Joint Probability-based Dissimilarity Measure for Discrete Variables

Harshvardhan

ASSIGNMENT

Instructor's Name: Prof. Wenjun Zhou

Course: Machine Learning

Date of Submission: April 4, 2022

Most clustering algorithms work for numerical variables where the variables are assumed to be continuous and random. In this short monograph, I propose a probability-based distance measure for computing dissimilarity between observations for discrete variables thought to be randomly distributed. As their probabilities are derived empirically, there is no underlying assumption on their distribution.

1 Background

Clustering is the task of grouping a set of observations together such that observations that are similar to one another belong to one group. Although its primary purpose is in grouping observations for exploratory data analysis, the method is currently used in many fields. The rise of machine learning has accentuated its popularity.

With only numerical data at hand, the problem is relatively simple. We choose a distance measure and keep observations "close" to one another, maximising the distance between the cluster centres. However, the distance methods with categorical data are not obvious.

Various researchers have studied the problem in detail. Gower's distance is a popular measure that can calculate the dissimilarity between logical, numerical, categorical, and even textual data [2]. The original paper by Gower [1] is lucid enough for readers to understand.

I propose a joint probability-based measure of dissimilarity between observations based on the joint probability of two discrete variables. The underlying assumption is that the discrete variables are independent of each other.

A generative model in statistics that assumes the observations were sampled from a natural population with a joint probability defined by $P(X, Y)$, where X is the observable variable, and Y is the target variable. This contrasts with a discriminative model of the conditional probability that assumes that the target Y was produced given an observation $X = x$, i.e. determined by $P(Y|X = x)$.

Clustering with unknown targets is a generative process rather than a discriminative process. Consider a natural generative process that creates a set of observations X. We do not know in advance which Y does an observation belong. Since this is mainly non-experimental data, we are trying to find which observations are likely similar.

2 Method

Consider two discrete random variables X_1 with u different classes and X_2 with v different classes. Let $\{c_{11}, c_{12}, \ldots, c_{1u}\}$ be the set of different classes of X_1. Similarly, let $\{c_{21}, c_{22}, \ldots, c_{2v}\}$ be the set of different classes of X_2. The empirical probability of event $X_1 = c_i$ is $\frac{m}{n}$, where m is the frequency of c_i observed in X_1 and n is the total number of observations.

Assuming that the sample is representative of the population, we can calculate the empirical probability of each class for each variable. Once we have those probabilities, we can calculate the joint probability for an observation that I call “score”. This score is a number between 0 and 1.

Interpretation The score of zero is asymptotically possible but impossible in real-world analysis. If the researcher assumes no prior knowledge about the variable, only the existing classes observed in the data can be used as a possible class. In that case, the score cannot be zero for any observation. However, if the researcher assigns a non-zero probability to a class that wasn’t observed in real data, we can have zero probability for some classes. A score of one is possible only when all observations are precisely the same.

In most general cases, the value for each observation would lie between zero and one. The closer the values are to each other, the closer they are to each other (although this is not guaranteed, as we will see in the following example.)

Pros and Cons

The proposed method is amazing if we do not assume any prior probabilistic distribution for the variables. Since it relies on empirical distribution, it estimates the class probability for a discrete variable only based on available observations. However, this benefit comes

Sl No	Sex	City	Colour	Executive	Score
1	Male	Shanghai	Blue	Yes	$12/625 = 0.0192$
2	Female	New York	Blue	No	$24/625 = 0.0384$
3	Female	New Delhi	Black	Yes	$9/625 = 0.0144$
4	Male	New York	White	Yes	$12/625 = 0.0192$
5	Female	Boston	Red	No	$6/625 = 0.0096$

Tab. 1: This dummy data was created for the example. As you can observe, different variables have different probabilistic distributions.

at a (potential) cost. A biased sample will significantly affect the empirical probability and thus the score. It may not be reliable in such cases.

It is also possible that this method will lead to combinatorial explosion and thus very small values of the score. When calculating the empirical probabilities, we will typically have small values — less than 0.3 if there are three classes, say. If there are five such variables, the "average" score would be $0.3^5 = 0.00243$, which is very small.

This limitation has an easy fix. We could easily scale the score by multiplying it by a large C to bring it on the same scale as the rest of the variables. This will ensure that the clustering algorithm doesn't penalise this variable for a small default value.

3 Example

Let me illustrate the method with a small example. Consider the following data with three discrete variables and no continuous variable (Table 1).

The variables have different probability distributions. The probability of being a Male is 2/5; being a Female is 3/5. The probability of the City being New York is 2/5; Shanghai, Boston or New Delhi are all equal to 1/5 each. The probability of the favourite colour being Blue is 2/5; Black, White or Red are at 1/3. Finally, being an executive is 3/5, and the probability of being a non-executive is 2/5.

Assuming that all variables are independent of each other, the probability that a person is Male who lives in Shanghai, whose favourite colour of Blue and who is an executive is $2/5 * 1/5 * 2/5 * 3/5 = 12/625 = 0.0192$. I call this joint probability an observation's score. We could repeat the exercise for all the observations, and we will obtain the results presented in the last column of Table 1.

This continuous measure that I call "Score" can measure dissimilarity between observations. Note that the method doesn't guarantee a differentiable score. Even observations with which we get precisely the same can differ from one another. However, observations with very different scores would inevitably be different observations. The latter property is more critical when deciding which cluster an observation belongs to.

4 Simulations

In this section, I will compare the clusters found using three methods: (1) using only continuous variables, (2) using continuous variables and the score, and (3) using Gower's distance. For the purpose of this simulation, I will use `flower` data available in `cluster` package in R. [1] It is a small dataset with 18 observations, six discrete variables and two continuous variables. A quick overview of the dataset is provided in Table 2.

Variable	Description	Type	# of Unique Values
V1	Indicates whether the plant may be left in the garden when it freezes.	Binary	2
V2	Shows whether the plant needs to stand in the shadow.	Binary	2
V3	Distinguishes between plants with tubers and plants that grow in any other way.	Binary	2
V4	Specifies the flower's color (1 = white, 2 = yellow, 3 = pink, 4 = red, 5 = blue).	Discrete	5
V5	Indicates whether the plant grows in dry (1), normal (2), or wet (3) soil.	Discrete	3
V6	Gives someone's preference ranking going from 1 to 18.	Discrete	18
V7	The plant's height in centimeters.	Continuous	—
V8	The distance in centimetres that should be left between the plants.	Continuous	—

Tab. 2: Summary of the dataset `flower`. We have 18 observations in total and eight variables. Six variables are discrete or binary (i.e. categorical) and two are continuous.

Results and Discussion

The clusters obtained from the continuous variables seem to have accounted only for V7 in differentiating between the observations. See Figure 1 for a scatter plot. The clusters obtained with the continuous variable and the score are presented in Figure 2. All the points in cluster 1 are identified as the same. However, points in cluster 2 in the second method include points that were represented by cluster 3 in the first method.

[1] For more details, see `https://cran.r-project.org/web/packages/cluster/cluster.pdf`. This dataset was first published by Struyf, Hubert and Rousseeuw (1996).

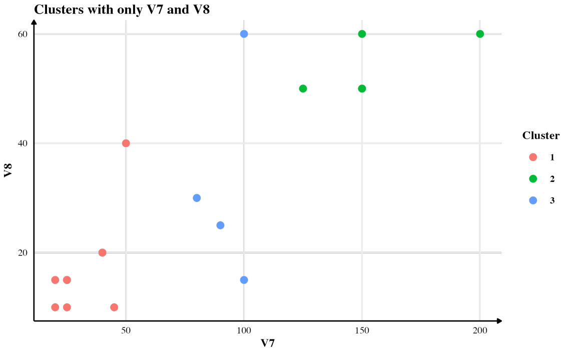

Fig. 1: Clusters obtained with only continuous variables: V7 and V8.

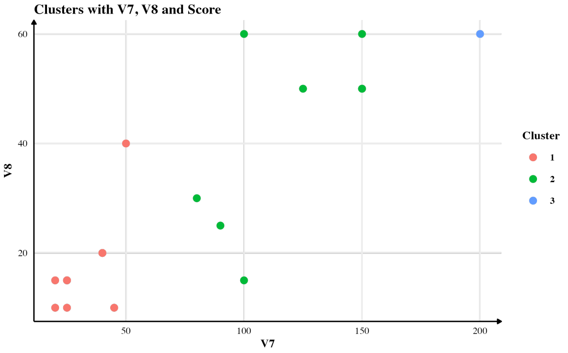

Fig. 2: Clusters obtained with only continuous variables: V7 and V8.

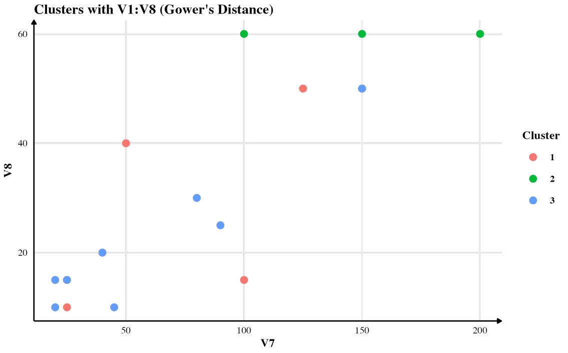

Fig. 3: Clusters obtained with all variables. Distance was calculated using Gower's distance metric.

5 Concluding Remarks

In this short monograph, I presented a new distance metric based on empirical joint probability. With a small simulation on `flower` data, I showed how effective it is as compared to not using categorical variables at all. I also compared the results with Gower's distance-based clustering. I found that the results from the three methods do not match exactly. However, my method shows some improvement over not using categorical variables.

References

[1] J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871, 1971.

[2] Mark van der Loo. *gower: Gower's Distance*, 2022. R package version 1.0.0.