Análise e sugestões de formalismo para artigo com pouco ou nenhum formalismo - FPCC 3

Bruna Stefany

Junho, 2024

1 Sugestões de Formalismo

1.1 Seção *Data Collection*

A seção de **Data Collection** é uma das seções mais defasadas com relação ao formalismo, mesmo sendo a que mais deveria apresentar o mesmo. Sendo assim, seguem as seguintes propostas de formalização.

1.1.1 Definição do período, conjunto de dados e tamanhos

Nesta subseção do artigo, deveriam ter sido definidos os períodos em dias e a definição do set de dados construído.

Original: "Data from the Top 50 Global lists—a.k.a. rankings—were collected using Spotify's Web API. The data were collected on a daily basis between November 2018 and April 2019."

Sugestão: "Data from the Top 50 Global lists—a.k.a. rankings—were collected using Spotify's Web API. The data were collected on a daily basis between November 2018 and April 2019 *(considering from 1st November to 30th April,* $daysPeriod = 180$*), creating a set composed by daily lists (each with size* $= 50$*) defined as* $DB = \{DL_1, DL_2, DL_3, ..., DL_{daysPeriod}\}$. Em inglês mês vem antes de dia. O correto é November 1st e April 30th.

Neste caso, foram adicionadas as definições necessárias, identificações dos tamanhos dos conjuntos e período.

1.1.2 Utilização do som como dado principal, definição do mesmo e suas informações

Cada linha dentro do set construído se refere a um som, portanto, o mesmo deveria ter sido definido, além de também indicar de forma mais direta a "aparência"de uma linha de informações relativa ao som.

Original: "For each daily list, we collected information from nine fields made available by the API. Namely, the entry's rank, the ranking date, the artists names, the song title, the song release date, its duration in milliseconds, and a URL for the song's 30-second sample. Additionally, each entry has an "explicit" flag, which indicates whether the song contains profanity, and a popularity score, which is a value in the [0, 100] interval that reflects how popular the song was on that day."

Sugestão: "For each *song* $= S \in DL_i$, we collected information from nine fields made available by the API:

$$SInfo = [rank,\ date,\ artists\_names,\ title,\ release\_date,\ duration,\ URL,\ explicit,\ popularity]$$

Definição de Sinfo OK. Namely, the entry's rank, the ranking date, the artists names, the song title, the song release date, its duration in milliseconds, a URL for the song's 30-second sample, *explicit indicating whether the song contains profanity and popularity* $\in [0, 100]$ that reflects how popular the song was on that day."

Sinto que faltou um pouco de explicação nas variáveis explicit e popularity. Sugestão: "Explicit indicating how much the song contains profanity and *explicit*) $\in [0, 100]$ where 0 indicates the song has no profanity and 100 the song contains profanity in every single line of the song; And *popularity*) $\in [0, 100]$ indicating how much the song is popular where 0 is the song is no played at all and 100 the song is played by all of our base."Acho que uma explicação mais completa poderia adicionar ao artigo original também.

Nesta sugestão, foi definido um símbolo para o dado sonoro e a indicação de que o mesmo pertence ao conjunto de uma lista diária. Além disso, foi definida a linha criada com as informações da API e uma explicação mais concisa sobre os campos *explicit* e *popularity*

1.1.3 Definição das features adicionais e adição das mesmas ao conjunto de informações de cada dado sonoro

Outro ponto que deveria ter sido definido era o conjunto de informações do som após a adição das *acoustic features* ao set.

Original: "[...] For the remaining songs, we used the Python package LibROSA to extract five acoustic features, namely Mel-Frequency Cepstral Coefficients (MFCC), spectral centroid, spectral flatness, zero crossings, and tempo. We chose this specific set of features because they appeared frequently during our literature overview."

Sugestão: "[...] For the remaining songs, we used the Python package LibROSA to extract five acoustic features:

$$F_{S_i} = [MFCC, centroid, flatness, zero\_cross, tempo]$$

Namely, Mel-Frequency Cepstral Coefficients (MFCC), spectral centroid, spectral flatness, zero crossings, and tempo. We chose this specific set of features because they appeared frequently during our literature overview. So, the final set of information for each sound was defined as follows:

OK, só adicionaria a definição destas variáveis para quem for leigo caso o artigo original não já tenha. Pessoalmente não sei o que é Mel-Frequency Cepstral Coefficients, spectral centroid, spectral Flatness nem zero crossing, consigo imaginar que são features de uma onda sonora na representação Mel Spectogram.

$$SInfo_i = SInfo_i + F_i$$

$SInfo_i$ possui duas definições. Em uma situação como essas acho que seria melhor sinalizar como outra variável como $SInfoEnriched_i$ para indicar que é uma versão enriquecida da variável da Sinfo.

$$SInfo_i = \begin{bmatrix} rank, & date, & artists\_names, & title, & release\_date, & duration, & URL, \\ explicit, & popularity, & MFCC, & centroid, & flatness, & zero\_cross, & tempo \end{bmatrix}$$ „

Nesta sugestão, foi definido o formato dos dados de uma *acoustic feature* de um som, a indicação de junção dos dois em um único set relativo ao som e a apresentação deste set após a junção.

1.2 Seção *Instance extraction*

1.2.1 Definir a representatividade dos símbolos desde o primeiro momento de utilização

Ao invés de apresentar δ e k como lag period e lagged features em **C. Training and Prediction**:

C. Training and Prediction - Original: "For our experiments, we defined the lag period to be δ=20 days and the number of lagged features to be $k = 3$."

Apresentar em **B. Instance extraction**, onde δ e k aparecem primeiro, desta forma já se é possível entender desde o início.

B. Instance extraction - Sugestão: "[...] The last feature added to each instance is the target value, which is a binary variable that indicates whether a song is popular $(k+1)\delta$ days after the date of the first instance, *where δ is the lag period and k is the number of lagged features.* For example, if we define $\delta = 30$ and $k = 4$, then for each instance the target will be its popularity status five months into the future. "

1.2.2 Atribuição da equação a uma função para reutilizar

A equação $(k + 1)\delta$ é utilizada em mais de uma seção diretamente. A mesma poderia ter sido atribuída a uma função para reutilizar.

Original: "[...] The last feature added to each instance is the target value, which is a binary variable that indicates whether a song is popular $(k + 1)\delta$ days after the date of the first instance. For example, if we define δ=30 and k=4, then for each instance the target will be its popularity status five months into the future."

Sugestão: "[...] The last feature added to each instance is the target value, which is a binary variable that indicates whether a song is popular $daysEstimated(k, \delta)$ after the date of the first instance. *Defining* $daysEstimated(k, \delta) = ((k + 1)\delta)$.

For example, if we define δ=30 and k=4, then for each instance the target will be its popularity status five months into the future. "

Sendo assim, na seção **C. Training and Prediction**, poderia ser reutilizada a função:

C. Training and Prediction - Sugestão: "For our experiments, we defined δ=20 k=3.[...] Each instance is lagged with popularity information from the next 60 days and is labeled according to the popularity of each song $daysEstimated(k, \delta) = 80$ days later[...]"

A definição a priori de (k+1)δ de fato deixa a interpretação mais clara e direta, mas acho que a definição da função algo desnecessário pois pode reduzir o entendimento do cálculo em si da função. Acho que seguir com (k+1)δ uma maneira de demonstrar diretamente o cálculo como melhor do que a opção de utilizar uma função.