

Aprendizado de Máquina

Web Conf. 1 - Vamos falar de Aprendizado de Máquina

Adriano Rivolli

rivolli@utfpr.edu.br


Especialização em Inteligência Artificial

Universidade Tecnológica Federal do Paraná (UTFPR)

Câmpus Cornélio Procópio

Departamento de Computação

Conteúdo

- 
- 1** Introdução
 - 2** Pré-processamento
 - 3** Regressores
 - 4** Exemplo prático
 - 5** IA Responsável

Introdução

Aprendizado de Máquina 101

- A aplicação de técnicas de AM requer **dados**
- Os dados precisam ser **obtidos**, **analisados** e **preparados**
- Há um ditado na área que diz:
 - ▶ **garbage in, garbage out**

Experiências

■ Professor:

- ▶ Experimentos acadêmicos
- ▶ Aplicação com dados processuais

■ Alunos:

- ▶ Levantem a mão :)

Análise de dados

■ Análise estatística dos dados:


- ▶ Frequências
- ▶ localidade
- ▶ Dispersão
- ▶ Distribuição
- ▶ Correlação

■ Visualização dos dados


- ▶ Gráfica
- ▶ Visualização de informação

Pré-processamento

Problemas que impactam a qualidade dos dados

- 
- Valores ausentes
 - Dados redundantes
 - Dados inconsistentes
 - Ruídos
 - *Outliers*


Causas dos valores ausentes

- 
- Atributo não é considerado quando os dados começaram a ser coletados
 - Alguns atributos são desconhecidos durante a coleta
 - Erro na captura do valor (humano ou mecânico)
 - Não existência dos valores para alguns casos


Dados redundantes

- Dados redundantes ocorre quando os dados são redundantes
- Podem estar relacionados com
 - ▶ Atributo
 - ▶ Instância
- Características:
 - ▶ Não contém novas informações
 - ▶ Irrelevantes para o problema
 - ▶ Podem ser derivados dos que já existem
 - ▶ Duplicados (caso extremo)

Dados inconsistentes

- 
- Podem estar nos atributos preditivos ou atributo alvo
 - Viola alguma restrição do próprio atributo ou da relação entre os atributos
 - ▶ Idade negativa
 - ▶ Data da última visita ao ginecologista (para um homem)


Ruídos nos dados

- 
- O ruído representa uma informação que não é verdadeira
 - Os ruídos podem ser causados por medições incorretas ou distorcidas, erro humano ou mesmo contaminação das amostras
 - Dados ruidosos podem ou não ser inconsistentes

Outliers

- Valores ou objetos anômalos
- São valores não usuais, embora corretos
 - ▶ Por isso não são inconsistentes e nem ruídos

Tratamento dos dados

- 
- Integração
 - Amostragem
 - Dados desbalanceados
 - Limpeza dos dados
 - Transformações


Integração dos dados

- Diferentes fontes de dados
 - ▶ Uso de um identificador único
 - ▶ Validação dos valores e obtenção de novos dados
- Cuidado com a inconsistência e redundância
 - ▶ Mesmo atributo com nomes diferentes
 - ▶ Mesma informação geradas em momentos diferentes
- Meta-dados podem ajudar neste caso

Amostragem

- Algoritmos de AM podem não escalar bem para grandes conjuntos de dados
 - ▶ Problema de memória
 - ▶ Alto custo computacional
 - ▶ Não conseguir ajustar os hiperparâmetros
- Opções:
 - ▶ Amostragem aleatória
 - ▶ Amostragem estratificada
 - ▶ Amostragem progressiva

Dados desbalanceados

- 
- Tendência dos modelos em enfatizar a classe majoritária
 - Adquirir mais dados das classes minoritárias
 - Uso de técnicas para balancear os dados artificialmente:
 - ▶ Sobreamostragem
 - ▶ Subamostragem
 - Uso de diferentes custos para as classes

Limpeza dos dados

- **Valores ausentes**
- Dados redundantes
- Dados inconsistentes
- Dados ruidosos
- Remoção dos *outliers*


Valores ausentes

- Ignorar (a técnica de AM precisa suportar)
- Remover as instâncias ou atributos
- Preencher usando alguma técnica de estimativa
 - ▶ Média
 - ▶ Regressor

Dados redundantes

- Remover instâncias e atributos duplicados
- Remover atributos altamente correlacionados
- Transformação dos dados

Dados inconsistentes


- 
- Primeiro, é necessário identificá-los
 - Depois, trate-os como valores ausentes

Dados ruidosos

- Existem técnicas de AM para identificar ruídos nos dados
- Os **filtros de ruídos** possuem vieses semelhante aos dos algoritmos
- Existem filtros de ruído para os atributos preditivos e para o **atributo alvo**
- Não é possível garantir que todos os ruídos sejam identificados

PIO, P. B.; GARCIA, L. P. F.; **RIVOLLI, A.** . Meta-Learning Approach for Noise Filter Algorithm Recommendation. In: Knowledge Discovery, Mining and Learning (KDMiLe), 2022, Campinas. Knowledge Discovery, Mining and Learning (KDMiLe), 2022.

Transformação dos dados

- 
- Conversão dos tipos de atributos
 - Alteração das escalas dos valores
 - Redução da dimensionalidade
 - ▶ Seleção de atributos
 - ▶ Transformação de atributos

Regressores

Resumo da semana

- Algoritmos de regressão
 - ▶ Regressão Linear
 - ▶ Regularização: Lasso e Ridge
 - ▶ Regressão Polinomial
- Metodologia de avaliação
- Medidas de avaliação


Regressores

- Baseado em distância: [KNeighborsRegressor](#)
- Simbólico: [DecisionTreeRegressor](#)
- Vetores de suporte: [SVR](#)
- Comitê de árvores (floresta aleatória): [RandomForestRegressor](#)


KNeighborsRegressoro

- Encontra os k vizinhos mais próximos
- Prediz a média do atributo alvo
- Pode ser ponderado pela distância (ordem de proximidade)

DecisionTreeRegressor

- 
- Gera uma árvore de decisão
 - ▶ Os atributos preditivos são usados nos nós internos
 - ▶ O atributo alvo é usado nas folhas
 - Prediz o valor médio das instâncias presentes na folha

Support Vector Regressor (SVR)


- 
- Uma extensão do algoritmo de SVM
 - Encontra os vetores de suporte
 - O preditor é feito a partir da distância do ponto em relação à margem

RandomForestRegressor

- Gera um conjunto de árvores regressoras
- Cada árvore prediz um valor
- Utiliza a média destes valores para a predição final

Exemplo prático

O que vamos ver?

- 
- Como carregar um dataset no Google Collab
 - Como usar os algoritmos de regressão
 - Como usar um regressor para preencher valores ausentes

IA Responsável

Processos recentes em IA (e AM)

- Progressos na IA estão melhorando a produtividade
 - ▶ Redução de custos e tempo de desenvolvimento
 - ▶ Rapidez e abrangência
 - ▶ Grandes oportunidades/riscos
- Não há restrições de fronteiras no mundo digital

Escopo da ia (e AM)

■ Hoje: estreita (fraca)

- ▶ Especializada para tarefas específicas
- ▶ Pode ser ligada/desligada a qualquer momento

■ Futuro: geral (forte)

- ▶ *Artificial General Intelligence* (AGI)
- ▶ Altas capacidades 'cognitivas' (desempenho similar ao humano)

IA (e AM) de fronteira

- Onde os avanços mais recentes estão ocorrendo
- Modelos de IA de propósito geral e de elevada capacidade
- IA generativa
 - ▶ Grandes modelos de linguagem (LLM)
 - ▶ Direção autônoma
 - ▶ Robótica
 - ▶ Tecnologias novas (a surgirem)


Riscos da IA (e AM)

- Resultados indesejados (viés algorítmico)
 - ▶ Carros desgovernados
 - ▶ Discriminação pelo gênero, origem, situação social
- Manipulação de informações e desinformação
- Ameaça à privacidade e liberdade individual
- Uso para objetivos indesejados
- Perda de habilidades humanas
- Desemprego tecnológico
- Desequilíbrio ambiental


Princípios da IA (e AM) responsável

- Justiça e Equidade
- Confiabilidade e segurança
- Privacidade e segurança
- Inclusão e diversidade
- Transparência e explicabilidade
- Responsabilidade e prestação de contas
- Sustentabilidade e impacto ambiental

Pontos chaves

- 
- Regulamentação/Legislação
 - Origem e qualidade dos dados usados
 - Avaliação dos sistemas inteligentes
 - Auditoria e monitoramento dos sistemas
 - Responsabilização dos usuários

Para finalizar...

- 
- Quem arrisca alguma estimativa de futuro?
 - Recomendação de filmes e séries sobre IA
 - Pense, logo exista!
 - Seja legal :)
 - #paz