Apresentação Projeto Ciência de dados

Alberto Rodrigues Ferreira
Orientador: Rafael Braz Farias
Co-Orientador: Juvêncio Santos Nobre

Universidade Federal do Ceará
Departamento de Estatística e Matemática Aplicada

1 de Março de 2019

Tipos De Dados Faltantes

- Dado faltante completamente aleatório: São os dados que são faltantes complemente ao acaso.
- Dado faltante aleatório: É quando a probabilidade dos dados serem faltantes só é explicada pelos dados não-faltantes.
- Dado faltante não-aleatório: É quando o dado faltante ocorre somente pela por motivo da própria variável.

Estimação de Dados Faltantes Para Árvore de Regressão

- Proporção de dados faltantes:
  Proporção $\leq$ 0,05 $\rightarrow$ Pode ser usada imputação única ou trabalhar com o banco de dados completo.
  Proporção entre 0,05 e 0,15 $\rightarrow$ Imputação única pode ser usada provavelmente sem problemas, mas o uso de imputação múltipla é indicado.
  Proporção $>$ 0,15 $\rightarrow$ Imputação múltipla é indicado na maior parte dos casos.

Método de retirada dos valores faltantes

- É quando simplesmente tira-se a observação que tenha pelo menos um valor faltante.

Método do valor de tendência central

- Se a distribuição for aproximadamente normal, onde as observações estão bem agrupadas em torno da média, a média deve ser selecionada.
- Para distribuições assimétricas, com muitos outliers, a mediada é uma medida mais robusta para ser imputada.
- No caso da variável ser nominal, a moda é selecionada.

Método de Regressão

- Explorando a correlação entre as variáveis, utilizando regressão linear simples ou múltipla com os dados completos para imputar com a variável resposta como sendo a variável com dados faltantes.

Método dos vizinhos mais próximos

- Verifica-se para cada observação com dado faltante quais os vizinho mais próximos, e é mensurado com a distância dada a seguir:

$$\text{d}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{p-a} \delta_i(x_i, y_i)}$$

$$\delta_i = \begin{cases} 1 & \text{se i é categorica e } x_i \neq y_i \\ 0 & \text{se i é categorica e } x_i = y_i \\ (x_i - y_i)^2 & \text{se i é numérica} \end{cases}$$

- Em seguida, pode ser usado dos k vizinhos mais próximos a mediana da variável com dado faltante ou uma média ponderada pela função kernel dada por e^{-d}.

Método da imputação múltipla

- São obtidos m banco de dados imputados por diferentes técnicas.
- Isoladamente são analisados por algum método estatístico.
- Os m resultados são combinados de modo simples para inferência.

Método Predictive Mean Matching

- Para cada possibilidade de dados faltantes é realizado um modelo de regressão linear com as outras variáveis preditoras e calculamos $\hat{\beta}$.
- Um novo $\hat{\beta}^*$ é obtido através de um sorteio aleatório de uma $N_p(\hat{\beta}, \Sigma)$, em que Σ é a matriz de covariâncias de $\hat{\beta}$.
- Agora são realizados todos os valores preditos do banco de dados inteiro, incluindo a resposta como faltante e não-faltante.
- Para cada valor predito cuja resposta é um dado faltante observamos os valores preditos de todas as outras respostas aonde não é dado faltante, e pegamos os k valores preditos mais próximos e sorteamos aleatoriamente e esse valor predito será imputado no dado faltante.

Exemplo 1

- **Descrição do Problema e Objetivos**:
  Altas concentrações de certas algas nocivas nos rios constituem um sério problema ecológico com um forte impacto não só nas formas de vida dos rios, mas também na qualidade da água. Ser capaz de monitorar e realizar uma previsão antecipada de algas nocivas é essencial para melhorar a qualidade dos rios.
  Com o objetivo de resolver este problema de previsão, várias amostras de água foram coletados em diferentes rios europeus em diferentes momentos durante um período de aproximadamente 1 ano. Para cada amostra de água, diferentes propriedades químicas foram medidos, bem como a frequência de ocorrência de sete algas nocivas.

Exemplo 1

- Algumas outras características do processo de coleta de água também foram armazenadas, como a estação do ano, o tamanho do rio e a velocidade do rio. Uma das principais motivações por trás desta aplicação reside no fato de que monitoramento químico é barato e facilmente automatizado, enquanto a análise biológica das amostras para identificar as algas que estão presentes na água envolve exame microscópico, requer mão de obra treinada e, portanto, é caro e lento. Como tal, obter modelos que sejam capazes de prever as frequências de algas com base nas propriedades químicas facilitaria a criação de sistemas baratos e automatizados para monitoramento de algas nocivas.

Exemplo 1

- Cada observação contém informações sobre 11 variáveis. Três dessas variáveis são nominais e descrevem a estação do ano em que as amostras de água a serem agregados foram coletados, bem como o tamanho e a velocidade do rio em questão. As oito variáveis restantes são valores de diferentes parâmetros medidos nas amostras de água que formam a agregação, nomeadamente:
- Valor pH máximo, valor mínimo de O_2, valor médio de Ch, valor médio de NO_3^-, valor médio de NH_4^+, média de PO_4^{3-}, média do total de PO_4 e média de clorofila (Chla).

Referências

- Assunção, F. (2011). Estratégias para tratamento de variáveis com dados faltantes durante o desenvolvimento de modelos preditivos.
- Ferreira, M F M. (1999). Árvores de regressão e generalizações -Aplicações-.
- T, Luis. Árvores de Regressão Métodos e Aplicações. Disponível em: http://www.dcc.fc.up.pt/ ltorgo/Presentations/MatAplicNov98/inde Acesso em 29 de Outubro de 2018.
- A, Paul. Imputation by Predictive Mean Matching: Promise Peril. Disponível em : https://statisticalhorizons.com/predictive-mean-matching. Acesso em 10 de Fevereiro de 2019.
- NUNES, Luciana N. Métodos de imputação de dados aplicados na área da saúde. 2007. 120 f. Diss. Tese (Doutorado em Epidemiologia)–Universidade Federal do Rio Grande do Sul, Porto Alegre, 2007.