Misturas Finitas de Especialistas de Regressão: Estimação, Agrupamento e Aplicações

Pedro Henrique Corrêa de Almeida
Orientadora: Dra. Camila Borelli Zeller

Outubro, 2023

Introdução

A modelagem baseada em misturas finitas de distribuições é uma área com uma quantidade enorme de aplicações, que vem experimentando um crescimento muito rápido nos últimos anos.

O objetivo geral deste projeto é dar continuidade às pesquisas que vêm sendo desenvolvidas. Este projeto trata do problema de capturar a heterogeneidade via análise de regressão utilizando misturas finitas de especialistas.

Agrupamento

Métodos de agrupamento são utilizados com o objetivo de agrupar observações em grupos homogêneos diferentes entre si. Nessa área existem métodos baseados em algoritmos (como $k\text{-}means$) e baseados em modelos estatísticos. Dentre esses baseados em modelos estão as misturas finitas que iremos nos aprofundar nessa apresentação.

Os métodos de agrupamentos são classificados como métodos não supervisionados, dado que não a priori qual grupo a observação pertence.

- Análise multivariada

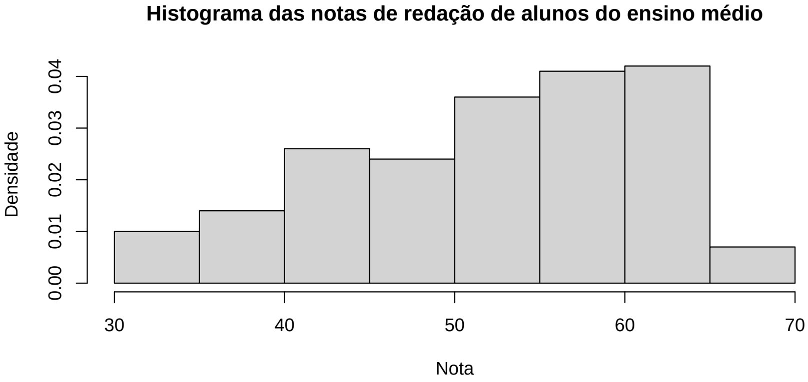
Misturas finitas

A expressão do modelo de misturas finitas é dada por

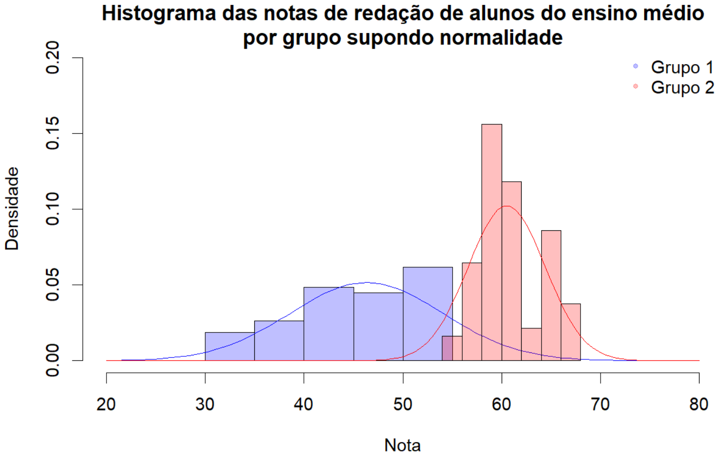
$$f(y_i;\boldsymbol{\Theta}) = \sum_{j=1}^{G} p_j \; g_j(y_i;\boldsymbol{\theta_j}) \tag{1}$$

onde G é o número de grupos, p_j é a probabilidade a priori de uma observação pertencer ao grupo j, em que $\sum_{j=1}^{G} p_j = 1$, e $g_j(.)$ é a função de densidade de probabilidade do grupo j, baseada na suposição de uma distribuição dos dados. Neste trabalho foram consideradas as distribuições Normal e t-Student, bem como suas respectivas distribuições assimétricas skew-Normal e skew-t.

Misturas finitas



Misturas finitas



Regressão Linear

Um modelo de regressão linear é definido por $Y_i = \boldsymbol{X_i}^T \boldsymbol{\beta} + \epsilon_i, i = 1, ..., n$, onde Y_i representa a variável resposta da observação i, $X_i = (x_i 1, ..., x_i p)^T$, $n > p$, é um vetor das covariáveis para o indivíduo i e $\boldsymbol{\beta}$ é um vetor p-dimensional desconhecido com os coeficientes de regressão.

Dessa forma, o interesse é estimar os coeficientes da regressão, e, para isso assumimos que ϵ segue uma distribuição, usualmente $\epsilon \sim N(0, \sigma^2)$.

- Análise de Regressão

Misturas finitas de regressão

As misturas finitas de regressão tem como objetivo combinar as técnicas de agrupamento e de regressão em um modelo. Dessa forma, consideramos que, seja Z_{ij} uma variável latente multinomial, que assume 1 se i pertence a j e 0 caso contrário, $Y_i|(Z_{ij}=1) = \boldsymbol{X_i^T \beta_j} + \epsilon_{ij}$, ou seja, isso é equivalente a termos um modelo de regressão diferente para cada grupo.

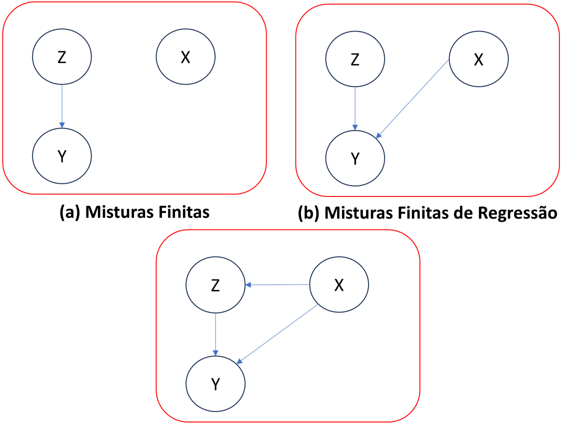
Misturas finitas de especialistas de regressão

As misturas finitas de especialistas de regressão são uma extensão onde consideramos que a probabilidade a priori depende das covariáveis (R) de cada observação. Com isso a função de densidade de probabilidade é dada por

$$f(y_i; \boldsymbol{\Theta}) = \sum_{j=1}^{G} \pi_j(\boldsymbol{r_i}; \Theta) \; g(y_i; \boldsymbol{x_i^T} \boldsymbol{\beta_j}, \Theta_j^*) \tag{2}$$

onde $\sum_{j=1}^{G} \pi_j(\boldsymbol{r_i}; \Theta) = 1$.

Misturas finitas de especialistas de regressão



(a) Misturas Finitas

(b) Misturas Finitas de Regressão

(c) Misturas Finitas de Especialistas de Regressão

Classificação

Uma questão fundamental a ser explorada é a classificação das observações. Aqui estamos assumindo a chamada classificação não supervisionada.

Conforme a regra de Bayes, tem-se que a probabilidade a posteriori de que a entidade pertença a j-ésima componente com y_i já observado, é

$$\hat{z_{ij}} = P(Z_{ij} = 1|yi) = \frac{p_j g_j(y_i)}{f(y_i)}, \tag{3}$$

onde a j-ésima proporção p_j da mistura, pode ser interpretada como a probabilidade a priori da i-ésima observação proveniente da j-ésima componente da mistura com $i = 1, ..., n$.

Classificação

No contexto de especialistas essa expressão será substituída por

$$\hat{z_{ij}} = P(Z_{ij} = 1|yi) = \frac{\pi_j(\boldsymbol{r_i}; \Theta) g_j(y_i)}{f(y_i)}, \tag{4}$$

- Cálculo de Probabilidades

Classificação

Nesse estudo foi considerado o modelo de regressão logística multinomial no ajuste das probabilidades π_j, onde

$$\pi_j(\boldsymbol{r_i};\Theta) = \pi_j(\boldsymbol{r_i};\alpha_j) = \frac{e^{\alpha_j^T \boldsymbol{r_i}}}{1+\sum_{l=1}^{G-1} e^{\alpha_l^T \boldsymbol{r_i}}}, \tag{5}$$

- Análise de dados categóricos
- MLG

Estimação

A partir da função de densidade de probabilidade do modelos, a fim de encontrar $\widehat{\beta}$, $\widehat{\alpha}$ e $\widehat{\Theta}^{*}$ vamos encontrar os estimadores de máxima verossimilhança. Dessa forma, todas as propriedades desses estimadores são válidas.

Na obtenção dos estimadores de máxima verossimilhança algoritmos do tipo EM foram obtidos via uma representação estocástica do modelo proposto. Toda essa estimação foi implementada ao longo do projeto no software R.

- Inferência Estatística Paramétrica

Algoritmo EM

Os algoritmos do tipo EM são algoritmos iterativos que tem como base a representação estocástica do modelo. Nesse sentido, precisamos estabelecer um critério de parada. Exitem algumas formas de definir o critério de parada, a mais comum utilizando a diferença relativa entre a log-verossimilhança entre o passo anterior e o passo atual

$$\frac{|\widehat{l}^{k+1} - \widehat{l}^{k}|}{|\widehat{l}^{k}|} \tag{6}$$

Nesse trabalho o critério de parada foi dado quando a diferença relativa fosse menor ou igual a 10^{-4}.

Algoritmo EM

Além disso, o algoritmo necessita de um chute inicial. Neste trabalho foi utilizado o algoritmo kmeans na clusterização dos grupos, e a regressão linear na estimação dos parâmetros de locação.

Seleção de modelos

Dado que, ao longo do projeto, foram considerados diferentes distribuições nas suposições do modelo, diferentes modelos podem ser estimados. Nesse sentido, a fim de escolher o melhor modelo de forma mais pacimoniosa possível, o critério de informação bayesiano (BIC) foi utilizado.

$$BIC = k\ ln(n) - ln(\hat{L}) \tag{7}$$

onde k é o numéro de parâmetros, n o tamanho amostral e $\hat{L}$ a verossimilhança.

Aplicação

Voltando ao exemplo apresentando anteriormente, os dados foram coletados por uma entidade federal para coleta e análise de dados relacionados a educação nos Estados Unidos da América. O conjunto é composto de 200 observações, foram aplicados os modelos de misturas de especialistas considerando a variável resposta *write* (y), a variável explicativa read (x_1) e as covariáveis de classificação *female* (r_1), *math* (r_2) e *science* (r_3).

Aplicação

Table 1: Estimativas considerando normalidade.

Estimativas	Grupo 1	Grupo 2
$\boldsymbol{\beta_1}$	44,76 (1,79)***	26,27 (2,21)***
$\boldsymbol{\beta_2(read)}$	0,26 (0,03)***	0,37(0,04)***
$\boldsymbol{\sigma^2}$	15,21(0,01)	35,64(0,07)
$\boldsymbol{\alpha_1}$	-12,89 (1,32)***	-
$\boldsymbol{\alpha_2(female)}$	2,42 (0,33)***	-
$\boldsymbol{\alpha_3(math)}$	0,13 (0,02)***	-
$\boldsymbol{\alpha_4(science)}$	0,10(0,02)***	-

0 *** 0,001 ** 0,01 * 0,05 . 0,1

Aplicação

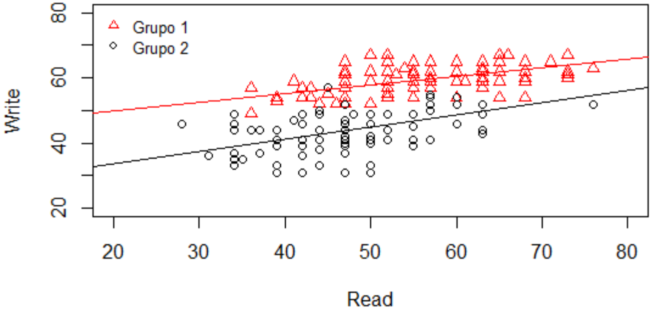


Figure 1: Ajuste modelo de misturas de especialistas de regressão Normal.

Aplicação

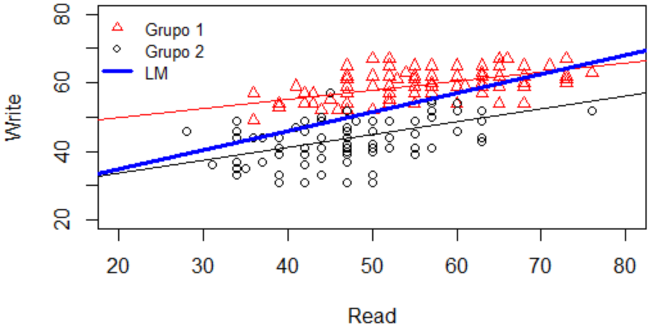


Figure 2: Ajuste modelo linear notas de redação.

Odds Ratio

Dado que utilizamos um modelo de regressão logística no ajuste das probabilidade π_j, e, que $G = 2$, nós conseguimos estimar a razão de chances dos grupos.

$$\text{Razão de Chances} = \frac{\text{Chance de pertencer ao grupo 1}}{\text{Chance de pertencer ao grupo 2}} \tag{8}$$

- Análise de dados categóricos

Aplicação

Table 2: Tabela de Razão de Chances.

Covariável	Odds Ratio
$\boldsymbol{female}$	11,25
$\boldsymbol{math}$	1,14
$\boldsymbol{science}$	1,11

0 *** 0,001 ** 0,01 * 0,05 . 0,1

Conclusão

A partir dos resultados obtidos no projeto concluímos que o modelo de misturas de regressão de especialistas é permite bastante flexibilidade, pois trata ao mesmo tempo da modelagem em problemas de regressão com multimodalidade, assimetria e/ou caudas pesadas.

Além disso, esses tipos de modelo apresentam um ganho na interpretabilidade de classificação dos grupos em relação aos modelos de misturas finitas usuais, dado que este utiliza as informações de cada observação no ajuste da classificação dos grupos.

Referências

Basso, Rodrigo Marreiro et al. 2010. "Misturas Finitas de Misturas de Escala Skew-Normal." PhD thesis, Dissertação de Mestrado-Universidade Estadual de Campinas, São Paulo.

Mirfarah, Elham, Mehrdad Naderi, and Ding-Geng Chen. 2021. "Mixture of Linear Experts Model for Censored Data: A Novel Approach with Scale-Mixture of Normal Distributions." *Computational Statistics & Data Analysis* 158: 107182.

Zeller, Camila B, Celso RB Cabral, and Víctor H Lachos. 2016. "Robust Mixture Regression Modeling Based on Scale Mixtures of Skew-Normal Distributions." *Test* 25: 375–96.