Data Transformation in Clustering

Maximilian Fiedler, supervised by Prof. Dr. Peter Gritzmann

Chair of Applied Geometry and Discrete Mathematics

TopMath Mathematik mit Promotion

Technische Universität München

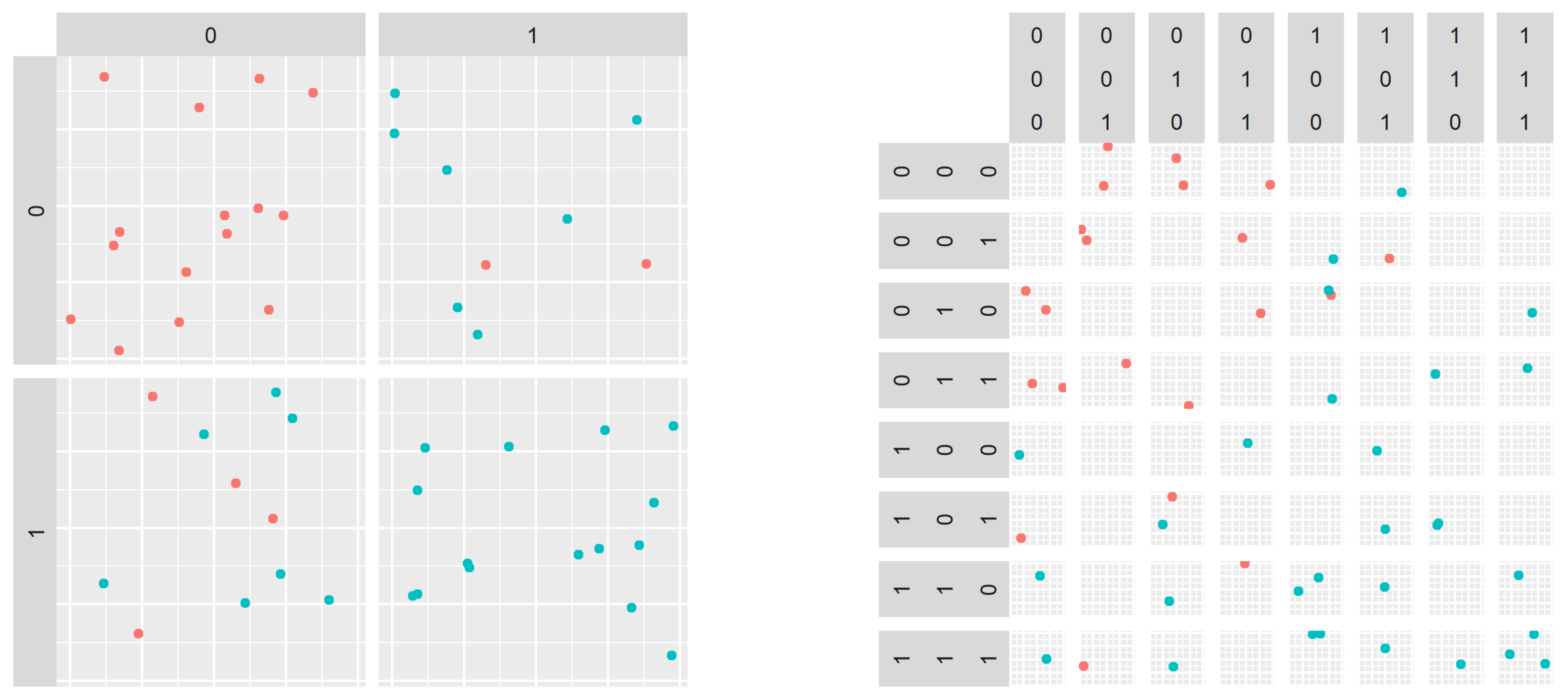
Elitenetzwerk Bayern

Universität Augsburg University

Abstract

In Learning, traditional methods like regression have mostly been developed for continuous input data and categorical variables are often recoded into several binary variables. When recoding, the dimension of the input data can significantly increase and lead to problems caused by the curse of dimensionality. The question arises whether we can transform categorical data more appropriately. The value difference metric is one way of computing a distance between both continuous and categorical variables. This metric learns the distance between input variables from a given data set by estimating conditional probabilities.

Curse of Dimensionality



Assume we are given a data set of 50 instances on two binary variables (left) and six binary variables (right). The instances are plotted in their corresponding cells, the color reflects their known label. For a new instance with unknown label we can predict its label by choosing the color that appears most often for the given input values of this new instance. While this is a reasonable prediction in low dimensions it does not work for high dimensions. In the right picture we see that for some input value combinations there are no instances in our data set available.

Curse of Dimensionality

To retain the quality of a prediction the number of necessary instances in the data set increases exponentially with the dimension.

Linear Modeling (Dummy Coding and Error)

Linearity assumption

Given a data set of instances $(y_i \in \mathbb{R}, \boldsymbol{x}_i \in \mathbb{R}^p)_{i\in[n]}$ with x_i the input vector and y_i the label of the i-th instance, linear modeling assumes:

$$\mathbb{E}[y_i] = \beta_0 + \beta_2 x_{i1} + ... + \beta_p x_{ip}$$

For categorical variables this assumption does not hold. Hence variables are recoded such that for each category a β-coefficient is estimated.

Dummy Coding

Let x be a categorical variable with k categories. Dummy coding replaces x by $k-1$ binary variables $x^1, ..., x^{k-1}$ which are coded as follows

$$x^j = \begin{cases} 1 & \text{if } x \text{ takes category } j \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } j \in [k-1].$$

However this recoding decreases the expected prediction error.

Expected Prediction Error in Linear Modeling

The expected prediction error of a new instance $\boldsymbol{x}_0$ in linear modeling is

$$\text{EPE}(\boldsymbol{x}_0) = \mathbb{E}[(y_0 - \beta_0 + \beta_2 x_{01} + ... + \beta_p x_{0p})^2]$$

and decomposes into an irreducible error the squared bias and the variance. This motivates the EPE as a reasonable quality criterion. Under some mild assumptions on the distribution of new instances $\boldsymbol{x}_0$, we compute

$$\mathbb{E}_{\boldsymbol{x}_0}\text{EPE}(\boldsymbol{x}_0) = \sigma^2 + \sigma^2\frac{(p+1)}{n}$$

where σ^2 is the variance of the assumed error in linear modeling. From the fact that the expected prediction error increases with the number of variables p in the model, we conclude that categorical variables with many categories can highly decreases the quality of the prediction of a linear model. Is there a way of learning that can handle categorical variables without increasing the dimension of the data?

Value Difference Metric for Categorical Variables

To apply geometric methods to data, often a notion of distance is necessary. But how can one define a distance between categorical variables that reflects the similarity of categories with respect to the label? The idea of the value difference metric is to compare the conditional probabilities of two categories to have the same label. Since the value difference metric has been developed for the purpose of classification, it allows in its original form only finitely many possible labels c_i, $i \in [C]$. In our example the label has two possible values, namely red and blue.

Value Difference Metric

Let x be a categorical variable and let $(x_i|c_i)$, $i \in [n]$, be n instances where the input vector contains data on the categorical variable x and the label takes one of a finite number $C \in \mathbb{N}$ of classes $c^j, j \in [C]$. The distance between two categories a, b of the categorical variable x measured by the value difference metric is given by:

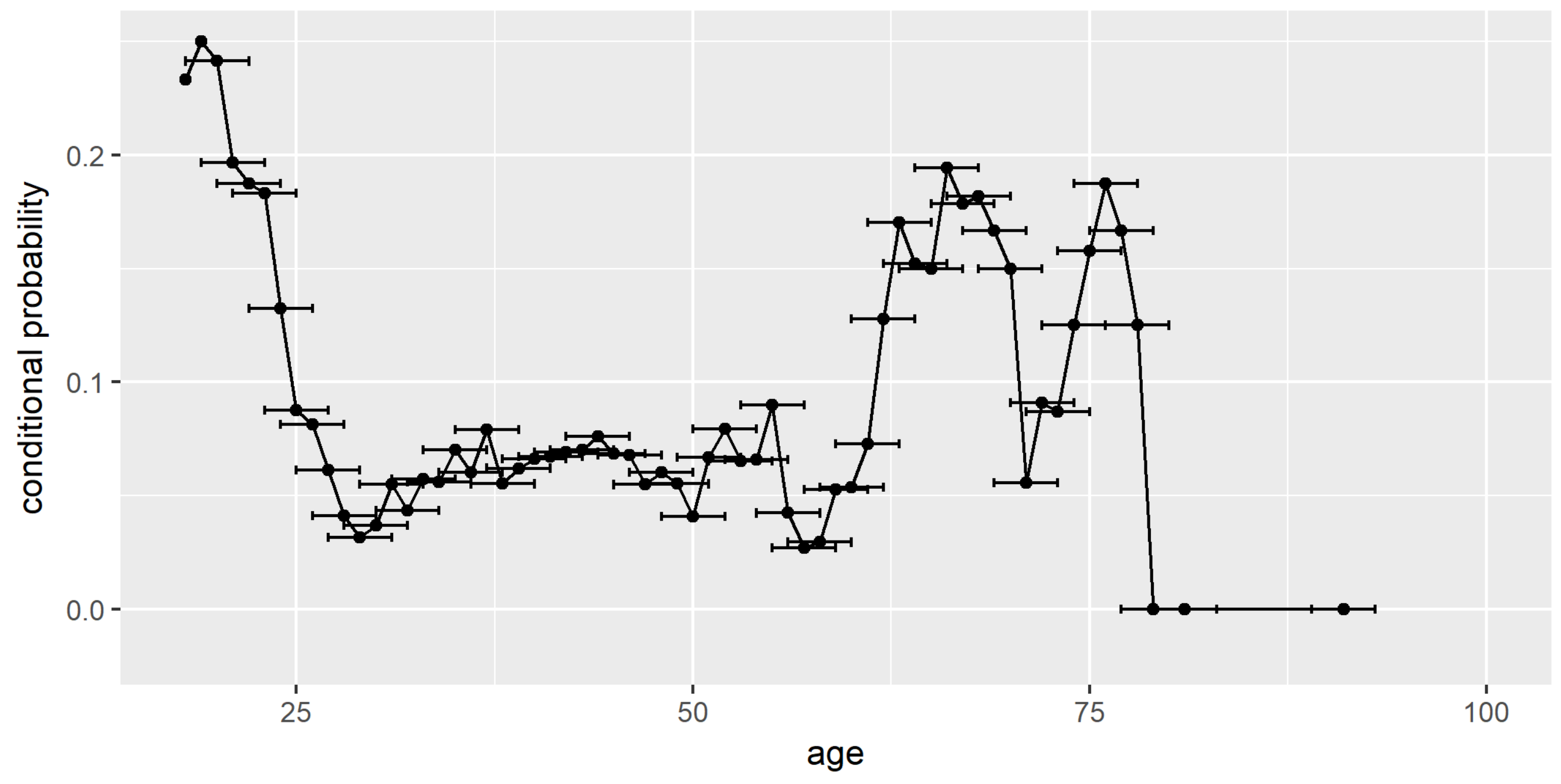
$$\text{vdm}_x(a,b) = \sum_{j=1}^{C} \left| P_{x,a,c^j} - P_{x,b,c^j} \right|$$

where

- $P_{x,a,c} = \frac{N_{x,a,c}}{N_{x,a}}$ is an estimate of the conditional probability that for an instance i with $x_i = a$ the label is given by $c_i = c$.
- $N_{x,a}$ is the number of instances $i \in [n]$ with $x_i = a$.
- $N_{x,a,c}$ is the number of instances $i \in [n]$ with $x_i = a$ and $c_i = c$.

Value Difference Metric for Continuous Variables

The above definition of the value difference metric is not reasonable for continuous variables. The estimation of the conditional probabilities from the data set requires a sample size large enough for each value of a variable. However, for continuous variables each value is most likely to be unique and hence the sample size will be one. A more sophisticated idea is to treat a window of observations around each value of a continuous variable as a category and then estimate the conditional probability of this window. The following image depicts this idea for the age variable in a data set.



Using the Value Difference Metric for Prediction

The value difference metric has the advantage of treating both categorical and continuous variables the same by comparing their conditional probabilities and thus does not need to recode categorical data. Rethinking our label estimation in dimension six, we could instead of only considering instances that are in the same cell as a new input vector also consider similar instances for the prediction of the color. The main question is how to identify similar instances. The introduced value difference metric gives us an appropriate way of quantifying this similarity and can improve the prediction.

References

1. Stanfill, Craig, and David Waltz. "Toward memory-based reasoning." Communications of the ACM 29.12 (1986): 1213-1228.
2. Wilson, D. Randall, and Tony R. Martinez. "Value difference metrics for continuously valued attributes." Proceedings of the International Conference on Artificial Intelligence, Expert Systems and Neural Networks. 1996.
3. Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning. Vol. 1. Springer, Berlin: Springer series in statistics, 2001.