日本行動計量学会第 48 回大会抄録集、2020、212－215.（原稿）

One Hot Encoding と共通数量化

カテゴリ変量の前処理

○岡本安晴
日本女子大学

1　はじめに

調査紙あるいは質問紙のデータはカテゴリ値で得られることが多い。カテゴリ値データを数値データに変換して分析するための前処理として機械学習において One Hot encoding と呼ばれている方法がある。これに類似した方法として共通数量化（Okamoto, 2015；岡本，2019）がある。One Hot encoding では、各カテゴリ値にダミー変数を割り当てられるが、共通数量化ではカテゴリ変量ごとに数量化が行われる。共通数量化においてもダミー変数が使われるが、その意味は異なる。具体的に説明する。

表１　カテゴリデータ

個人	カテゴリ値
1	A
2	C
3	B
4	A

表１はカテゴリ値 A、B、C を取る変量の４人の値である。カテゴリ値に対応してダミー変数 DA、DB、DC を用意すると、表１のデータは表２のようになる。ダミー変数は、データ値が対応するカテゴリ値であれば値１をとり、対応する値でなければ値０をとる数値変量であり、カテゴリ変数をダミー変数で置き換えて０と１に数値化することは One Hot encoding と呼ばれている。

表２　ダミー変数

個人	DA	DB	DC
1	1	0	0
2	0	0	1
3	0	1	0
4	1	0	0

これに対して共通数量化では、カテゴリ A、B、C に数値w_A、w_B、w_Cが割り当てられる（表３）。この数値の割り当ては、データ（表１）の情報が最もよく反映されるように行われる。すなわち、数量化されたデータ値w_Xの分散が最大になるように数値が割り当てられる。数量化されたデータ（表３）は、次式で表すことができる。

表３　数量化（カテゴリ：数値）（A：w_A）（B：w_B）（C：w_C）

個人	数量化
1	w_A
2	W_C
3	w_B
4	w_A

（１）
$$\begin{bmatrix} w_A \\ w_C \\ w_B \\ w_A \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} w_A \\ w_B \\ w_C \end{bmatrix} = G \begin{bmatrix} w_A \\ w_B \\ w_C \end{bmatrix}, G = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

ここで、Gは indicator matrix と呼ばれるものである（Gifi, 1990）。Gは、One Hot encoding（表２）と同じ０と１の並びであるが、One Hot encoding では、０と１はカテゴリ値に対応する数値として与えられている。共通数量化では、与えられている数値はw_Xであり、Gにおける０あるいは１は、カテゴリ値Xに数値w_Xを割り当てる数式を構成するための形式として設定されている。

One Hot encoding における 1 は対応するカテゴリ値の値であり、共通数量化においてカテゴリ値 Xに与えられる数値はw_Xである。数値計算上は、式（1）の数量化の最大化は、Gの主成分分析の数値計算と同じである。しかし、共通数量化では、最大化はカテゴリ変量ごとであり、その解釈は当該変量におけるカテゴリ値全体にわたって行われる。One Hot encoding では、カテゴリ変量にわたってカテゴリ値のダミー変数の分析が行われ、解釈はカテゴリ値ごとである。

共通数量化と One Hot encoding には以上のような違いがあるが、いずれが優れているかという観点から考えるべきではない。一般に機械学習では所与のデータに対して複数の分析法が試みられ、それらの分析結果が比較されるが、これはある特定の１つの分析法が常に優れているわけではないことによる（Raschka & Mirjalili, 2019）。カテゴリ変量の前処理も、One Hot encoding だけではなく共通数量化も試みることにより、データからより豊かな分析結果の得られることが期待される。One Hot encoding と共通数量化の比較のために、全変量のカテゴリ値のダミー変量をまとめて分析した場合と、共通数量化の分析の場合との比較例を以下に報告する。

2　データ分析例

岡本（2019）の表 8.2.1 のデータで比較してみる。カテゴリ項目５問、連続量項目２問の質問紙における心理学科１年生３９名のデータであり、質問項目は以下の通りである。

表４　共通数量化の値

	Q1-0	Q2-0	Q3-0	Q4-0	Q5-0
w_A	0.40825	-0.61920	0.34889	0.09878	-0.80075
w_B	-0.81650	0.77052	-0.81375	-0.75130	0.53859
w_C	0.40825	-0.15132	0.46486	0.65252	0.26216
	Q1-1	Q2-1	Q3-1	Q4-1	Q5-1
w_A	-0.70711	0.53222	-0.73820	-0.81050	-0.15960
w_B	-0.00000	0.27013	0.06695	0.31970	-0.61367
w_C	0.70711	-0.80235	0.67125	0.49080	0.77327

問１　心理学という分野があることを、A.小学生の頃には知っていた。B.中学生の頃に知った。C.高校生の頃に知った。D.その他。

問２　大学受験のとき、A.心理学≒臨床心理学と思っていた。B.心理学には臨床心理学以外の分野があることを知っていた。C.その他（「心理学とは何をする分野なのか知らなかった。」など）。

問３　大学受験のとき、卒業後は何になりたいと考えていましたか？A.カウンセラーになりたい。B.カウンセラーではないが、心理学を活かした職業に就きたい。C.その他。　**問４**　いま、卒業後は何になりたいと考えていますか？A.カウンセラーになりたい。B.カウンセラーではないが、心理学を活かした職業に就きたい。C.その他。　**問５**　自分の得意分野は A.文系であると思う。B.理系であると思う。C.その他。　**問６**　心理学科に入学したことにどの程度満足していますか。非常に満足しているなら１００点、全然満足していないならー１００点、どちらでもないなら０点、という具合に、－１００点から１００点までの数値で心理学科に進学したことの満足度を表すと、あなたの満足度は何点になりますか。（　　）点

問７　受験生から、どこに進学するのが良いか相談を受けたとしたら、本学の心理学科をどの程度薦めますか。強く薦めるなら１００点、強く反対するならー１００点、薦めも反対もしないなら０点、という具合に、－１００点から１００点までの数値で、本学心理学科をどの程度薦めた

いか評定して下さい。（　　）点

問１において回答として「その他」を選んだ者はいなかったので、問１も実質的には回答カテゴリ数は３カテゴリである。問１から問５までの共通数量化の値を表４に示す。カテゴリ数が３であるので、共通数量化の値の組は、各問いについて２組である。問iの２つの共通数量化変量を Qi-0 および Qi-1 と表す。Qi-0 の方が Qi-1 より対応する固有値が大きく、説明率が高い。共通化変量 Q1-0 から Q5-1、および問６と問７（それぞれ Q6 と Q7 で表記）をまとめて主成分分析すると図１の結果を得る。

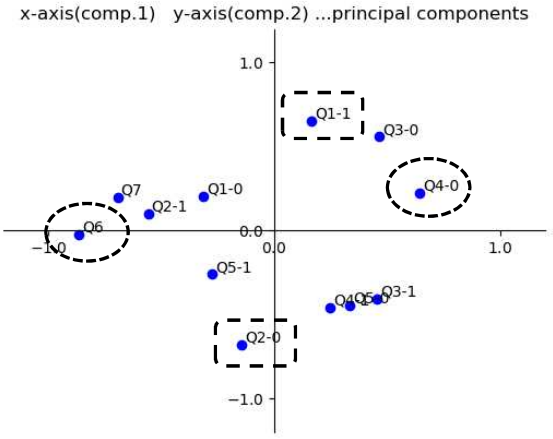

図１　数量化変数の主成分分析

共通数量化 Q4-0 と連続変量 Q6 が左右反対側にあるという関係が表されているので、Q4-0 を独立変数、Q6 を従属変数として回帰直線を求めると図２のようになる。共通数量化変量 Q4-0 の表すカテゴリ値が横軸にそって記されている。卒業後、カウンセラーではないが心理学を活かした職業に就きたいという学生の満足度が、他のグループより高いという傾向が表されている。

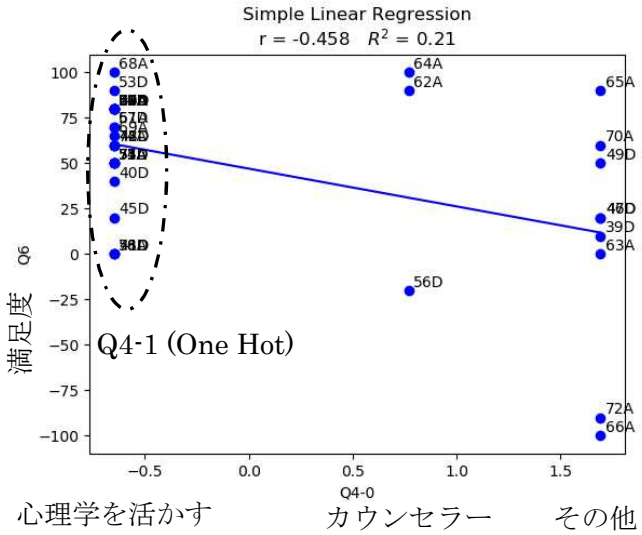

図２　数量化変量 Q4-0 と Q６の関係

図１では、共通数量化変数 Q1-1 と Q2-0 が反対側に位置している。カテゴリ変数 Q1 と Q2 のクロス表を、それぞれの共通化変量w_xの大きさの順に並べたカテゴリ値で作成すると表５のようになる。破線で囲まれた度数のまとまりの分布が読み取れる。問１の共通数量化 Q1-1 は、数量化の大小関係が、心理学を知った年齢に対応している（表４）。問２の共通数量化 Q2-0 の順序は、表４に示されているように表５の列の順序である。中学生までに心理学という学問があることを知っていた場合は、臨床心理学以外にも心理学の分野があることを知っていたが、高校生になってから心理学を知った学生は、心理学とは臨床心理学であると理解していたものの多いことが示されている。

次に、カテゴリ変量を One Hot encoding した場合の分析結果を見る。問iの３つのカテゴリ値

を表すダミー変数を Qi-0、Qi-1、Qi-2 とおく。ダミー変数 Q1-0 から Q5-2、および量的変数 Q6 と Q7 に対して主成分分析を行うと図３の結果を得る。量的変量 Q6 と One Hot encoding の変数 Q4-1 が近くに位置しているが、これは図２において「心理学を活かした職業に就きたい」という学生の満足度が他のグループより高い傾向として表されている。One Hot encoding の変数 Q1-2 と Q2-0 が近くに位置していることは、表５の対応するセルの度数が 6 と高いことに対応しており、Q2-1 と Q1-0 および Q1-1 の位置が互いに近いことは、これらの組み合わせの度数がそれぞれ 10 と高いことに対応している。

表５　共通数量化 Q1-1 と Q2-0、および One Hot encoding Q1-j と Q2-k のクロス表

	心理学≒臨床進学 -0.61920 Q2-0	その他 -0.15132 Q2-2	心理学＞臨床心理学 0.77052 Q2-1
小学生 -0.70711 Q1-0	1	1	10
中学生 0.00000 Q1-1	3	2	10
高校生 0.70711 Q1-2	6	2	4

3　まとめ

カテゴリ変量は、One Hot encoding あるいは共通数量化により量的変量と同じように分析できる。One Hot encoding では個々のカテゴリ値の振る舞いが扱われ、共通数量化では各カテゴリ変量ごとにカテゴリ値がまとめて扱われるので、当該の変量におけるカテゴリ値全体の分布が反映される。個々のカテゴリ値の振る舞いの情報を求めるか、変量ごとにカテゴリ値の分布を反映した情報を求めるか、それぞれの特徴を踏まえた分析が工夫される。

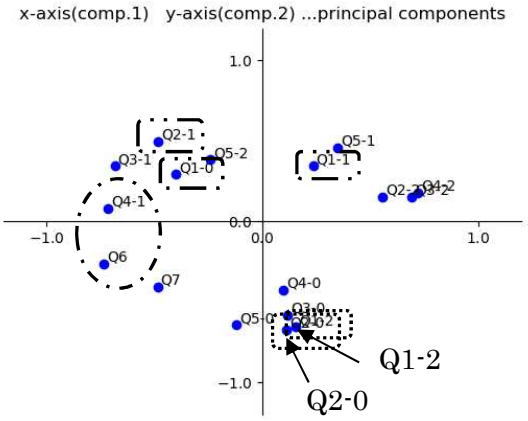

図３　One Hot encoding の主成分分析

4　引用文献

Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester: John Wiley & Sons.

Raschka, S. & Mirjalili, V. (2019). *Python machine learning, third ed.* Birmingham: Packt.

Okamoto, Y. (2015). A two-step analysis with common quantification of categorical data. *Japan Women's Univarsity: Faculty of Integrated Arts and Sciences Journal.* 26, 99-112.

岡本安晴（2019）いまさら聞けない Python でデータ分析．丸善出版.