MINI PROJETO 1

MODELOS PROBABILÍSTICOS E DADOS

IDENTIFICAÇÃO DE DISTRIBUIÇÕES

OBJETIVO:

O objetivo deste projeto é identificar quais distribuições (funções de densidade de probabilidade - no caso contínuo, ou funções de probabilidade - no caso discreto) descrevem bem variáveis quantitativas extraídas de um *datasets*.

O resultado final esperado é um relatório que identifique, com bons argumentos, a escolha de um ou mais modelos probabilísticos para ajuste de uma variável quantitativa extraída de *dataset*.

Este projeto é **estritamente individual**.

O QUE DEVE SER FEITO:

Você precisa escolher uma variável quantitativa em *datasets* públicos de sua escolha. A sua variável pode ser discreta ou contínua.

Limpe e prepare os dados para processamento (tratando valores `NaN` ou `N/A`, por exemplo). Fique atento ao dicionário de dados (se houver) para identificar quais colunas do *dataset* de fato são quantitativas e eventualmente remover valores inválidos.

A seguir, estude a variável escolhida e procure identificar uma função adequada que descreva as probabilidades de ocorrência dos valores que essa variável pode assumir.

Sugerimos fortemente que o trabalho siga as seguintes fases:

1. Seleção de um *dataset* e escolha uma variável quantitativa adequada.
    - Não há restrições em relação à base de dados a utilizar, ***desde que não seja as mesmas bases da PNAD já usadas na disciplina***. Aconselha-se evitar variáveis de bases com pequeno tamanho amostral.
    - Tornamos disponível uma **Lista de datasets** que pode ajudar nesta fase do trabalho. **Atenção:** nem todas as bases de dados desta lista têm variáveis quantitativas, analise com cuidado. Você não precisa ficar restrito a esta lista
    - Indique o *dataset* e a variável que escolheu no **piazza. IMPORTANTE!!!**

2. Limpeza da variável escolhida, se necessário.
3. Inspeção visual da distribuição dos valores da variável escolhida - usando um histograma, por exemplo.
4. Formulação de hipóteses sobre o formato da distribuição dos dados (simetria, assimetria positiva e assimetria negativa) e escolha **PELO MENOS DUAS DISTRIBUIÇÕES TEÓRICAS DIFERENTES PARA MODELAR SUA ÚNICA VARIÁVEL QUANTITATIVA** definida no item 1. Justifique por que escolheu suas distribuições teóricas.
5. Tentativa de estimar os parâmetros da família de distribuições escolhida no item acima a partir dos dados.
6. As distribuições do pacote `scipy.stats` têm uma função chamada `fit()` que procura estimar os parâmetros a partir do conjunto de dados.
    - Use o `fit()` para fazer estimativa dos parâmetros da família de distribuições escolhida no item 4.
    - Compare os parâmetros estimados a partir do `fit()` com os parâmetros estimados por você no item 5. Para cada uma das suas distribuições teóricas, opte por um ajuste: o do item 5 ou o obtido pelo comando `fit()`.
7. Construa o histograma dos dados junto com a fdp de cada distribuição teórica e analise.
8. Construa o QQ-Plot (quantil amostral vs quantil teórico) e analise. **Dica:** veja Exemplo 6.8 do Magalhães e Lima (7ª. edição) de como obter as frequências relativas acumuladas a partir de uma amostra de tamanho n e de como obter os quantis teóricos.
9. Construa um gráfico com a frequência relativa acumulada (a partir dos dados) vs a função de distribuição acumulada e analise.
10. Faça um teste de aderência para a distribuição (veja o arquivo `MiniProjeto1 Aderencia Numpy Pseudocodigo.ipynb`). Teste de aderência é útil para mensurar a qualidade do ajuste do modelo teórico aos dados.
11. Elabore uma tabela que contrasta sua variável com as distribuições teóricas escolhidas e a qualidade do ajuste em cada caso. Analise essa tabela e faça a escolha da melhor das distribuições teóricas para o ajuste dos dados.

ENTREGÁVEIS ESPERADOS E DATAS:

Turmas A, B e C:

Item	**Data**	**Descrição**
Indicação de dataset	19/09/2016	Indicar *dataset* e variável de interesse em post no **Piazza** de Ciência dos dados
Entrega intermediária (check)	20/09/2016	Histogramas das variáveis candidatas e possíveis distribuições adequadas (Itens 1 a 4 completos).
Relatório final	23/09/2016	Relatório enviado na pasta MiniProjeto1 no Github.

FÓRUM DE DISCUSSÃO:

Um fórum de discussão foi criado no **Piazza** - procure participar para tirar suas dúvidas e ajudar seus colegas: (https://piazza.com/insper.edu.br/fall2016/cd2016_2)

Não aceitaremos mesma variável quantitativa analisada por dois alunos da mesma sala ou de turmas diferentes. Assim, aproveite o fórum para descrever as variáveis quantitativas que irá trabalhar. Uma vez publicadas, um outro aluno não poderá mais utilizá-las neste projeto. Esse fórum será único para as três turmas, fazendo com que isso seja válido para todas as turmas.

RUBRICS DE AVALIAÇÃO DO OBJETIVO DE APRENDIZADO

Objetivo de aprendizado	Insatisfatório (I)	Em desenvolvimento (D)	Essencial (C)	Proficiente (B)	Avançado (A)
Especificar as distribuições de probabilidades adequadas para as variáveis	Apresentou entregas insuficientes ou atrasadas	Conseguiu fazer a leitura dos dados mas não avançou na análise Escolheu um conjunto de dados que já tinha sido escolhido pelo colega Não indicou adequadamente a URL do dataset escolhido ou os nomes específicos das variáveis	Para a variável quantitativa escolhida: - Leu os dados adequadamente - Traçou um histograma considerando densidade no eixo y (normed=True). - Elegeu pelo menos duas distribuições teóricas e justificou escolhas. - Estimou os parâmetros das distribuições teóricas a partir dos dados ou utilizou comando fit(), mas justificou no item 6 escolha final para estimativas. (Até item 6!)	Realizou os comportamentos de C de maneira excelente e: - - Traçou adequadamente as fdp's ou fp's junto aos histogramas. - Construiu e avaliou adequadamente o ajuste dos dados usando QQ-plot. - Construiu e avaliou adequadamente função de distribuição acumulada e frequência relativa acumulada. - Analisou gráficos impecavelmente. (Até item 9!)	Realizou os comportamentos de B e C de maneira excelente e: - Avaliou entre pelo menos duas distribuições alternativas usando um teste de aderência e formulou uma conclusão coerente em relação a escolha da distribuição que gera o melhor ajuste.