ME732 / ME921 / MI407: Análise Multivariada e Aprendizado Não-Supervisionado de Máquinas, atividade prática II

Guilherme Ludwig
gvludwig@ime.unicamp.br

25 de Maio de 2021

- A atividade é individual. A consulta de material eletrônico e livros será permitida, mas as respostas devem ser desenvolvidas inteiramente na solução entregue, e devem estar completas. Referencie o trabalho citado, mas desenvolva as soluções na atividade. Explique resultados com suas próprias palavras.
- A solução deve ser desenvolvida em detalhe. Todas as contas devem estar explicadas e desenvolvidas. Não é preciso fazer operações elementares manualmente mas as contas devem estar claramente indicadas.
- Cada aluno deverá ser responsável por garantir que não há cópia do seu trabalho. Indícios de plágio podem levar à anulação da atividade de todos os envolvidos. Use seu bom senso para guiar as conversas com seus colegas.
- A atividade deve ser entregue através do Moodle, em um único documento em formato `PDF`.
  - As páginas da sua solução devem estar numeradas, e as linhas também devem estar numeradas. Use o pacote no LaTeX `\usepackage{lineno}` e `\linenumbers`. Isso é importante pois me ajudará a dar *feedback* à sua solução.
  - Se você usar o `Word` ou `LibreOffice`, numere linhas e páginas e imprima o documento em formato `PDF` antes de enviá-lo.
- Use a linguagem que você preferir, de preferência com o pacote `listings` do LaTeX para exibir código: `R`, `C`, `python`, `julia`, ou outra qualquer. Todos os códigos devem estar listados no final da atividade, com instruções de execução (à parte das linguages listadas).
- Se você quiser, pode usar `Rmarkdown`, `Jupyter notebooks` ou o que preferir para mesclar código e texto.
- Organize os resultados mas não deixe de entregar soluções parciais, se você não conseguiu finalizar o exercício.
- **Prazo de entrega:** 01/06/2021, às 23:59, via Moodle. Vou tentar configurar o Moodle para permitir múltiplos envios, a última versão enviada é a final.

O conjunto de dados "`Beethoven_compositions.csv`" foi extraído na data de hoje do site `https://imslp.org/wiki/List_of_works_by_Ludwig_van_Beethoven` e corresponde a lista de composições publicadas por Ludwig van Beethoven (1770–1827). Essas composições não correspondem à totalidade da música feita por Beethoven, apenas a que ele publicou e recebeu um índice *Opus* (que é como um catálogo oficial de composições de autores clássicos). Eu removi as composições sem número de Opus, bem como algumas linhas redundantes. A lista de atributos é a seguinte:

```
> head(beethoven <- read.csv("Beethoven_compositions.csv", encoding = "latin1"), 10)

    Op.              Title        Key     Date     Scoring     Genre
1  1/1     Piano Trio No.1 Eb major  1792-93    vn vc pf   Chamber
2  1/2     Piano Trio No.2   G major  1792-94    vn vc pf   Chamber
3  1/3     Piano Trio No.3   C minor  1792-94    vn vc pf   Chamber
```

```
4  2/1  Piano Sonata No.1    F minor  1793-95          pf  Keyboard
5  2/2  Piano Sonata No.2    A major  1794-95          pf  Keyboard
6  2/3  Piano Sonata No.3    C major  1794-95          pf  Keyboard
7     3       String Trio  Eb major      1794   vn va vc   Chamber
8     4    String Quintet  Eb major      1795 2vn 2va vc   Chamber
9  5/1  Cello Sonata No.1    F major       1796      vc pf   Chamber
10 5/2  Cello Sonata No.2    G minor       1796      vc pf   Chamber
```

O *Opus* (indicado por `Op.`) do autor está organizado de forma crescente, mas pode conter mais de uma obra: por exemplo, o Opus 2 contém as Sonatas para Piano 1, 2 e 3. A coluna `Title` tem o título em inglês, comum, da peça. Tanto `Op.` quanto `Title` são identificadores dos indivíduos, e não devem ser usadas como atributo para clustering. O atributo `Key` indica o tom da música, `Date` é o ano de composição (mas há composições que demoraram mais de um ano, e outras particularidades). `Genre` é a classificação da música (como de câmara, orquestral, teclado, vocal etc.), mas os atributos mais importantes estão codificados na coluna `Scoring`. A coluna `Scoring` é composta pelos instrumentos envolvidos na composição, e tem uma legenda em `https://imslp.org/wiki/IMSLP:Abbreviations_for_Instruments`. Por exemplo, a primeira composição (Piano Trio No.1, Op.1) tem um `Scoring` dado por `vn vc pf`, ou seja, Violino, Violoncelo e Piano. Para simplificar a análise, eu transferi os nomes dos instrumentos em inglês para "`Beethoven_instrumentos.csv`":

```
> head(instruments <- read.csv("Beethoven_instruments.csv", encoding = "latin1"), 10)
```

```
   Abbreviation                    English
1           acc                  Accordion
2           afl                 Alto flute
3           alt   Alto (voice) (contralto)
4           arp                 Arpeggione
5           bag                    Bagpipe
6           bar           Baritone (voice)
7          bass               Bass (voice)
8          bbar      Bass baritone (voice)
9            bc  Continuo (Basso continuo)
10          bcl              Bass clarinet
```

Atividade:

- Reuna os dados dos `data.frames` `beethoven` e `instruments`. Minha sugestão é examinar quantros instrumentos únicos são utilizados na obra (i.e., codificá-los em 0 e 1), mas você também pode colocar a contagem do número de instrumentos (veja que no String Quintet Op. 4 temos 2 violinos e 2 violas).
- Justificando amplamente, decida os atributos mais interessantes (pode incluir `Key`, `Genre` e `Date`, se você organizar a última), e faça a análise de clusteres usando a distância de Gower e clustering tipo hierárquico (divisivo ou aglomerativo).
- Repita a análise do item anterior usando algum método baseado em modelos (por exemplo, do `Rmixmod`). Você pode selecionar atributos diferentes, se precisar (justifique!). Como os resultados se comparam?
- Como você determinou o número de clusteres?