FUNDAÇÃO EDSON QUEIROZ
UNIVERSIDADE DE FORTALEZA
ENSINANDO E APRENDENDO

Pré-Modelagem em Ciência de Dados

Prof. Rilder S. Pires

MBA em Ciência de Dados

Pré-Modelagem em Ciência de Dados

Encontros:

- Módulo 1: 09, 10 e 11 de dezembro de 2021
- Módulo 2: 13, 14 e 15 de janeiro de 2022
- Módulo 3: 27, 28 e 29 de janeiro de 2022

Projeto Final:

- Análise de Dados Sócio-Econômicos das Mesoregiões Cearenses

Pergunta Norteadora:

- Quão diferente são as Mesoregiões Cearenses?

Observações:

- Dados da Plataforma SIDRA-IBGE
- Produção Agrícola Municipal (https://sidra.ibge.gov.br/tabela/5457)
- Produto Interno Bruto dos Municípios (https://sidra.ibge.gov.br/tabela/5938)
- Estimativas de População: (https://sidra.ibge.gov.br/tabela/6579)
- Entregar os **notebooks com códigos e explicações**.

Pré-Modelagem em Ciência de Dados

Ementa:

- Conceitos de Axiomas da Probabilidade
- Atribuições das Probabilidades
- O que é uma variável aleatória?
- Distribuição de Probabilidade Discretas:
  - Distribuição de Bernoulli,
  - Distribuição Binomial,
  - Distribuição de Poisson,
  - Distribuição Geométrica e Hipergeométrica
- Distribuições Contínuas:
  - Distribuição Uniforme,
  - Distribuição Exponencial,
  - Distribuição Normal ou Gaussiana,
  - Cálculo de Probabilidade em Distribuições Normais e Funções lineares de Distribuições Normais.
- Inferência Estatística: Noções de amostragem e estimação.

Na aula passada...

Parte Teórica:

- Inferência Estatística
  - Conceitos Fundamentais
    - Conjuntos de Confiança
    - Teste de Hipótese

Parte Prática:

- Exemplos
- Exploração dos Dados

Correlação

Covariância:

- Seja X e Y variáveis aleatórias com médias μ_X e μ_Y e desvios padrões σ_X e σ_Y, definimos a covariância entre X e Y, por

$$\operatorname{cov}(X, Y) = \mathbb{E}\left((X - \mu_X)(Y - \mu_Y)\right)$$

Correlação:

- O **coeficiente de correlação de Pearson** é definido como

$$\rho = \rho_{X,Y} = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y}$$

- o coeficiente de correlação de Pearson mede o quão forte é relação linear entre X e Y.
- **Cuidado!** Correlação não significa causalidade.

Correlação de Ranking:

- A **coeficiente de correlação de Spearman** é definido como

$$\rho = \rho_{R(X),R(Y)} = \frac{\operatorname{cov}(R(X), R(Y))}{\sigma_{R(X)} \sigma_{R(Y)}}$$

- a correlação de Spearman mede o quanto a relação entre X e Y pode ser descrito utilizando uma **função monotônica**.

Máxima Likelihood (verossimilhança)

Definição:

- É um método para estimar parâmetros de um modelo paramétrico.
- A **função likelihood** é definida por

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i, \theta)$$

- A **função log-likelihood** é definida como $\ell_n(\theta) = \log \mathcal{L}_n(\theta)$.
- A função likelihood é apenas a densidade conjunta dos dados, exceto que a tratamos como uma função do parâmetro θ.

Estimador:

- O **estimador de máxima verossimilhança** (maximum likelihood estimator MLE) denotado por $\hat{\theta}_n$, é o valor de θ que maximiza $\mathcal{L}_n(\theta)$.

Projeto Final:

Projeto Final

Projeto Final:

Perguntas

Projeto Final

Projeto Final:

Perguntas

1. Qual a distribuição da “diversidade” dos municípios da sua região?

Projeto Final

Projeto Final:

Perguntas

1. Qual a distribuição da "diversidade" dos municípios da sua região?

2. Qual a distribuição dos valores de produção agrícola dos municípios da sua região?

Projeto Final

Projeto Final:

Perguntas

1. Qual a distribuição da "diversidade" dos municípios da sua região?

2. Qual a distribuição dos valores de produção agrícola dos municípios da sua região?

3. Qual a distribuição dos valores de produção do principal produto para municípios da sua região?

Projeto Final

Projeto Final:

Perguntas

1. Qual a distribuição da "diversidade" dos municípios da sua região?

2. Qual a distribuição dos valores de produção agrícola dos municípios da sua região?

3. Qual a distribuição dos valores de produção do principal produto para municípios da sua região?

4. e para o Ceará?

Projeto Final

Projeto Final:

Perguntas

1. Qual a distribuição da "diversidade" dos municípios da sua região?

2. Qual a distribuição dos valores de produção agrícola dos municípios da sua região?

3. Qual a distribuição dos valores de produção do principal produto para municípios da sua região?

4. e para o Ceará?

5. Quais outras variáveis podemos considerar?

Fim

Obrigado pela atenção!