Universidade Federal Fluminense
Instituto de Computação
Programa de Pós-Graduação em Computação
Aprendizado de Máquina
2019.1

Prof.: Flavia Bernardini (fcbernardini@ic.uff.br)
Estagiário: Eduardo Andrade (eandrade@ic.uff.br)

Especificação do Trabalho – Prático

Para a execução do trabalho prático, cada grupo deve ter no máximo 5 participantes cada. Espera-se que o trabalho seja o resultado da aplicação dos conceitos e práticas apresentados ao longo do período. O trabalho deve estar intimamente relacionado aos temas propostos, mas há diversas vias de exploração que podem ser consideradas, desde que os objetivos sejam cumpridos.

O trabalho deve ser completo. Isto significa que os grupos não podem ignorar questões relevantes de aprendizado de máquina, discutidas em sala. No relatório final, não deve estar escrito apenas o que foi feito, mas também deve ser apresentado o *rationale* utilizado para chegar à solução apresentada, à luz dos conceitos de aprendizado de máquina.

O relatório final deve ser construído em formato de artigo. Espera-se que o trabalho esteja contido em 8-10 páginas. Deve ser observado que a compreensão e a aplicação das questões envolvidas no uso de aprendizado de máquina são mais importantes do que obter resultados "excelentes" nos experimentos realizados.

Ao final do texto, é importante que cada grupo descreva o papel de cada pessoa no desenvolvimento do trabalho. Um dos pontos de avaliação será a execução dos projeto conforme o tamanho dos grupos, caso o número de pessoas não seja o mesmo para todos os grupos.

Todos os arquivos para a elaboração do trabalho podem ser encontrados neste link: https://www.dropbox.com/sh/e3pc5lsbydj1ut3/AADoBLs5C-Zscv_C0YgzN9yRa?dl=0. Os *datasets* e *templates* para o relatório final estão em suas versões originais. Será da responsabilidade de cada grupo alterá-los para o obtenção dos resultados esperados.

São 3 propostas distintas de trabalho prático para a pós-graduação e apenas 1 proposta de trabalho prático para a graduação. Haverá um sorteio para a pós-graduação no qual cada grupo ficará com apenas 1 das 3 propostas apresentadas neste documento. Não será possível alterar a proposta após o sorteio.

Alunos de Graduação

Análise de Sentimentos

O *Internet Movie Database* (IMDB) é a fonte mais popular e autoritária do mundo para conteúdo de filmes, programas de TV e celebridades. No IMDB há milhares de *ratings* e *reviews* feitos por usuários de diversos lugares do mundo. A língua mais utilizada é o inglês e o *dataset* para esta proposta apresenta somente informações neste idioma.

O *dataset* consiste de 50.000 *reviews* para filmes distintos. Nenhum filme apresenta mais de 30 *reviews*. Resumidamente, a análise de sentimentos consiste de técnicas que buscam extrair informações de textos em linguagem natural, com o objetivo de obter a polaridade destes textos ou de suas sentenças. Há apenas duas classes consideradas no *dataset*: *review* positivo ou *review* negativo.

Há três campos de dados nos arquivos presentes no *dataset*: id (um identificador para cada *review*), *sentiment* (1 para *reviews* positivos e 0 para *reviews* negativos) e *review* (o texto de cada *review*). Também há três arquivos distintos:

- labeledTrainData: é o *dataset* de treinamento rotulado. Há 25.000 linhas contendo os campos id, *sentiment* e o texto de cada *review*.
- testData: é o *dataset* de teste. Novamente são 25.000 linhas contendo os campos id, *sentiment* e o texto de cada *review*.
- unlabeledTrainData: é um outro *dataset* de treinamento sem rótulos. Há 50.000 linhas contendo os campos id e o texto de cada *review*.

Os objetivos que devem ser atingidos com o trabalho são:

- Realização da classificação binária com no mínimo 2 algoritmos apresentados em sala de aula (*naive bayes* e SVM, por exemplo)
- Realização de *clustering* com o *dataset* sem rótulos com no mínimo 1 algoritmo apresentado em sala de aula (KNN, por exemplo)
- Discussão dos resultados obtidos com as métricas utilizadas (*precision*, *recall*, *f-score*, por exemplo), comparando os algoritmos escolhidos

Alunos de Pós-Graduação

Proposta 1
Reconhecimento de Objetos

O *dataset* Caltech 256 contém 30.607 imagens de 256 categorias de objetos tomadas em variadas orientações, diferentes condições de iluminação e com origens distintas. A área de reconhecimento de objetos faz parte de um campo chamado computação visual e tem como objetivo, a procura pela identificação de objetos em vídeos ou imagens.

O *dataset* é autoexplicativo pois cada pasta contém as imagens correspondentes a uma determinada categoria. Os objetivos que devem ser atingidos com o trabalho são:

- Realização da classificação multiclasse com no mínimo 2 algoritmos apresentados em sala de aula (SVM, por exemplo), sendo no mínimo 1 rede neural entre esses 2 algoritmos (CNN, por exemplo). Não precisam ser consideradas todas as classes pois pode ser computacionalmente custoso. No mínimo considerar 10 classes distintas.
- Realização de *clustering* com no mínimo 2 algoritmos apresentados em sala de aula (KNN e DBSCAN, por exemplo)
- Discussão dos resultados obtidos com as métricas utilizadas (*precision*, *recall*, *f-score*, por exemplo), comparando os algoritmos escolhidos
- Apesar da breve descrição nesta proposta, ela não pode ser necessariamente considerada a mais fácil. O reconhecimento de objetos nem sempre é trivial, incluindo a extração/seleção de *features*

Proposta 2
Detecção de Anomalias e Regressão

Os esportes dos EUA são bastante conhecidos por terem um grande histórico de informações que são armazenadas para o uso de ciência dos dados. Com a NBA (*The National Basketball Association*) e a ABA (*The American Basketball Association*) isto não é diferente. O basquete é um dos esportes mais populares do país, sendo famoso pelos comentários e apostas que procuram acertar resultados de partidas, por exemplo. Todas as equipes também possuem especialistas que procuram nestas informações os pontos fracos de seus adversários.

O *dataset* possui os seguintes dados:

- *Player regular season stats* (estatísticas da temporada regular dos jogadores)
- *Player regular season career totals* (estatísticas totais de carreira da temporada regular dos jogadores)
- *Player playoff stats* (estatísticas da fase final dos jogadores)
- *Player playoff career totals* (estatísticas totais de carreira da fase final dos jogadores)
- *Player all-star game stats* (estatísticas do jogo das estrelas dos jogadores)
- *Team regular season stats* (estatísticas da temporada regular das equipes)
- *Complete draft history* (histórico completo do *draft*[1])
- *Coaches season* (recordes dos treinadores da NBA por temporada)
- *Coaches career* (recordes de carreira dos treinadores da NBA)
- *Teams* (lista de todos os times)
- *Team season* (estatísticas da temporada regular dos times)
- *Players* (lista de todos os jogadores)

Os objetivos que devem ser atingidos com o trabalho são:

- Detecção de anomalias como “quais jogadores foram muito superiores aos outros?”, por exemplo
- Prever o resultado dos jogos através de algoritmos de regressão, por exemplo
- Elaborar pelo menos mais 4 perguntas e predições no mínimo como o primeiro e segundo itens, respectivamente
- Discussão dos resultados obtidos, utilizando diferentes algoritmos apresentados em sala de aula
- Esta é a proposta com maior liberdade, porém é a proposta que mais demanda criatividade e capacidade de procurar informações úteis no dados

[1] https://pt.wikipedia.org/wiki/Draft_da_NBA

Proposta 3
Agrupamento e Regressão

A Netflix é uma empresa global de seriados e filmes via *streaming*. Atualmente o serviço de *streaming* oferecido pela Netflix conta com mais de 100 milhões de usuários. A Netflix é uma referência em sistemas de recomendação e de tempos em tempos, lança alguns desafios com prêmios valiosos para programadores do mundo todo que conseguem melhorar algum tipo de algoritmo utilizado pela empresa.

O *dataset* desta proposta de trabalho é o maior entre todas as propostas. Ele inclui 100 milhões de registros de avaliação na forma "usuário X avaliou o filme Y com um *rating* de 4.0 na data 12/02/05". Este *dataset* foi originalmente utilizado para um desafio da Netflix com premiação de 1 milhão de dólares. Os objetivos que devem ser atingidos com o trabalho são:

- O grupo pode prever a classificação que um usuário dará em um filme em relação aos filmes que o usuário avaliou no passado, bem como as avaliações que usuários semelhantes deram a filmes semelhantes?
- O grupo consegue descobrir *clusters* de filmes ou usuários semelhantes?
- Elaborar pelo menos mais 2 predições como o primeiro item
- Realização de *clustering* do segundo item com no mínimo 2 algoritmos apresentados em sala de aula (KNN e DBSCAN, por exemplo)
- Discussão dos resultados obtidos com as métricas utilizadas e comparação dos algoritmos escolhidos
- Esta é a proposta que ficaria no meio entre a proposta 1 e proposta 2. Também não é necessário utilizar todos os 100 milhões de registros no experimento mas no mínimo utilizar 1 milhão