



THE DEVELOPER'S CONFERENCE

Trilha – Data Science

Renan Faraon Cintra

Graduado em Estatística e Credit Data Science Analyst no Agibank
2019

Utilização da Análise de Classes Latentes para estimar similaridades na detecção de propostas fraudulentas

Agenda



- MOTIVAÇÃO
- ANÁLISE DE CLASSES LATENTES
- TÁ...MAS QUAL É A IDEIA?
- CONSIDERAÇÕES FINAIS



THE
DEVELOPER'S
CONFERENCE

MOTIVAÇÃO

MOTIVAÇÃO



- Quando encontramos algum caso inconsistente ou fraudulento: como revisar casos passados para verificar se o caso encontrado é único?

Por onde começar a procurar?!



THE
DEVELOPER'S
CONFERENCE



MOTIVAÇÃO



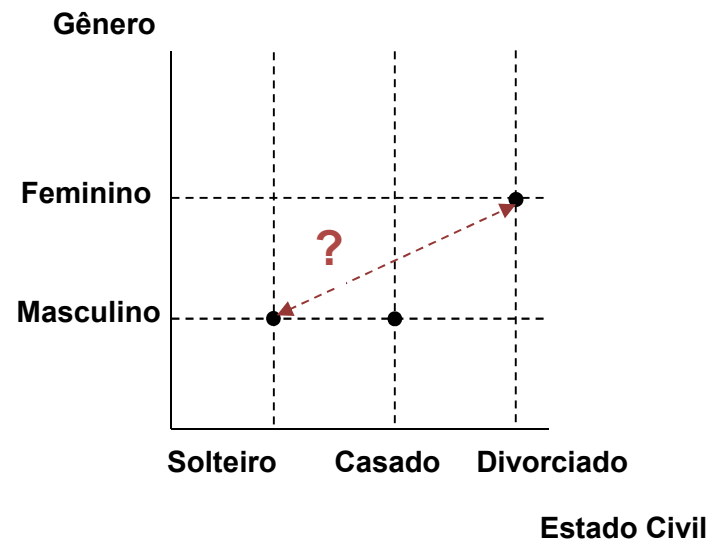
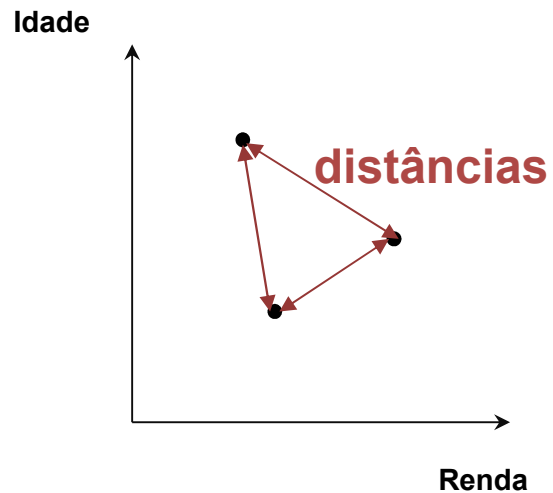
- Utilizar algum método que consiga encontrar **similaridade** entre os diversos casos e, assim, priorizar aqueles que mais se "pareçam" com o caso detectado.



ANÁLISE DE CLASSES LATENTES



Análise de Classes Latentes



Exemplo retirado de: <https://www.youtube.com/watch?v=Anzoq3VDEB0>

Análise de Classes Latentes



THE
DEVELOPER'S
CONFERENCE



Paul Felix Lazarsfeld

The logical and mathematical foundation of latent structure analysis & The interpretation and mathematical foundation of latent structure analysis. 1950



Leo A. Goodman

The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: A modified latent structure approach. 1974

Análise de Classes Latentes



Clusterização

Formação de grupos a partir de padrões de resposta dos indivíduos

Análise Fatorial

Vincula-se à literatura de variáveis latentes

Propício a variáveis categóricas

Probabilidades na definição de grupos

Estatísticas de Ajuste

AIC, BIC...

Análise de Classes Latentes



**Estimando Probabilidade de
pertencimento a uma classe, dado um
padrão de respostas**

$$P(X = x \mid \mathbf{Y} = \mathbf{y}) = \frac{P(X = x)P(\mathbf{Y} = \mathbf{y} \mid X = x)}{P(\mathbf{Y} = \mathbf{y})}$$

**TÁ...MAS QUAL É A
IDEIA?**

Tá...mas qual é a ideia?!



Um banco de dados com muitas
variáveis categóricas

	A	B	C	D
1	ID	Var1	Var2	Var3
2	ID563	CATEGORIA1	FATOR13	NÍVEL3
3	ID579	CATEGORIA2	FATOR14	NÍVEL2
4	ID734	CATEGORIA2	FATOR19	NÍVEL1
5	ID955	CATEGORIA3	FATOR5	NÍVEL1
6	ID48	CATEGORIA2	FATOR25	NÍVEL1
7	ID170	CATEGORIA2	FATOR3	NÍVEL3

Tá...mas qual é a ideia?!



	A	B	C	D	E	F
1	ID	Var1	Var2	Var3	Classe1	Classe2
2	ID563	CATEGORIA1	FATOR13	NÍVEL3	0,718098	0,281902
3	ID579	CATEGORIA2	FATOR14	NÍVEL2	0,424907	0,575093
4	ID734	CATEGORIA2	FATOR19	NÍVEL1	0,447519	0,552481
5	ID955	CATEGORIA3	FATOR5	NÍVEL1	0,902616	0,097384
6	ID48	CATEGORIA2	FATOR25	NÍVEL1	0,042554	0,957446

**Estimar a Probabilidade de
Pertencimento a Classes criadas
através do Análise
De Classes Latentes**

Tá...mas qual é a ideia?!



**Ocorrência
Fraudulenta/Inconsistente**

	A	B	C	D	E	F
1	ID	Var1	Var2	Var3	Classe1	Classe2
2	ID563	CATEGORIA1	FATOR13	NÍVEL3	0,718098	0,281902
3	ID579	CATEGORIA2	FATOR14	NÍVEL2	0,424907	0,575093
4	ID734	CATEGORIA2	FATOR19	NÍVEL1	0,447519	0,552481
5	ID955	CATEGORIA3	FATOR5	NÍVEL1	0,902616	0,097384
6	ID48	CATEGORIA2	FATOR25	NÍVEL1	0,042554	0,957446
7	ID170	CATEGORIA2	FATOR3	NÍVEL3	0,835579	0,164421

Tá...mas qual é a ideia?!



	ID563	ID579	ID734	ID955
ID563	0	0,414634	0,382656	0,260946913
ID579	0,414634	0	0,031978	0,675581286
ID734	0,382656	0,031978	0	0,643603214
ID955	0,260947	0,675581	0,643603	0
ID48	0,955363	0,540729	0,572707	1,216310
ID170	0,166143	0,580777	0,548799	0,094804415

**Estimar a distância euclidiana
A partir das probabilidades**

Tá...mas qual é a ideia?!



**Ranking dos casos mais
Próximos à primeira
detecção**

Ranking	ID48
0,1293651	ID826
0,3393749	ID677
0,5407290	ID579
0,5727071	ID734
0,9553634	ID563
1,1215059	ID170
1,1720078	ID18
1,2115554	ID170

Tá...mas qual é a ideia?!

1.

Partir de um banco de dados cheio de variáveis categóricas

	A	B	C	D
1	ID	Var1	Var2	Var3
2	ID563	CATEGORIA1	FATOR13	NÍVEL3
3	ID579	CATEGORIA2	FATOR14	NÍVEL2
4	ID734	CATEGORIA2	FATOR19	NÍVEL1
5	ID955	CATEGORIA3	FATOR5	NÍVEL1
6	ID48	CATEGORIA2	FATOR25	NÍVEL1
7	ID170	CATEGORIA2	FATOR3	NÍVEL3

2.

Estimar a Probabilidade de Pertencimento a Classes criadas através do Análise De Classes Latentes

	A	B	C	D	E	F
1	ID	Var1	Var2	Var3	Classe1	Classe2
2	ID563	CATEGORIA1	FATOR13	NÍVEL3	0,718098	0,281902
3	ID579	CATEGORIA2	FATOR14	NÍVEL2	0,424907	0,575093
4	ID734	CATEGORIA2	FATOR19	NÍVEL1	0,447519	0,552481
5	ID955	CATEGORIA3	FATOR5	NÍVEL1	0,902616	0,097384
6	ID48	CATEGORIA2	FATOR25	NÍVEL1	0,042554	0,957446

3.

Pegar um caso que sabemos Que é uma ocorrência atípica

	A	B	C	D	E	F
1	ID	Var1	Var2	Var3	Classe1	Classe2
2	ID563	CATEGORIA1	FATOR13	NÍVEL3	0,718098	0,281902
3	ID579	CATEGORIA2	FATOR14	NÍVEL2	0,424907	0,575093
4	ID734	CATEGORIA2	FATOR19	NÍVEL1	0,447519	0,552481
5	ID955	CATEGORIA3	FATOR5	NÍVEL1	0,902616	0,097384
6	ID48	CATEGORIA2	FATOR25	NÍVEL1	0,042554	0,957446
7	ID170	CATEGORIA2	FATOR3	NÍVEL3	0,835579	0,164421

4.

Estimar a distância euclidiana A partir das probabilidades

	ID563	ID579	ID734	ID955
ID563	0	0,414634	0,382656	0,260946913
ID579	0,414634	0	0,031978	0,675581286
ID734	0,382656	0,031978	0	0,643603214
ID955	0,260947	0,675581	0,643603	0
ID48	0,955363	0,540729	0,572707	1,216310
ID170	0,166143	0,580777	0,548799	0,094804415

5.

Ranking dos casos mais Próximos à primeira detecção

Ranking	ID48
0,1293651	ID826
0,3393749	ID677
0,5407290	ID579
0,5727071	ID734
0,9553634	ID563
1,1215059	ID170
1,1720078	ID18
1,2115554	ID170

CONSIDERAÇÕES FINAIS

CONSIDERAÇÕES FINAIS



- Dada a descoberta de uma fraude ou inconsistência, como rever casos passados com potencial de apresentar o mesmo tipo de problema?
- Uma abordagem que destaque a similaridade entre diversas ocorrências permitiria definir quais são as mais próximas à primeira ocorrência com fraude ou inconsistência encontrada.
- A análise de classes latentes possibilita a criação de *clusters*. Essa técnica também possui diversas vantagens: variáveis categóricas, estatísticas de ajuste e estimação de probabilidades.
- Através das probabilidades estimadas de pertencimento a classes geradas, pode-se usar um método de distância para ordenar quais casos são mais similares e diminuir custos operacionais.

Referências Interessantes

Mastella, Juliana Obino. Análise de classes latentes : da teoria à prática. Trabalho de Conclusão de Curso, UFRGS, 2015.

Link: <https://lume.ufrgs.br/handle/10183/133682>

Latent Class Analysis (LCA) in R with poLCA package for beginner .

Link: <https://www.youtube.com/watch?v=LL6tio9V-Vw>

Latent Class Analysis for Excel.

Link: <https://www.youtube.com/watch?v=Anzoq3VDEB0&t=123s>

poLCA: Polytomous variable Latent Class Analysis package.

Link: <https://cran.r-project.org/web/packages/poLCA/poLCA.pdf>

Statistics How To. Latent Class Analysis / Modeling: Simple Definition, Types.

Link: <https://www.statisticshowto.datasciencecentral.com/latent-class-analysis-definition/>

Vermunt J K, Magidson, J. Latent Class Analysis. Link:

<https://pdfs.semanticscholar.org/8149/dc67491a4fc68a2cb4281d26fc42033ced0f.pdf>

CONTATO



<https://github.com/renancintra>



<https://www.linkedin.com/in/renan-cintra-2b830011a/>



THE DEVELOPER'S CONFERENCE