Projeto 3

Ciência dos Dados 2018

Este documento apresenta as premissas do projeto final de Ciência dos Dados.

Objetivos

O principal objetivo do Projeto 3 é conduzir uma análise de dados com grau elevado de autonomia e liberdade de escolha de tema e de técnica.

Para que este fim possa ser alcançado, os estudantes deverão se aprofundar na técnica escolhida enquanto realizam o projeto.

É importante que o trabalho produza uma conclusão analítica e vá além da análise exploratória. Esta conclusão analítica deve ter a forma de classificação (supervisionada ou não supervisionada) ou regressão.

Grupos

O projeto pode ser realizado em grupos de no máximo 3 alunos.

Datas

Data	Entregável
30/10	Definição de grupo e até 3 propostas de tema (técnica e dataset) por grupo
1/11	Até fim do dia ter escolhido dataset e tema
5/11	Devolutiva dos professores
12/11	Check intermediário - dataset lido, exploratória e exemplo de aplicação da técnica
21/11	Análises concluídas - algoritmo gera alguma resposta
21/11	FIM: Relatório com explicação detalhada da aná;ise, conclusões e referências para fundamentação teórica

Sugestões de temas a utilizar

1. Regressão (linear ou logística)

Prever o valor de uma coluna de um dataset em função das outras. Pode ser uma regressão linear (se a variável de saída for quantitativa) ou regressão logística

(se a variável de saída for qualitativa)

Exemplos de datasets:

Predição de preços de casas em King County, Seattle

Predição de por quanto uma casa vai ser vendida

Predição de se funcionário vai deixar empresa ou não

Predição de qual *rating* alguém vai dar para um filme no Netflix

2. Classificadores - extensão do Naive Bayes

Baseado em todos os dados existentes, classificar em categorias

Exemplos de datasets:

Porto Seguro - cliente vai acionar o seguro?

Deteção de fraude no cartão de crédito

Deteção de fraude financeira

Predição de se funcionário vai deixar empresa ou não

Predição de sucesso de um filme

3. Clusterização

Agrupe os dados de um conjunto baseado em similaridade. Neste problema em geral pode-se escolher o número de *clusters* e o algoritmo precisa fazer o agrupamento.

Datasets interessantes para esta técnica

Pokémon

Fifa 18

Datasets interessantes

Ainda não há pergunta definida, mas são datasets interessantes

Futebol Europeu

Reviews de smartphones na Amazon

Filtro Anti Spam

Dataset da Enron. Mensagens classificadas em relação a assunto e sentimento

Predição se um produto entrou em falta - *backorder* - ou não

Lista de todos os datasets do Kaggle

Alguns datasets disponíveis publicamente

Rubricas

Veja a tabela com a rubrica geral do projeto. Postada no Blackboard e também no Github.

Dimensões de trabalho em equipe

1/3 do conceito do Projeto 3 vem de uma avaliação de trabalho em equipe.

Para ter a nota máxima, é preciso ter contribuições relevantes no Github do grupo **e** ter preenchido um formulário de avaliação dos colegas.

Nível	Descrição do nível
5	Produz mais trabalho ou trabalho de mais qualidade do que é esperado Faz contribuições importantes que mellhoram o trabalho do time Ajuda colegas que estão em dificuldade a completarem sua parte do trabalho
4	Demonstra um misto dos comportamentos acima e abaixo
3	Completa uma parte justa do trabalho com qualidade aceitável Respeita compromissos e completa tarefas a tempo Ajuda colegas que estão em dificuldade se a tarefa for fácil ou muito importante
2	Demonstra um misto dos comportamentos acima e abaixo
1	Não faz trabalho na proporção justa esperada Entrega trabalho de qualquer jeito ou incompleto Perde prazos Atrasa, falta ou chega despreparado para reuniões e trabalho Não ajuda os colegas nem os mantém informado sobre o que está fazendo Abandona tarefas difícies

Figure 1: Contribuir com o trabalho do time

Atenção:
A nota de trabalho em equipe nunca aumenta a nota geral do projeto.

Em outras palavras, não adianta ter A em trabalho em equipe e D em projeto. A nota final ainda será D.

Referências

Além dos materiais da disciplina e dos livros-texto, sugerimos as seguintes obras para uma visão geral de Machine Learning / Classificação em Python.

Nível	Descrição do nível
5	Monitora questões que afetam o time e acompanha a evolução do trabalho e das pessoas Assegura-se de que os companheiros de grupo estão progredindo em suas tarefas Dá aos colegas feedback específico, construtivo e na hora certa (não quando é tarde demais)
4	Demonstra comportamentos do 5 e do 3
3	Percebe mudanças que afetam o sucesso do time Sabe o que todos da equipe deveriam estar fazendo e percebe problemas Alerta os colegas ou sugere soluções quando o sucesso estiver ameaçado
2	Demonstra comportamentos acima e abaixo
1	Não sabe se o time está atingindo as metas Ignora a evolução do trabalho dos colegas Evita discutir problemas do time, mesmo quando são óbvios

Figure 2: Manter o time no rumo certo

Nível	Descrição do nível
5	Solicita e demonstra interesse nas idéias e contribuições dos colegas de equipe Certifica-se de que os colegas de time estejam informados e se entendam mutuamente Encoraja e deixa o time entusiasmado Pede feedback aos colegas de equipe e use suas sugestões para se aperfeiçoar
4	Demonstra comportamentos descritos acima e abaixo
3	Respeita o feedback dos colegas de grupo e responde a ele Participa integralmente das atividades da equipe Comunica-se com clareza. Compartilha informação com colegas de equipe Ouve aos colegas de time e respeita suas contribuições
2	Demonstra comportamentos descritos acima e abaixo
1	Interrompe, ignora, oprime, ou faz chacotas com os colegas de grupo Toma ações que impactam os colegas sem perguntar a eles antes. Não compartilha informação Reclama, arranja desculpas e não interage com os colegas de grupo É defensivo. Não aceita ajuda ou sugestões de colegas

Figure 3: Interação com os colegas de time

Hands-on Machine Learning

[Python Data Science Handbook - Capítulo 5] (https://jakevdp.github.io/PythonDataScienceHandbook/)

Python Machine Learning

Dica: Encontre um *dataset* primeiro, depois formule uma pergunta, e daí busque uma técnica condizente.