Dados Categóricos

1. Introdução

Os dados categóricos são os tipos de variáveis definidos anteriormente como variáveis **qualitativas** e que poderiam ser separados em dois principais tipos, **nominais** ou **ordinais** como definidos abaixo:

- **qualitativa nominal**: as variáveis do tipo qualitativas não apresentam valores mensuráveis. No caso das variáveis **qualitativas** e **nominais**, as variáveis **não apresentam uma ordenação ou hierarquia** entre as categorias. **Exemplo:** Sexo, País, estado civil etc.
- **qualitativa ordinal**: Já para as variáveis **qualitativas** e **ordinais**, as variáveis **apresentam uma ordenação ou hierarquia** entre as categorias. **Exemplo:** escolaridade, faixa salarial, período do dia etc.

De forma análoga ao processo feito com as variáveis **quantitativas** (no caso numéricas), existem testes de hipóteses que são capazes de dizer a qualidade de uma determinada variável categórica em relação ao poder de separação em um modelo. Os testes que são aplicados em variáveis categóricas são os chamados de **testes não-paramétricos**.

Nos tópicos a seguir, será discutido a respeito de dois testes muito utilizados no processo de **seleção de atributos**, estes testes são o **Qui-quadrado** e **ANOVA**

2. Teste Qui-Quadrado

O Teste **Qui-quadrado** (*Chi-Squared* em inglês) é um teste não-paramétrico que mede a relação de dependência entre duas variáveis categóricas, verificando se os valores esperados estão muito distantes dos valores observados para estas métricas. Ou seja, dados o vetor de contagens observadas $Oij=(O_{11},O_{12}, ..., O_{rc})$, r E_{ij} representa os valores

esperadas e admitindo válida a hipótese de independência dos critérios de classificação, a estatística para o teste de Qui-quadrado pode ser definida a seguir:

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Para valores altos da estatística do Qui_quadrado (respectivamente um valor baixo para o *p-value*), significa uma evidência forte que os valores observados e esperados são diferentes, portanto, possuem dependência entre si. Esse grau de dependência entre as variáveis está relacionado diretamente com o valor do Qui_quadrado, quanto maior o valor da estatística, maior a dependência.

A implementação em *Python* para o teste Qui-quadrado pode ser feito conforme o código a seguir:

```
# Carrega a função do Qui Quadrado
from sklearn.feature_selection import chi2

# Aplicando o teste Qui quadrado
chi_scores = chi2(X, # Todas as variáveis categóricas
                  y) # Variável resposta
```

Dessa forma, é possível determinar dentre todas as variáveis categóricas em um conjunto de dados, quais que têm forte dependência com a variável resposta, em modelos de classificação no caso. Vale ressaltar que as variáveis categóricas devem ser convertidas para uma identificação numérica antes de aplicar o teste, como por exemplo utilizar o `LabelEncoder`

3. ANOVA

O teste **ANOVA** (*Analysis of Variance* em inglês) é um teste não-paramétrico para verificar se existem diferenças significativas entre as médias de grupos de dados, sendo possível inferir se as variáveis são dependentes uma sobre a outra. Para isso, calcula-se a relação entre a variância entre grupos S_B^2 com a variância dentro dos grupos S_W^2, conforme a fórmula a seguir:

$$F = \frac{S_B^2}{S_W^2}$$

A partir da estatística do valor F, quanto maior for a variância entre os grupos, mais diferentes as duas variáveis serão entre si. Dessa forma, de acordo com o valor F, pode se inferir a respeito das variáveis serem diferentes e exercerem influência entre si.

A implementação em *Python* para o teste ANOVA pode ser feito conforme o código a seguir:

```
# Carrega a função ANOVA
from sklearn.feature_selection import f_classif

# Aplicando o teste Qui quadrado
F_scores = f_classif(X, # Todas as variáveis categóricas
                     y) # Variável resposta
```

De forma análoga ao Qui-quadrado, com este teste é possível determinar dentre todas as variáveis categóricas em um conjunto de dados, quais que têm forte dependência com a variável resposta, em modelos de classificação no caso. Vale ressaltar novamente que as variáveis categóricas devem ser convertidas para uma identificação numérica antes de aplicar o teste, como por exemplo utilizar o `LabelEncoder`

Materiais Complementares

Documentação no Scikit-Learn sobre o chi2;

Documentação no Scikit-Learn sobre o ANOVA;

Artigo publicado por Gustavo Santos no *Data Hackers* - *Estatística para Seleção de Atributos*.

Referências

James, Gareth, et al. An Introduction to Statistical Learning: With Applications in R. Alemanha, Springer New York, 2013;

Bruce A., Bruce P. Estatística Prática para Cientistas de Dados. Segunda Edição, Alta books, 2019;

Tópico anterior