Machine Learning
Preparação de dados: ruido

Prof. Hugo de Paula

Preparação dos dados: dados ruidosos

Podem ser inconsistentes, inexatos ou incompletos.

- Aplicar técnicas estatísticas para identificar e separar os diferentes tipos de ruídos.
- Ruído de atributos (ruído de medição ou aleatório)
- Ruído de classe

Preparação dos dados: omissão e ruído

Importante identificar como a precisão dos dados está sendo afetada pela omissão e ruído.

Exemplos:

- João e Joao podem ser consideradas entradas iguais ou diferentes.

Preparação dos dados: omissão e ruído

Exemplos:

- Diferentes entidades representadas pelo mesmo nome em diferentes sistemas:
  - Sistema folha de pagamento: Empregado é quem recebe salário (pode ser terceirizado).
  - Sistema de RH: empregado é quem possui número de pessoa.

Preparação dos dados: codificação

Importante garantir padronização e aplicar técnicas estatísticas para analisar consistência dos modelos encontrados.

- Nem todos usam o mesmo formato.

Preparação dos dados: codificação

Exemplos:

- Datas podem ser especialmente problemáticas:
  - 25/12/15
  - 25/dez/2015
  - 25-12-2015
  - 25 de dezembro de 2015

Preparação dos dados: codificação

Exemplos:

- Preços em moedas diferentes: R$ versus US$
  - Câmbio pode variar
  - Possibilidade: armazenar na moeda de origem e converter no momento da análise

Aviso legal

O material presente nesta apresentação foi produzido a partir de informações próprias e coletadas de documentos obtidos publicamente a partir da Internet. Este material contém ilustrações adquiridas de bancos de imagens de origem privada ou pública, não possuindo a intenção de violar qualquer direito pertencente à terceiros e sendo voltado para fins acadêmicos ou meramente ilustrativos. Portanto, os textos, fotografias, imagens, logomarcas e sons presentes nesta apresentação se encontram protegidos por direitos autorais ou outros direitos de propriedade intelectual.

Ao usar este material, o usuário deverá respeitar todos os direitos de propriedade intelectual e industrial, os decorrentes da proteção de marcas registradas da mesma, bem como todos os direitos referentes a terceiros que por ventura estejam, ou estiveram, de alguma forma disponíveis nos slides. O simples acesso a este conteúdo não confere ao usuário qualquer direito de uso dos nomes, títulos, palavras, frases, marcas, dentre outras, que nele estejam, ou estiveram, disponíveis.

É vedada sua utilização para finalidades comerciais, publicitárias ou qualquer outra que contrarie a realidade para o qual foi concebido. Sendo que é proibida sua reprodução, distribuição, transmissão, exibição, publicação ou divulgação, total ou parcial, dos textos, figuras, gráficos e demais conteúdos descritos anteriormente, que compõem o presente material, sem prévia e expressa autorização de seu titular, sendo permitida somente a impressão de cópias para uso acadêmico e arquivo pessoal, sem que sejam separadas as partes, permitindo dar o fiel e real entendimento de seu conteúdo e objetivo. Em hipótese alguma o usuário adquirirá quaisquer direitos sobre os mesmos.

O usuário assume toda e qualquer responsabilidade, de caráter civil e/ou criminal, pela utilização indevida das informações, textos, gráficos, marcas, enfim, todo e qualquer direito de propriedade intelectual ou industrial deste material.

PUC Minas

Virtual