Projeto 3 - Ciência dos Dados

Ciência dos Dados 2019

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

Baixe o PDF para os links funcionarem

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.

As técnicas a ser utilizadas serão: regressão linear, regression tree, random forest regression, *multinomial naive bayes*, regressão logística, *decision tree*, *random forest* e clusterização (k-means).

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 4 alunos.

Datas

Data	Entregável
31/10	Kickoff do projeto
5/11	Definição de grupo e 2 propostas de tema (técnica e dataset) por grupo. Deixe claro em cada proposta qual o tipo de variável predita
12/11	Datasets lidos e análise exploratória concluída (entrega via Blackboard)
14/11	Algortimo gera alguma resposta (check em aula)
19/11	Entrega dos resultados (entrega via blackboard)
21/11	Entrega do relatório com explicação detalhada da análise, conclusões e referências para fundamentação teórica

Sugestões de temas a utilizar

Nota: há muito mais datasets listados neste link, os exemplos abaixo servem só para dar uma ideia .

1. Regressão

As técnicas que se prestam a este tipo de análise: regressão linear, *regression tree*, *random forest regression*

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 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. Técnicas que fazem classificação: *multinomial naive bayes*, regressão logística, *decision tree* e *random forest*.

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 - não supervisionado

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. A técnica que implementa são os k-means

Datasets interessantes para esta técnica

Pokémon

Fifa 18

Datasets interessantes

Ainda não há pergunta definida, mas são datasets interessantes

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

Vão ser cobradas e anotadas a cada aula: * Contribuição de cada membro do grupo * Participação ativa em aula de cada membro do grupo * Frequência dos membros do grupo

A falta de atender aos critérios acima implica em descontos na nota do projeto para todos os membros.

Além disso, haverá uma dimensão de trabalho em equipe avaliada por questionário e que modula a nota final.

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.

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

Referências

Aurelien Géron. Hands-on Machine Learning with Scikit-Learn and Tensorflow, O'Reilly, 2017. Capítulos 1 e 2. **Disponível na biblioteca**

DANTAS, D. Comparação Entre Técnicas de Regressão Logística, Árvore de Decisão, Bagging e Random Forest Aplicadas a um Estudo de Concessão de Crédito - Trabalho de Conclusão de Curso. UFPR, Curitiba, 2013 - Capítulo 2

Introduction to Statistical Learning - capítulos 4 e 10

No livro acima, leitura recomendada: * Capítulo 1 * Capítulo 2, seções 2.1 e 2.2 * Capítulo 3, seções 3.1 3 3.2 * Capítulo 4, seções 4.1, 4.2 e 4.3 * Capítulo 8, seções 8.1 e 8.2 * Capítulo 10, seção 10.3.

Existe uma tradução completa do código do livro acima para linguagem Python

Hands-on Machine Learning - notebooks Python. Temos o livro na biblioteca

Python Data Science Handbook - Capítulo 5

Python Machine Learning

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