

Descrição

É muito comum realizarmos comparação de algoritmos durante o desenvolvimento de soluções que utilizam Machine Learning. No caso de Aprendizado Supervisionado, existem muitas formas de avaliar os resultados, tais como Matriz de Confusão, Medida F1 e Curva ROC. No entanto, a avaliação de algoritmos de Aprendizado Não Supervisionado tem desafios adicionais, pois não temos um objetivo explícito sendo otimizado. Isso torna a avaliação mais complexa e um tanto quanto subjetiva, principalmente no caso de Agrupamento de Dados.

Neste trabalho, iremos exercitar esses conceitos, realizando tarefas muito similares às que Cientistas de Dados fazem no seu dia a dia. Além disso, vamos recapitular alguns conceitos importantes relacionados à Deep Learning. Este trabalho consiste de três partes principais:

1. Construção da base de dados
2. Aplicação de algoritmos de agrupamento de dados
3. Avaliação do resultado dos agrupamentos de dados

1 Construção da Base de Dados

Iremos realizar o agrupamento de dados em um conjunto de imagens. Para isso, é necessário realizar um processo intermediário para extração das *features* das imagens, com o objetivo de termos uma representação mais compacta (e “significativa”) dos dados antes de aplicarmos o agrupamento de dados.

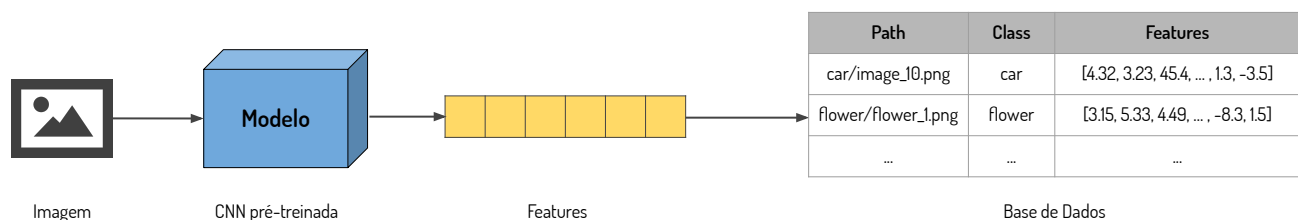


Figura 1: Fluxo de exemplo para construção da base de dados.

1.1 Requisitos

- a) Utilizar a base de dados *Natural Images 100*
 - https://drive.google.com/open?id=1PIqaoz15aqExRxJouFv7tRkwaaRms_jS
- b) Escolher uma rede neural convolucional pré-treinada disponível no Pytorch ou no Keras (e.g. ResNet18, ResNet50, Inception-v3, etc.). Escolher uma das últimas camadas da rede antes da camada de classificação para realizar a extração das features.

- [Exemplo de extração de features com Keras](#)
 - [Exemplo de extração de features com Pytorch](#)
- c) Criar rotina que compute as *features* para cada uma das imagens. A base de dados resultante pode ser de acordo com sua preferência. Alguns exemplos:
- *Forma 1*: Gerar um arquivo com o caminho/nome das imagens e outro arquivo com as features.
 - *Forma 2*: Gerar um arquivo único com o caminho/nome das imagens e com as features.
- d) **Atenção!** Como estamos utilizando modelos pré-treinados, preste muita atenção no pré-processamento das imagens. Os mesmos passos que foram feitos no treinamento do modelo devem ser utilizados ao aplicarmos o modelo em novos dados!

2 Aplicação de Algoritmos de Agrupamento de Dados

2.1 Requisitos

- Escolher três algoritmos diferentes de agrupamento de dados (utilizando as bibliotecas *scikit-learn* e *scipy*), para realizar agrupamento de diferentes imagens de uma base de dados.
- Idealmente (não é obrigatório) os algoritmos devem ser de tipos diferentes (por exemplo, um hierárquico, um particional e um por densidade). Podem ser utilizados algoritmos não vistos em aula, como Mean-Shift, Spectral Clustering, entre outros.

3 Avaliação do resultado dos agrupamentos de dados

3.1 Requisitos

- Comparação dos agrupamentos de dados
 - Quantitativa: Para comparação dos algoritmos, devem ser escolhidos **três** métodos de avaliação diferentes, sendo um *interno*, um *externo* e um *relativo*. O objetivo é avaliar a diferença dos índices para diferentes hiperparâmetros de um mesmo algoritmo e para comparar os resultados de diferentes algoritmos. Lembre-se das diferenças de cada um deles e quando eles se aplicam.
 - Subjetiva: Para a avaliação subjetiva, deve-se criar uma visualização das imagens dos grupos. Um exemplo pode ser visto na Figura 2.
- A partir dos experimentos realizados, deve-se escrever um relatório detalhando as escolhas feitas, os experimentos realizados e, principalmente, o aprendizado. Os diferentes hiperparâmetros utilizados (e.g. ϵ do DBSCAN) em diferentes experimentos não precisam ser todos colocados, porém o racional para ter chegado em tais valores deve ficar claro (o que implica em explicar por onde começou). É importante colocar os desafios e dificuldades enfrentados.
- Atenção!** Existe um número “correto” de grupos para essa base de dados, porém a análise não deve se concentrar apenas neste número. Use a criatividade e explore diferentes formas em que os dados podem ser agrupados, fugindo do que seriam os agrupamentos mais “naturais”!

Extras

Caso você se sinta encorajado a explorar um pouco mais, seguem três ideias de experimentos extras que você pode fazer para aprender e exercitar ainda mais seus conhecimentos. (Obs: ao realizar os experimentos extras, sua nota pode passar de dez 😊).

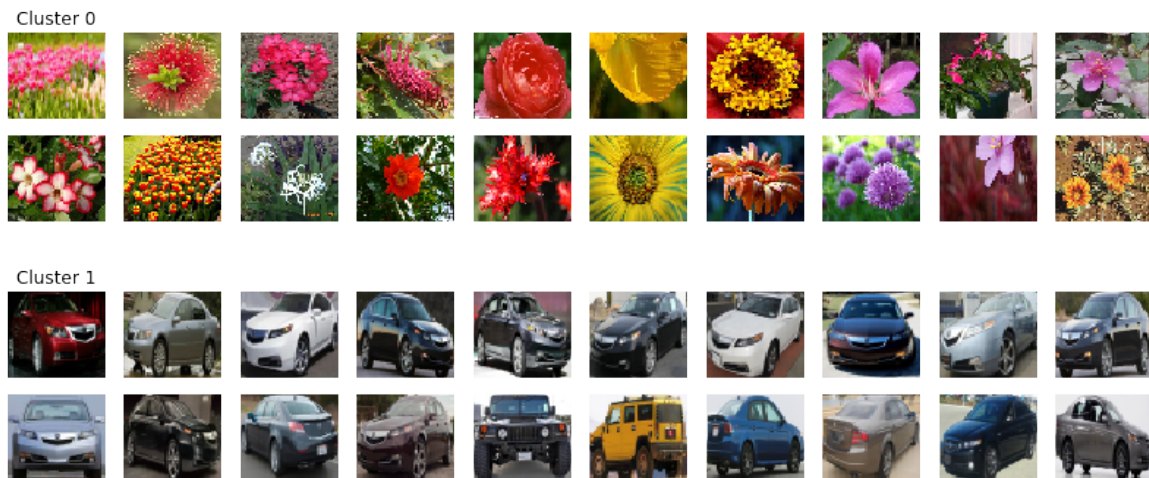


Figura 2: Exemplo de avaliação subjetiva.

- **Extra 1:** Escolha uma segunda rede neural convolucional pré-treinada e repita o processo de extração de features. Após, utilize os algoritmos de agrupamento com os melhores resultados e compare com os anteriores. O que você pode observar?
- **Extra 2:** Vamos aprender nas próximas aulas sobre redução de dimensionalidade. Faça experimentos com redução de dimensionalidade e compare com os agrupamentos anteriores.
- **Extra 3:** Escolha um conjunto de imagens de sua preferência (e.g. imagens de uma viagem que você fez) e realize o mesmo processo que você fez para os experimentos acima (extração de features e aplicação de algoritmos de agrupamento). Avalie os agrupamentos de forma subjetiva.

Entrega Final

- Jupyter notebook com os experimentos realizados (.ipynb e um .html ou .pdf).
 - Fique à vontade caso queira dividir as análises em mais de um notebook.
- Relatório em formato .pdf com os resultados e conclusões. Caso prefira, os resultados e conclusões podem ser colocados diretamente no Jupyter notebook, desde que fique bem explicado e claro.