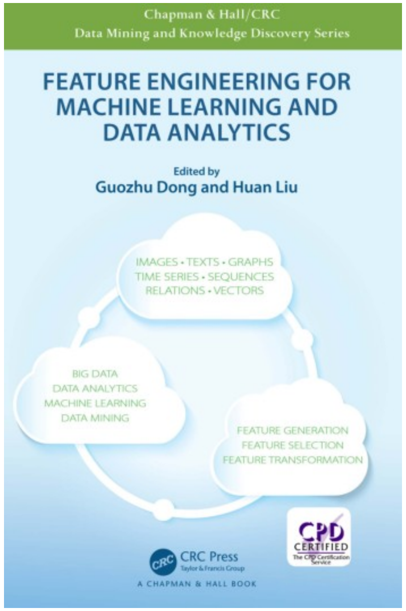
Vlad Gladkikh

IBS CMCM



THE ART OF
FEATURE ENGINEERING
Essentials for Machine Learning

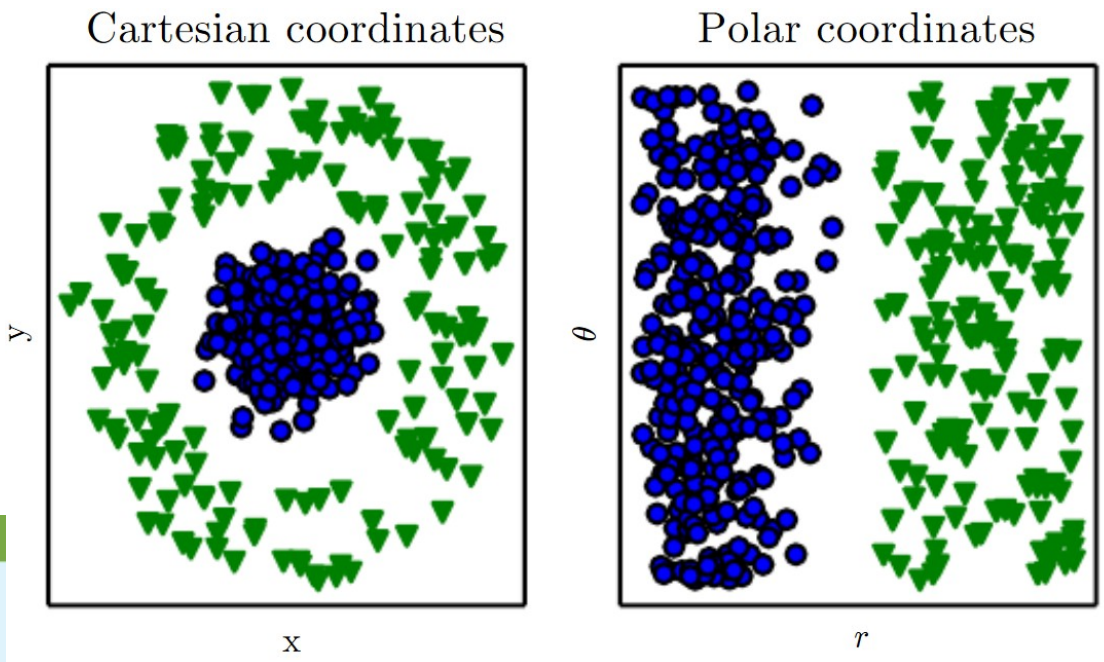
PABLO
DUBOUE



Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series

FEATURE ENGINEERING FOR MACHINE LEARNING AND DATA ANALYTICS

Edited by
Guozhu Dong and Huan Liu



http://www.deeplearningbook.org

> People are always subjective. They choose only features they like or find "more important". Please, avoid being human.

https://vas3k.com/blog/machine_learning/

Candidate features: $X = (X_1, X_2, \ldots, X_p) \ : \ n \times p$

Targets: $Y \ : \ n \times 1$

Primary features: may be properties of isolated atoms

or properties of the materials (composition and geometry).

group and period numbers,
number of valence electrons,
atomic mass,
electron affinity,
thermal conductivity,
heat capacity,
enthalpies of atomization,
fusion and vaporization,
ionization potentials,
effective atomic charge,
molar volume,
chemical hardness,
covalent and van der Waals radii,
electronegativity,
polarizability

Then build a large feature space *X* by combining the primary features via analytic formulas in a ‘grammatically correct’ way.

The goal of feature selection is to find a small subset of features that explain the targets well:

$$Y = f(X_1, X_2, \ldots, X_{\tilde{p}}) \ : \ \tilde{p} < p$$

Which candidate features can be safely removed?

Variance threshold: remove features that have the same value in all or most of the samples

Remove features that are highly correlated with each other.

Consider features correlated with the target.

Pearson correlation coefficient between two variables $x = \{x_1, \ldots, x_n\}, \quad y = \{y_1, \ldots, y_n\}$

(Auguste Bravais, 1846; Francis Galton, 1888)

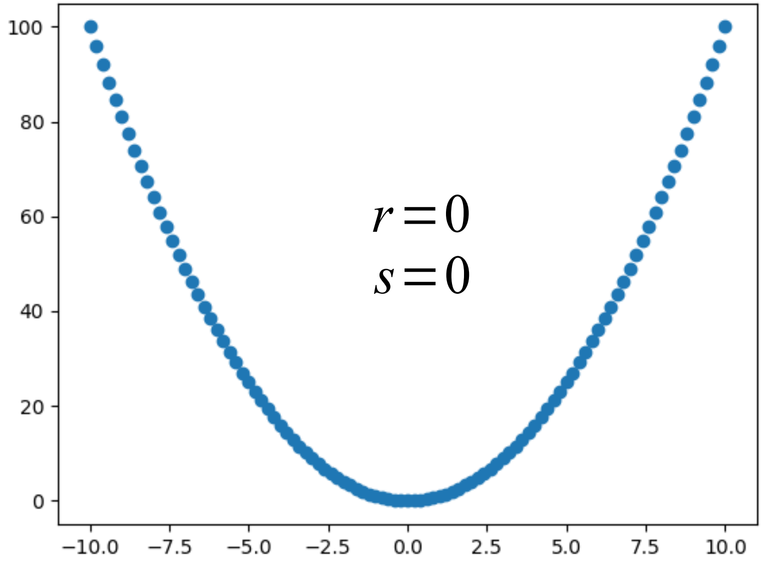
$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

– measures the linear relationship between *x* and *y*

Spearman's rank correlation coefficient (1904)

– the Pearson correlation coefficient between the ranked variables

– assesses monotonic relationships between *x* and *y*



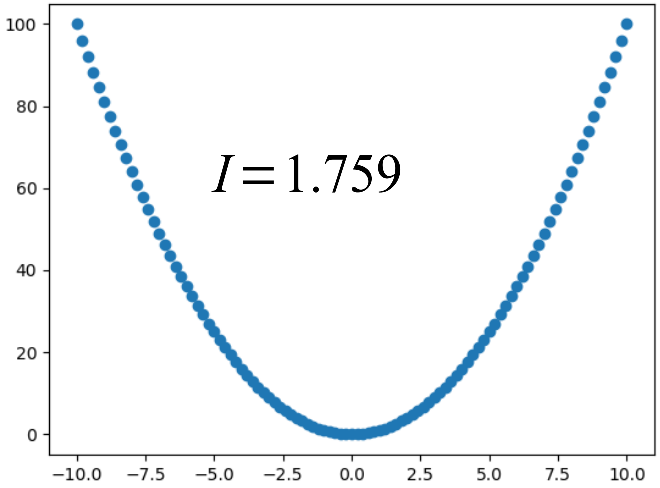
Mutual information between two random variables *X* and *Y*

$$I(X,Y)=\int_{x\in X}\int_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

May detect non-monotonic, more complicated relationships.

Mutual information measures the information that *X* and *Y* share. $I(X,Y)\geq 0$

If *X* and *Y* are independent, then $I(X,Y)=0$: knowing *X* does not give any information about *Y* and vice versa.



If *X* is a deterministic function of *Y* and *Y* is a deterministic function of *X* then all information conveyed by *X* is shared with *Y*: knowing *X* determines *Y* and vice versa.

In this case the mutual information is the same as the uncertainty contained in *Y* (or *X*) alone, namely the entropy of *Y* (or *X*).

$$I(X,Y)=H(Y)-H(Y|X)$$

$H(Y)=-\int_{y\in Y} p(y)\log p(y)$ – marginal entropy, a measure of uncertainty about *Y*

Conditional entropy

$H(Y|X)=\int_{x\in X} p(x)\int_{y\in Y} p(y|x)\log p(y|x)$ – the amount of uncertainty remaining about *Y* after *X* is known

Mutual information is the amount of information (reduction in uncertainty) that knowing either variable provides about the other.

Mutual information is *not* a measure of mutual dependence

It is a measure of mutual entropy! $I(X,Y)=H(X)+H(Y)-H(X,Y)$

$$I(X,Y)=H(X,Y)-H(X|Y)-H(Y|X)$$

Example: $Y=\sin(X)$

$I(X,Y)$ – symmetric, $I(X,Y)=I(Y,X)$, however

We can predict *Y* from *X* with 100% but cannot *X* from *Y*

If mutual information was a measure of mutual dependence, we would have

$$I[y(x)]>I[x(y)]$$

https://stats.stackexchange.com/questions/322814/mutual-information-for-feature-selection

Example: $Y = X\,, \quad x_i = \{0,1\}$

1) $$p(x_i=0)=\frac{1}{2},\ \ p(x_i=1)=\frac{1}{2}\ \Rightarrow\ p(x,y)=\begin{pmatrix}0.5 & 0\\ 0 & 0.5\end{pmatrix}$$

Then $I(X,Y)=\log(2)\approx 0.693$

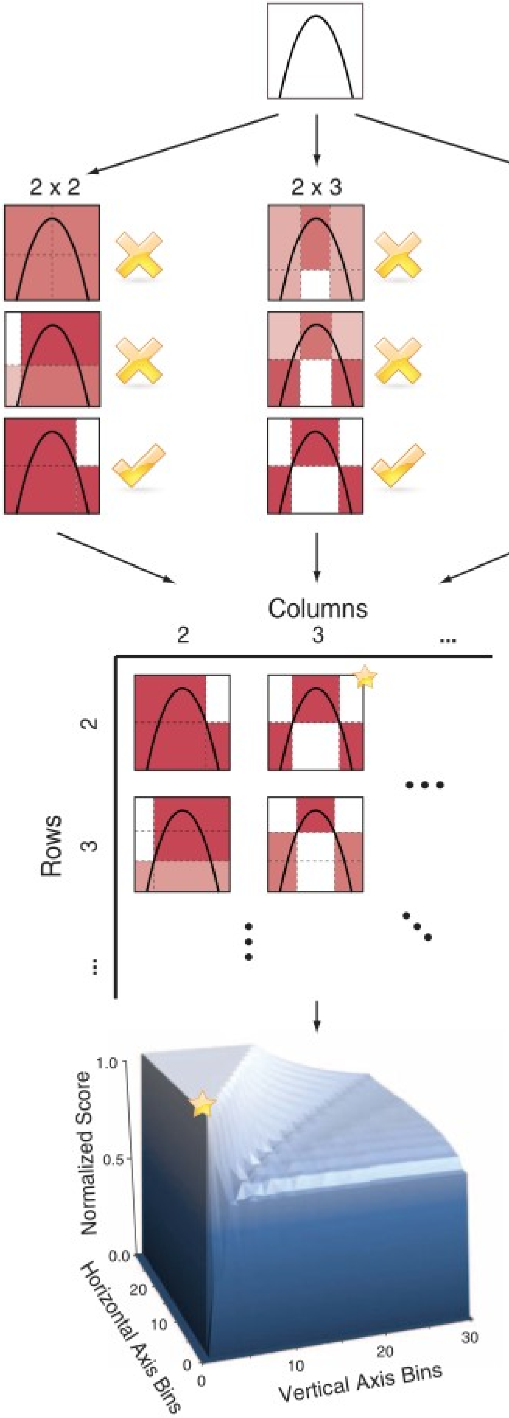
2) $$p(x_i=0)=\frac{1}{10},\ \ p(x_i=1)=\frac{9}{10}\ \Rightarrow\ p(x,y)=\begin{pmatrix}0.1 & 0\\ 0 & 0.9\end{pmatrix}$$

Then $$I(X,Y)=\frac{1}{10}\log(10)+\frac{9}{10}\log\left(\frac{10}{9}\right)\approx 0.325$$

Perfect prediction in both cases but the MI is different.

In the 2nd example the perfect correlation remained, but the entropy of each signal decreased and their mutual entropy decreased also.

https://stats.stackexchange.com/questions/322814/mutual-information-for-feature-selection



The maximal information coefficient (MIC)

Reshef et. al, Detecting Novel Associations in Large Data Sets

http://science.sciencemag.org/content/334/6062/1518

Originally done by Done by Wassily Hoeffding in 1948

Annals of Mathematical Statistics 19:546 https://projecteuclid.org/euclid.aoms/1177730150

https://stats.stackexchange.com/questions/20011/can-the-mic-algorithm-for-detecting-non-linear-correlations-be-explained-intuiti

G – grid

I_G – mutual information of the probability distribution induced on the boxes of *G*

The probability of a box is proportional to the number of data points inside the box.

$$m_{xy} = \frac{max\{I_G\}}{\log min\{x, y\}}$$

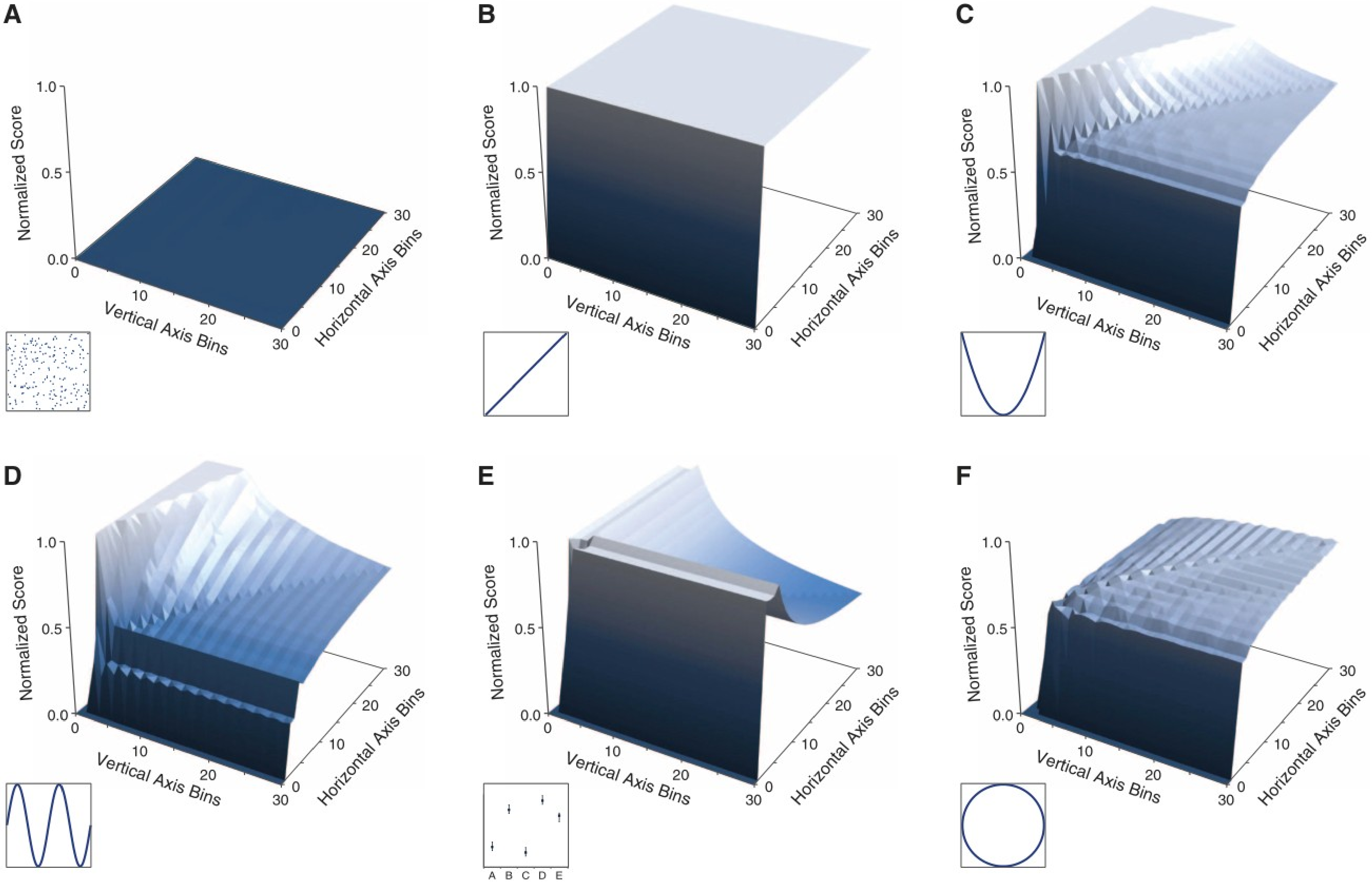
the maximum is taken over all *x*-by-*y* grids *G*

$$MIC = \underset{xy<n^{0.6}}{max}\{m_{xy}\}$$

n – sample size

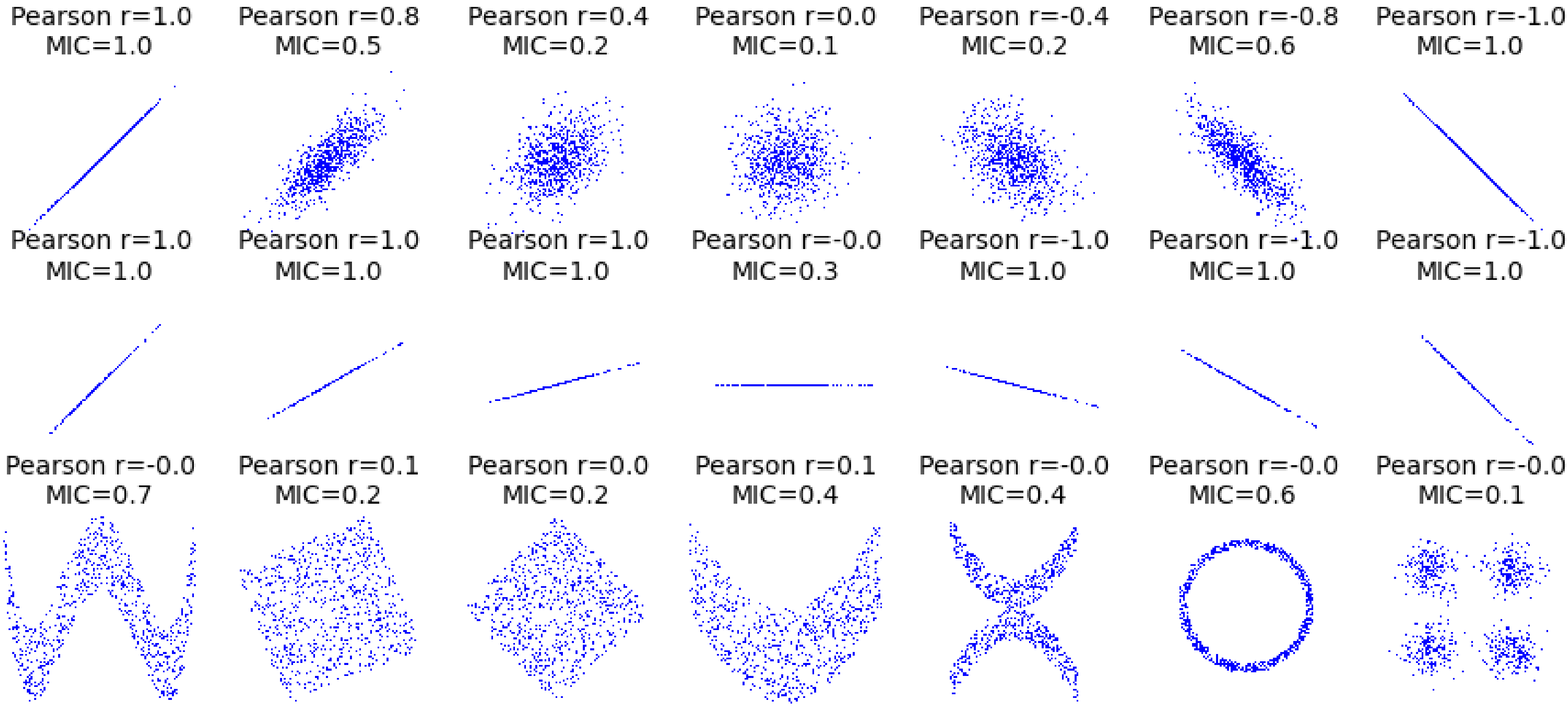
For computational efficiency, a dynamic programming is used that optimizes over a subset of the possible grids.

$$MIC(X, Y) = MIC(Y, X)$$



http://science.sciencemag.org/content/334/6062/1518 https://minepy.readthedocs.io/en/latest/index.html

C, C++, Python, MATLAB/OCTAVE, R



XOR: each columns has low MIC

$$X=\begin{pmatrix}0 & 0\\ 0 & 1\\ 1 & 0\\ 1 & 1\end{pmatrix} \qquad y=\begin{pmatrix}0\\ 1\\ 1\\ 0\end{pmatrix}$$

$$I(X[:,0],y)=0$$
$$I(X[:,1],y)=0$$

https://stats.stackexchange.com/questions/277710/how-to-be-absolutely-sure-that-features-do-have-predictive-power-to-predict-the

Question:

I calculated the MI of features with the target and sorted the features in descending order of mutual information.

The results seemed wierd to me as the highest MI that I get from any feature is 0.0063. Does this mean that my data is worthless and I should probably look for more data?

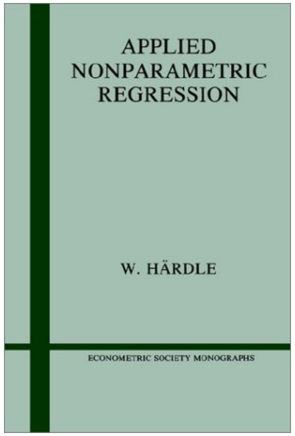
Answer:

No, a small mutual information between a target variable and single features does not render your dataset worthless since it neglects the information contained in the combination of features.

Multivariate mutual information https://en.wikipedia.org/wiki/Multivariate_mutual_information

Alternating Conditional Expectations (ACE)

Breiman, L. and Friedman, J. H. (1985) Estimating optimal transformations for multiple regression and correlation. J. Am. Stat. Assoc., 80(391):580–598. https://www.tandfonline.com/doi/abs/10.1080/01621459.1985.10478157


APPLIED NONPARAMETRIC REGRESSION
W. HÄRDLE
ECONOMETRIC SOCIETY MONOGRAPHS

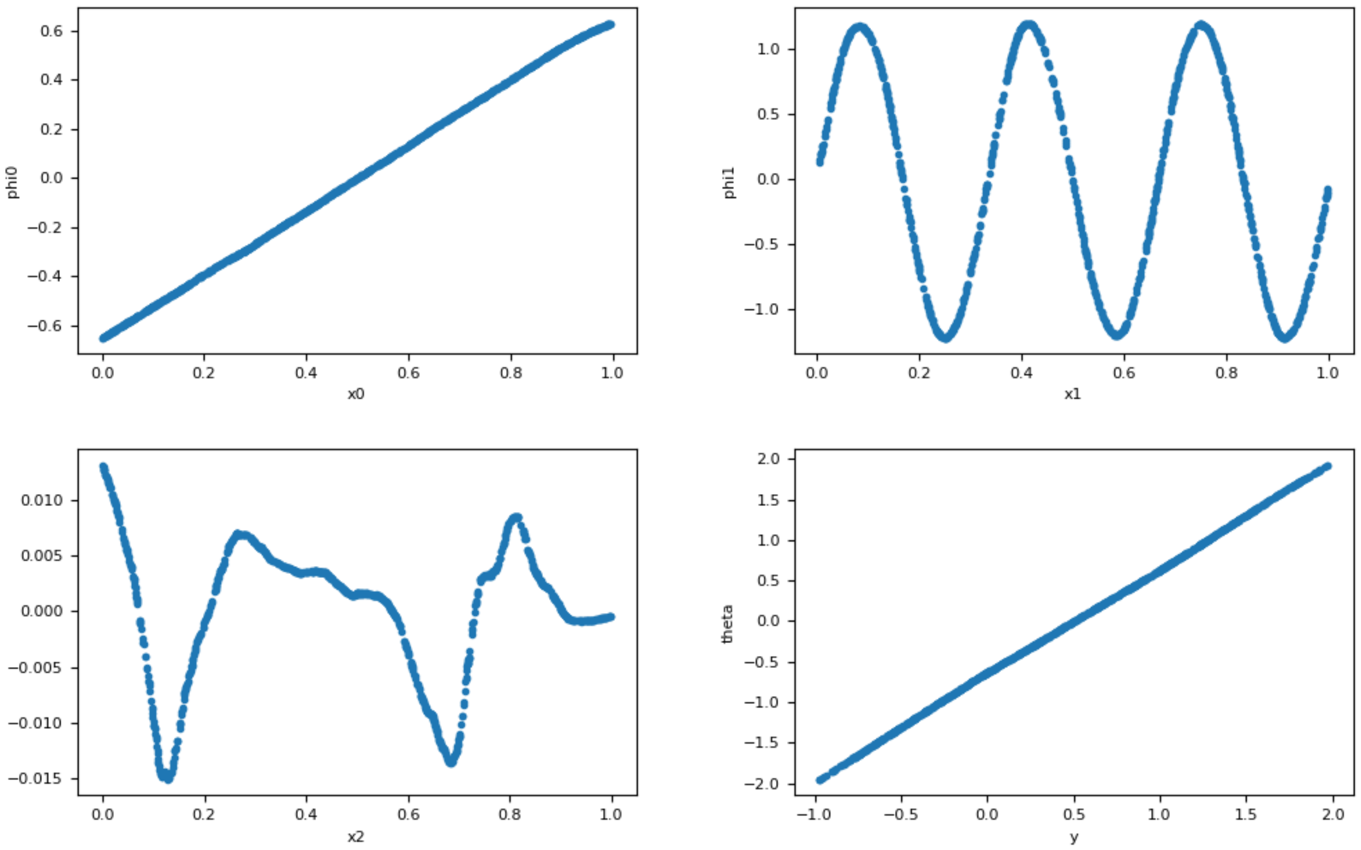
$$\theta(Y)=\alpha+\sum_{i=1}^{p} c_i\phi_i(X_i)+\epsilon$$

Fortran http://lib.stat.cmu.edu/general/ace

R https://cran.r-project.org/web/packages/acepack/index.html

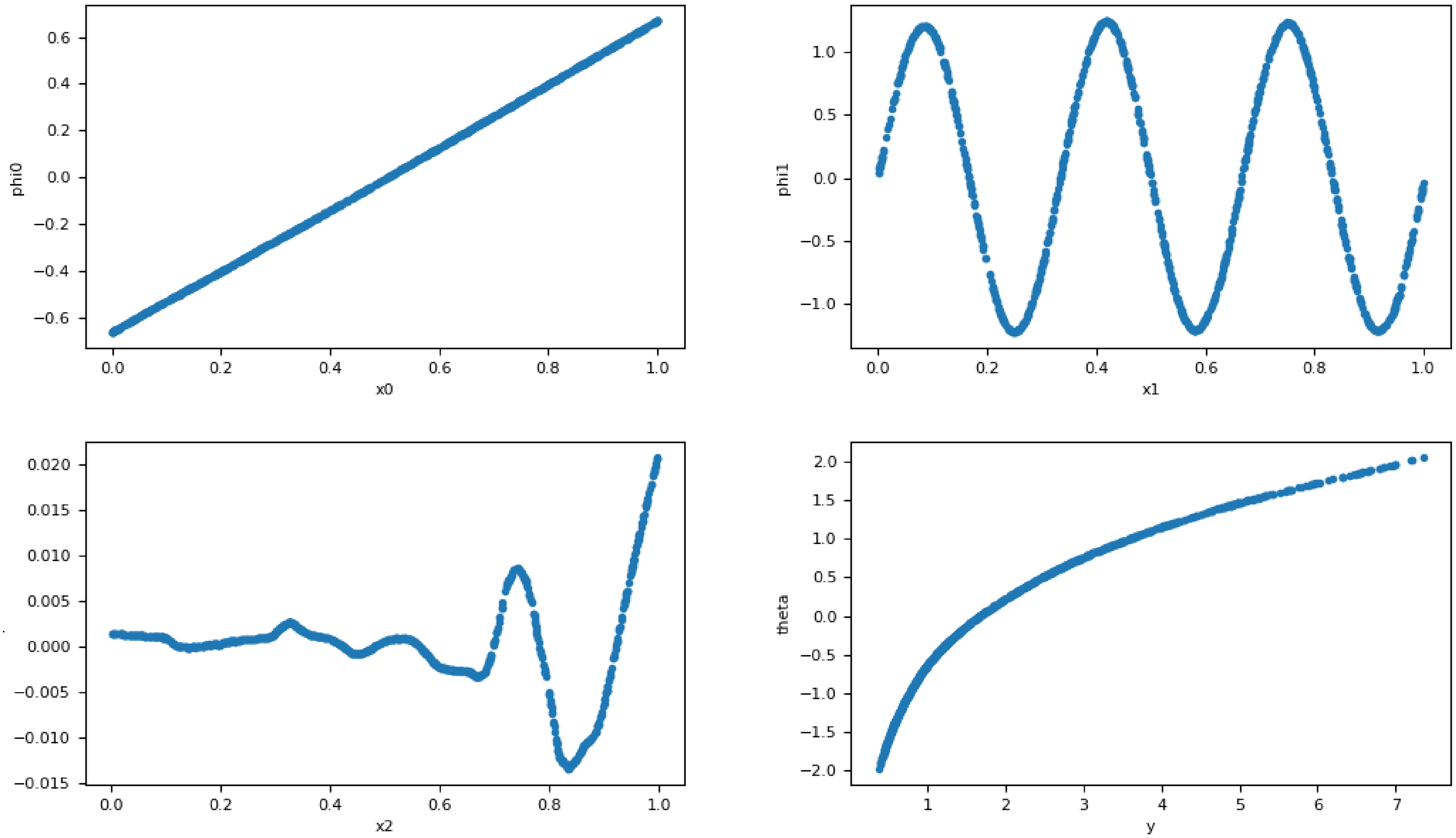
$$y=x_0+\sin(6\pi x_1)$$

```
library("acepack")

x0 <- runif(1000)
x1 <- runif(1000)
x2 <- runif(1000)

x <- cbind(x0,x1,x2)
y <- x0 + sin(6*pi*x1)
a <- ace(x,y)

par(mfrow=c(2,2))
plot(a$x[1,],a$tx[,1])
plot(a$x[2,],a$tx[,2])
plot(a$x[3,],a$tx[,3])
plot(a$y,a$ty)
```



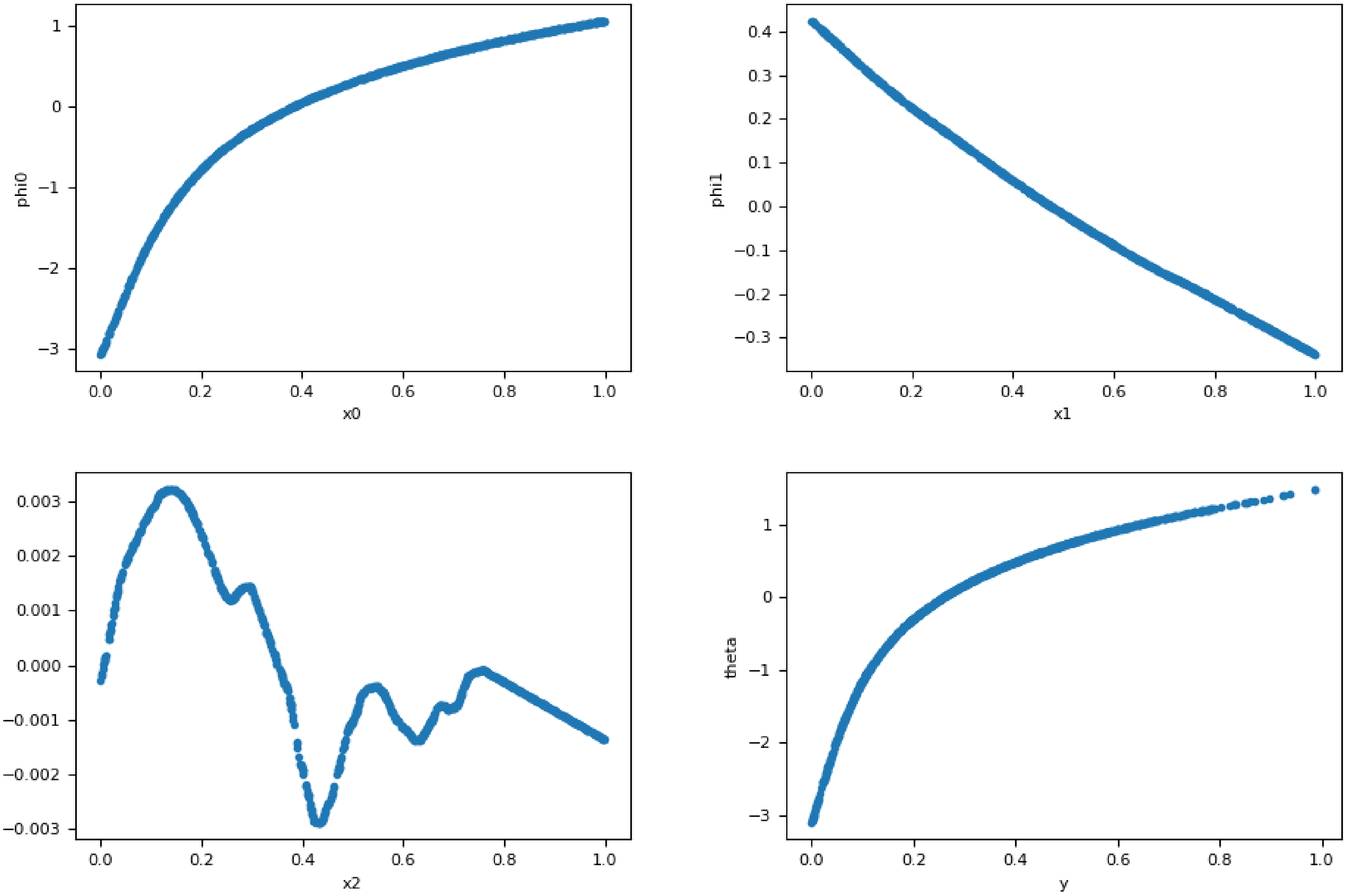
$$\theta(Y)=\alpha+\sum_{i=1}^{p} c_i\phi_i(X_i)+\epsilon$$

$$y=e^{x_0+\sin(6\pi x_1)}$$



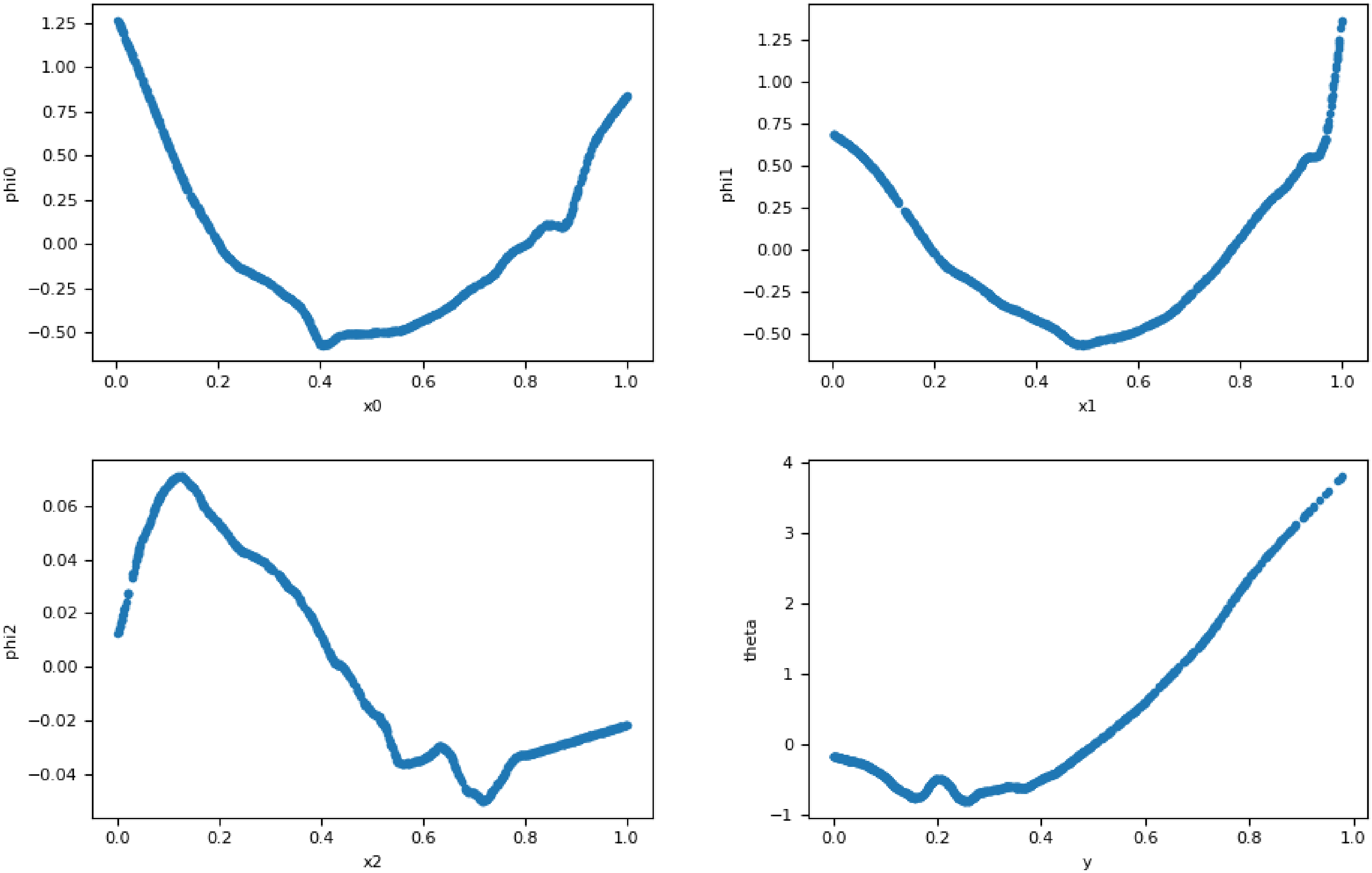
$$\theta(Y)=\alpha+\sum_{i=1}^{p} c_i\phi_i(X_i)+\epsilon$$

$$y=\frac{x_0}{1+x_1}$$



$$\theta(Y)=\alpha+\sum_{i=1}^{p}c_i\phi_i(X_i)+\epsilon$$

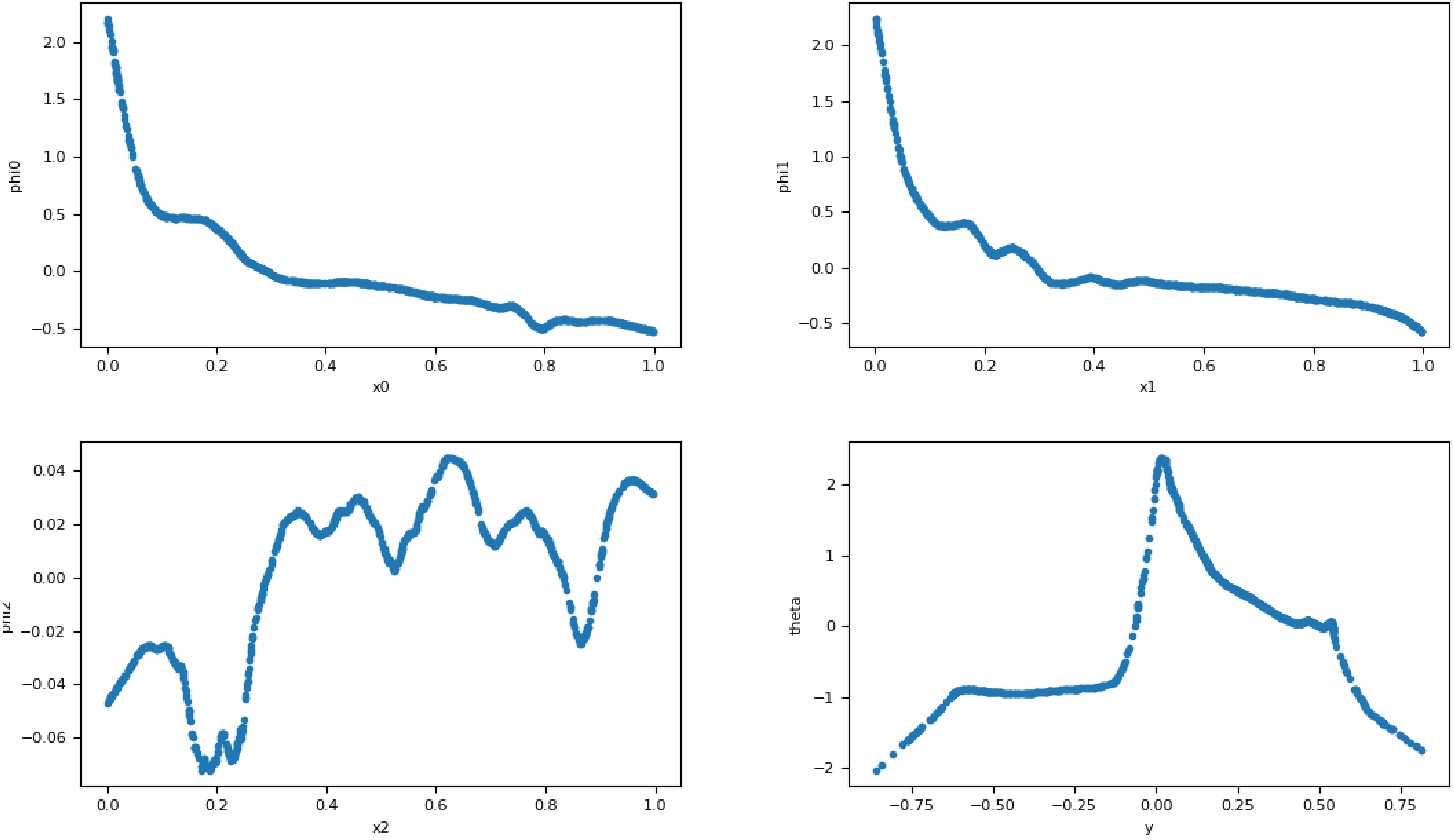
$$y=|x_0-x_1|$$



$$ay^2+by=b(x_0^2-x_0+x_1^2-x_1)+c\quad ???$$

$$\theta(Y)=\alpha+\sum_{i=1}^{p} c_i \phi_i(X_i)+\epsilon$$

$$y=\frac{\sin(6\pi x_0 x_1)}{1+\exp(-2x_0 x_1)}$$



Distance correlation between two paired random vectors of arbitrary, not necessarily equal, dimension (2005, Gábor J. Székely)

$$0 \leq dC(X, Y) \leq 1$$

Theory: https://dcor.readthedocs.io/en/latest/theory.html

R https://cran.r-project.org/web/packages/energy/

XOR

```
X (400, 2) y (400,)
MI: [0 0]
dcor(x0,y): -0.0025
dcor(x1,y): -0.0025
dcor(X,y): 0.252
```

$$y = x_0 + \sin(6\pi x_1)$$

```
dcor(x0,y): 0.124
dcor(x1,y): 0.069
dcor(x2,y): 0.00027
dcor([x0,x1],y): 0.127
dcor([x0,x2],y): 0.0897
dcor([x1,x2],y): 0.0368
```

$$y = |x_0 - x_1|$$

```
dcor(x0,y): 0.027
dcor(x1,y): 0.030
dcor(x2,y): 0
dcor([x0,x1],y): 0.145
dcor([x0,x2],y): 0.012
dcor([x1,x2],y): 0.014
```

$$y = e^{x_0 + \sin(6\pi x_1)}$$

```
dcor(x0,y): 0.126
dcor(x1,y): 0.075
dcor(x2,y): 0.00079
dcor([x0,x1],y): 0.133
dcor([x0,x2],y): 0.091
dcor([x1,x2],y): 0.043
```

$$y = \frac{x_0}{1 + x_1}$$

```
dcor(x0,y): 0.868
dcor(x1,y): 0.085
dcor(x2,y): 0.0008
dcor([x0,x1],y): 0.688
dcor([x0,x2],y): 0.614
dcor([x1,x2],y): 0.062
```

$$y = \frac{\sin(6\pi x_0 x_1)}{1 + \exp(-2 x_0 x_1)}$$

```
dcor(x0,y): 0.073
dcor(x1,y): 0.058
dcor(x2,y): 0
dcor([x0,x1],y): 0.106
dcor([x0,x2],y): 0.05
dcor([x1,x2],y): 0.04
```

Feature Selection

Step Forward Feature Selection

1st step: The performance of the classifier is evaluated with respect to each feature.

The feature that performs the best is selected out of all the features.

2nd step: The first feature is tried in combination with all the other features.

The combination of two features that yield the best algorithm performance is selected.

nth step: The process continues until the specified number of features are selected.

Step Backwards Feature Selection

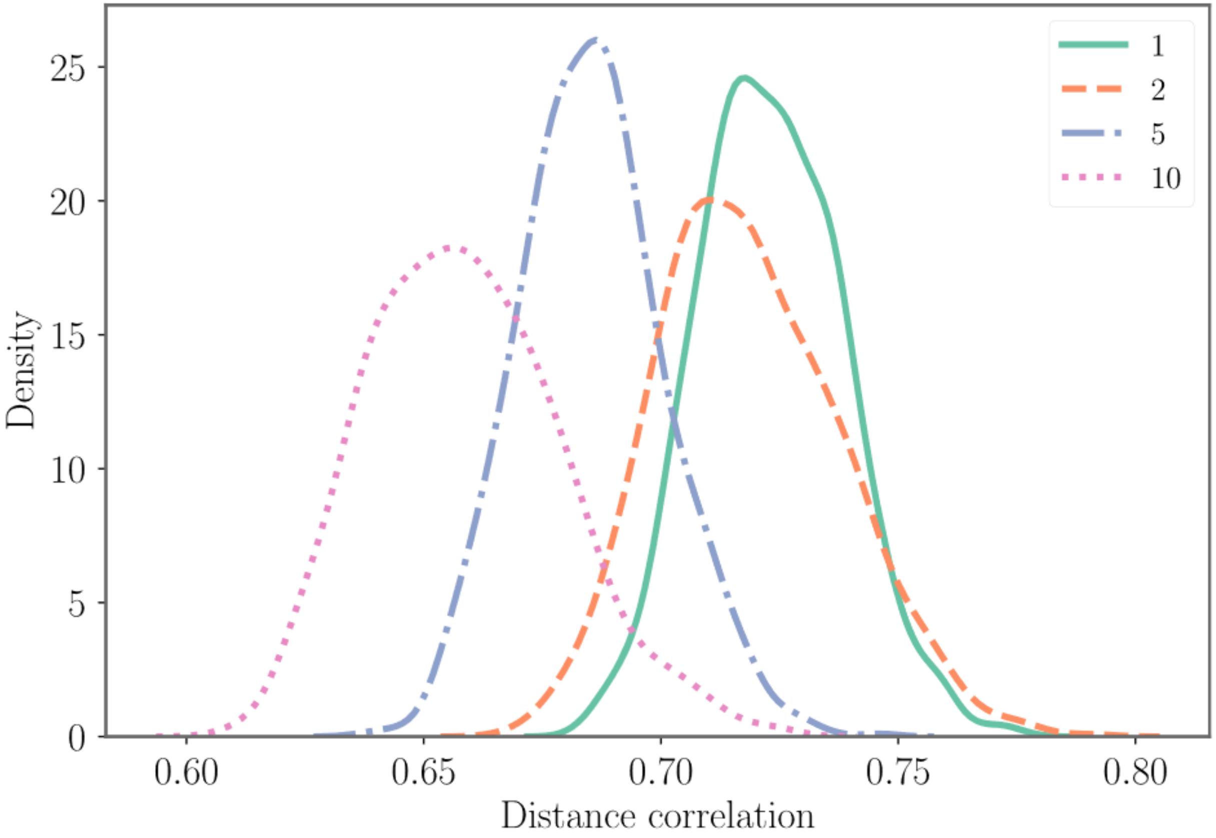
1st step: one feature is removed from the feature set and the performance of the classifier is evaluated.

The feature set that yields the best performance is retained.

2nd step: one feature is removed and the performance of all the combination of features except the 2 features is evaluated.

nth step: This process continues until the specified number of features remain in the dataset.

Uncertainty in feature importance scores coming from the choice of the samples in the data set.



Beam search

beam width

Similar to the step-wise selection but *B* best features are selected during each step.

Each of these *B* features is tried in combination with all the other features.

Then the *B* best combinations are selected for the next step.

Links

Correlation coefficient between a (non-dichotomous) nominal variable and a numeric (interval) or an ordinal variable

https://stats.stackexchange.com/questions/73065/correlation-coefficient-between-a-non-dichotomous-nominal-variable-and-a-numer

Correlations with unordered categorical variables

https://stats.stackexchange.com/questions/108007/correlations-with-unordered-categorical-variables

Different feature importance results between DNN, Random Forests and Gradient Boosted Decision Trees

https://datascience.stackexchange.com/questions/85242/different-feature-importance-results-between-dnn-random-forests-and-gradient-bo

Circular features

https://stats.stackexchange.com/questions/148380/use-of-circular-predictors-in-linear-regression