

Análise de Redes Sociais

Obtenção e Organização dos Dados

Prof. Luciano Antonio Digiampietri
EACH-USP

Obtenção dos Dados

Considerando que uma rede social é um conjunto de indivíduos e suas relações, existe uma gama muito grande de possibilidades de redes a serem construídas e analisadas.

Obtenção dos Dados

Historicamente, muitas das primeiras redes sociais analisadas eram compostas por pequenos grupos de indivíduos (com laços familiares, geográficos, profissionais etc.).

Obtenção dos Dados

Historicamente, muitas das primeiras redes sociais analisadas eram compostas por pequenos grupos de indivíduos (com laços familiares, geográficos, profissionais etc.).

Porém, com a popularização da Internet e, em especial, das redes sociais online se tornou possível analisar milhares ou milhões de indivíduos e suas relações.

REDE SOCIAL	USUÁRIOS ATIVOS¹ (no mundo)
1. Facebook	2,91 bilhões
2. YouTube	2,56 bilhões
3. WhatsApp	2 bilhões
4. Instagram	1,45 bilhão
5. WeChat	1,26 bilhão
6. Messenger	1 bilhão
7. TikTok	970 milhões
8. LinkedIn	830 milhões
9. Douyin	613 milhões
10. Snapchat	589 milhões

¹ <https://resultadosdigitais.com.br/marketing/redes-sociais-mais-usadas-no-brasil/>

REDE SOCIAL	USUÁRIOS ATIVOS² (no Brasil)
1. WhatsApp	165 milhões
2. YouTube	138 milhões
3. Instagram	122 milhões
4. Facebook	116 milhões
5. TikTok	73,5 milhões
6. Messenger	65,5 milhões
7. LinkedIn	56 milhões
8. Pinterest	30 milhões
9. Twitter	19 milhões
10. Snapchat	7,6 milhões

²<https://resultadosdigitais.com.br/marketing/redes-sociais-mais-usadas-no-brasil/>

Obtenção dos Dados

Redes sociais podem ser compostas de diferentes formas, por exemplo, a partir de dados de publicações científicas.

Obtenção dos Dados

Redes sociais podem ser compostas de diferentes formas, por exemplo, a partir de dados de publicações científicas.

A Plataforma Lattes³ conta com o currículo de 7,3 milhões de pessoas.

³<https://lattes.cnpq.br/>

Obtenção dos Dados

Redes sociais podem ser compostas de diferentes formas, por exemplo, a partir de dados de publicações científicas.

A Plataforma Lattes³ conta com o currículo de 7,3 milhões de pessoas.

Contém relações de coautoria, orientação, participação em bancas, participação em projetos, relações profissionais, entre outras.

³<https://lattes.cnpq.br/>

Obtenção dos Dados

A base de dados do DBLP⁴ (Digital Bibliography & Library Project (DBLP)) conta com informações de mais de 3 milhões de pesquisadores que publicaram artigos relacionados à ciência da computação.

⁴<https://dblp.org/>

Obtenção dos Dados

APIs

Gratuitas

Licenças Acadêmicas

Pagas

Web crawlers

Conjuntos de dados disponíveis

Obtenção dos Dados

Conjuntos de Dados Disponíveis

Disponibilizados pelos mantenedores dos dados
(por exemplo, dados do DBLP);

Organizados em diferentes sites: Kaggle⁵, UCI⁶,
Azure⁷

Disponibilizados de forma esparsa.

⁵<https://www.kaggle.com/datasets>

⁶<https://archive.ics.uci.edu/ml/datasets.php>

⁷<https://azure.microsoft.com/pt-br/services/open-datasets/>

Cuidados com o uso dos dados

Lei Geral de Proteção de Dados Pessoais (LGPD)

⁸<https://www.serpro.gov.br/lgpd/menu/protecao-de-dados/dados-publicos-lgpd>

Cuidados com o uso dos dados

Lei Geral de Proteção de Dados Pessoais (LGPD)

A lei cita “dados pessoais cujo acesso é público”. Como essa categoria de dados deve ser tratada?

⁸<https://www.serpro.gov.br/lgpd/menu/protecao-de-dados/dados-publicos-lgpd>

Cuidados com o uso dos dados

Lei Geral de Proteção de Dados Pessoais (LGPD)

A lei cita “dados pessoais cujo acesso é público”. Como essa categoria de dados deve ser tratada?

“Deve ser tratada considerando a finalidade, a boa-fé e o interesse público que justificaram a sua disponibilização. A LGPD define, por exemplo, que uma organização pode, sem precisar pedir novo consentimento, tratar dados tornados anterior e manifestamente públicos pelo titular.”⁸

⁸<https://www.serpro.gov.br/lgpd/menu/protecao-de-dados/dados-publicos-lgpd>

Cuidados com o uso dos dados

Comitê de Ética em Pesquisas⁹

⁹<https://www.gov.br/defesa/pt-br/assuntos/hfa/ensino-e-pesquisa/comite-de-etica-em-pesquisa-cep-hfa-1>

Cuidados com o uso dos dados

Comitê de Ética em Pesquisas⁹

Um CEP é um colegiado interdisciplinar e independente, de relevância pública, de caráter consultivo, deliberativo e educativo, criado para defender os interesses dos participantes da pesquisa em sua integridade e dignidade e para contribuir no desenvolvimento da pesquisa dentro de padrões éticos.

⁹<https://www.gov.br/defesa/pt-br/assuntos/hfa/ensino-e-pesquisa/comite-de-etica-em-pesquisa-cep-hfa-1>

Cuidados com o uso dos dados

Plataforma Brasil¹⁰

¹⁰<https://www.gov.br/defesa/pt-br/assuntos/hfa/ensino-e-pesquisa/comite-de-etica-em-pesquisa-cep-hfa-1>

Cuidados com o uso dos dados

Plataforma Brasil¹⁰

A Plataforma Brasil é o sistema oficial de lançamento de pesquisas para análise e monitoramento do Sistema CEP/CONEP. O pesquisador deve realizar cadastro na Plataforma Brasil para efetuar a submissão do projeto ao CEP.

¹⁰<https://www.gov.br/defesa/pt-br/assuntos/hfa/ensino-e-pesquisa/comite-de-etica-em-pesquisa-cep-hfa-1>

Cuidados com o uso dos dados

**Termo de Consentimento Livre e Esclarecido
(TCLE)¹¹**

¹¹<http://www5.each.usp.br/tcle/>

Cuidados com o uso dos dados

Termo de Consentimento Livre e Esclarecido (TCLE)¹¹

“O Termo de Consentimento Livre e Esclarecido (TCLE) é documento no qual é explicitado o consentimento livre e esclarecido do participante e/ou de seu responsável legal, de forma escrita, devendo conter todas as informações necessárias, em linguagem clara e objetiva, de fácil entendimento, para o mais completo esclarecimento sobre a pesquisa a qual se propõe participar (...) .”

¹¹<http://www5.each.usp.br/tcle/>

Limpeza de Dados

Limpeza de Dados

Limpeza de dados é o processo de corrigir ou remover dados corrompidos, incorretos, inválidos ou fora do contexto das análises que serão realizadas.

Enriquecimento do Conjunto de Dados

Enriquecimento do Conjunto de Dados

Busca por informações complementares (potencialmente de outras fontes), que poderão ser utilizadas na análise de dados.

Resolução de Entidades (*Entity Resolution*)

Resolução de Entidades (*Entity Resolution*)

Tarefa de desambiguação do conjunto de dados de forma a identificar corretamente as entidades tratadas, citadas ou analisadas.

Resolução de Entidades

Múltiplas aplicações, dependendo da análise que se pretende realizar.

Resolução de Entidades

Múltiplas aplicações, dependendo da análise que se pretende realizar.

Identificar os diversos registos que se referem a uma entidade.

Resolução de Entidades

Múltiplas aplicações, dependendo da análise que se pretende realizar.

Identificar os diversos registros que se referem a uma entidade.

Identificar informações duplicadas.

Resolução de Entidades

Múltiplas aplicações, dependendo da análise que se pretende realizar.

Identificar os diversos registros que se referem a uma entidade.

Identificar informações duplicadas.

Padronização na forma na qual uma entidade é referenciada.

Resolução de Entidades

Entidades podem ser:

Resolução de Entidades

Entidades podem ser:

Pessoas

Resolução de Entidades

Entidades podem ser:

Pessoas

Locais

Resolução de Entidades

Entidades podem ser:

Pessoas

Locais

Instituições

Resolução de Entidades

Entidades podem ser:

Pessoas

Locais

Instituições

Produtos

Resolução de Entidades

Entidades podem ser:

Pessoas

Locais

Instituições

Produtos

Datas ...

Resolução de Entidades

**Exemplo considerando dados da Plataforma
Lattes x Projeto DBLP**

Resolução de Entidades

**Exemplo considerando dados da Plataforma
Lattes x Projeto DBLP
Dados de Currículo x Dados de Publicação**

Resolução de Entidades - Lattes

**DIGIAMPIETRI, L. A.; ALCAZAR, J. J. P. ; MEDEIROS, C. M. B. .
AI Planning in Web Services Composition: a review of current
approaches and a new solution. In: VI Encontro Nacional de
Inteligência Artificial, 2007, Rio de Janeiro. Anais do XXVII
congresso da SBC, 2007.**

**DIGIAMPIETRI, L. ; ALCÁZAR, José de Jesús Pérez ;
MEDEIROS, C. B. . AI Planning in Web Services Composition: a
review of current approaches and a new solution. In: Encontro
Nacional de Inteligencia Artificial - ENIA, 2007, Rio de Janeiro.
ANAIS DO XXVII CONGRESSO DA SBC, 2007. v. XXVII.**

Resolução de Entidades - Lattes

DIGIAMPIETRI, L. A.; ALCÁZAR, J. J. P. ; MEDEIROS, C. M. B. .

DIGIAMPIETRI, L. ; ALCÁZAR, José de Jesús Pérez ;
MEDEIROS, C. B. .

Resolução de Entidades - Lattes

DIGIAMPIETRI, L. A.; **ALCAZAR, J. J. P.**; **MEDEIROS, C. M. B.** .

AI Planning in Web Services Composition: a review of current approaches and a new solution.

DIGIAMPIETRI, L. ; **ALCÁZAR, José de Jesús Pérez** ;

MEDEIROS, C. B. . AI Planning in Web Services Composition: a review of current approaches and a new solution.

Resolução de Entidades - Lattes

DIGIAMPIETRI, L. A.; **ALCAZAR, J. J. P.**; **MEDEIROS, C. M. B.** .

AI Planning in Web Services Composition: a review of current approaches and a new solution. In: VI Encontro Nacional de Inteligência Artificial,

DIGIAMPIETRI, L. ; **ALCÁZAR, José de Jesús Pérez** ;

MEDEIROS, C. B. . AI Planning in Web Services Composition: a review of current approaches and a new solution. In: Encontro Nacional de Inteligencia Artificial - ENIA,

Resolução de Entidades - Lattes

DIGIAMPIETRI, L. A.; **ALCAZAR, J. J. P.**; **MEDEIROS, C. M. B.** .

AI Planning in Web Services Composition: a review of current approaches and a new solution. In: VI Encontro Nacional de Inteligência Artificial, 2007, Rio de Janeiro. Anais do XXVII congresso da SBC, 2007.

DIGIAMPIETRI, L. ; **ALCÁZAR, José de Jesús Pérez** ;

MEDEIROS, C. B. . AI Planning in Web Services Composition: a review of current approaches and a new solution. In: Encontro Nacional de Inteligencia Artificial - ENIA, 2007, Rio de Janeiro. ANAIS DO XXVII CONGRESSO DA SBC, 2007.

Resolução de Entidades - Lattes

DIGIAMPIETRI, L. A.; **ALCAZAR, J. J. P.**; **MEDEIROS, C. M. B.** .

AI Planning in Web Services Composition: a review of current approaches and a new solution. In: VI Encontro Nacional de Inteligência Artificial, 2007, Rio de Janeiro. Anais do XXVII congresso da SBC, 2007.

DIGIAMPIETRI, L. ; **ALCÁZAR, José de Jesús Pérez** ;

MEDEIROS, C. B. . AI Planning in Web Services Composition: a review of current approaches and a new solution. In: Encontro Nacional de Inteligencia Artificial - ENIA, 2007, Rio de Janeiro. ANAIS DO XXVII CONGRESSO DA SBC, 2007. v. XXVII.

Resolução de Entidades - DBLP

Juan Carlos Zúñiga, José de Jesús Pérez Alcázar, Luciano A. Digiampietri, Silvio Ernesto Barbin: **A loosely coupled architecture for automatic composition of web services applications. Int. J. Metadata Semant. Ontologies 9(3): 241-251 (2014)**

Jesús Pascual Mena-Chalco, Luciano Antonio Digiampietri, Fabrício Martins Lopes, Roberto Marcondes Cesar Junior: **Brazilian bibliometric coauthorship networks. J. Assoc. Inf. Sci. Technol. 65(7): 1424-1445 (2014)**

Resolução de Entidades - DBLP

Juan Carlos Zúñiga, José de Jesús Pérez Alcázar, Luciano A. Digiampietri, Silvio Ernesto Barbin: A loosely coupled architecture for automatic composition of web services applications. Int. J. Metadata Semant. Ontologies 9(3): 241-251 (2014)

Jesús Pascual Mena-Chalco, Luciano Antonio Digiampietri, Fabrício Martins Lopes, Roberto Marcondes Cesar Junior: Brazilian bibliometric coauthorship networks. J. Assoc. Inf. Sci. Technol. 65(7): 1424-1445 (2014)

Inteligência Artificial e ARS

**Muitos problemas de análise de redes sociais são
(ou podem ser) tratados utilizando IA:**

Inteligência Artificial e ARS

**Muitos problemas de análise de redes sociais são
(ou podem ser) tratados utilizando IA:**

Classificação

Inteligência Artificial e ARS

Muitos problemas de análise de redes sociais são (ou podem ser) tratados utilizando IA:

Classificação - e.g., identificação de notícias falsas

Inteligência Artificial e ARS

**Muitos problemas de análise de redes sociais são
(ou podem ser) tratados utilizando IA:**

Classificação

Regressão

Inteligência Artificial e ARS

Muitos problemas de análise de redes sociais são (ou podem ser) tratados utilizando IA:

- Classificação

- Regressão - e.g., análise de tendências

Inteligência Artificial e ARS

**Muitos problemas de análise de redes sociais são
(ou podem ser) tratados utilizando IA:**

Classificação

Regressão

Agrupamento

Inteligência Artificial e ARS

Muitos problemas de análise de redes sociais são (ou podem ser) tratados utilizando IA:

- Classificação

- Regressão

- Agrupamento - e.g., formação de bolhas ideológicas

Inteligência Artificial e ARS

**Muitos problemas de análise de redes sociais são
(ou podem ser) tratados utilizando IA:**

Classificação

Regressão

Agrupamento

Detecção de Anomalias (*outliers*)

Inteligência Artificial e ARS

Muitos problemas de análise de redes sociais são (ou podem ser) tratados utilizando IA:

- Classificação

- Regressão

- Agrupamento

- Detecção de Anomalias (*outliers*) - e.g., identificação de bots

Inteligência Artificial e ARS

Muitos problemas de análise de redes sociais são (ou podem ser) tratados utilizando IA:

- Classificação

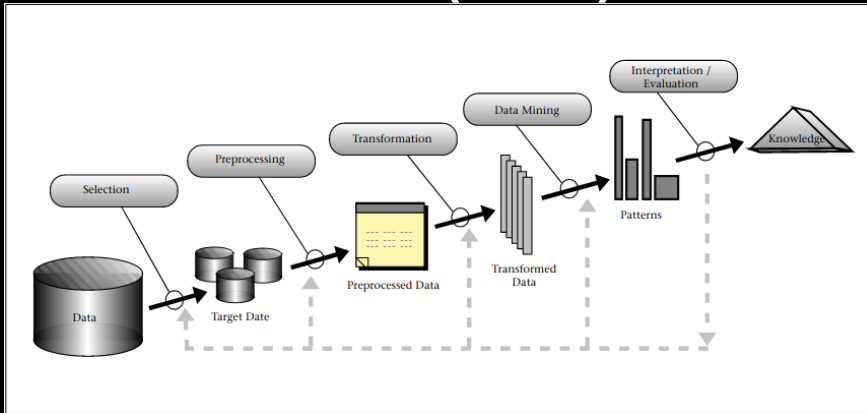
- Regressão

- Agrupamento

- Detecção de Anomalias (*outliers*)

- ...

Descoberta de Conhecimento em Bancos de Dados (KDD)¹²



¹²Fayyad et al. (1996)

KDD - Etapas/Tarefas

KDD - Etapas/Tarefas

Seleção

Pré-processamento

Transformação

Mineração de Dados

Avaliação/Interpretação

Seleção

**Seleção do subconjunto dos dados de interesse
(registros e atributos/características);**

Seleção

Seleção do subconjunto dos dados de interesse (registros e atributos/características);

Tarefa intimamente ligada ao problema que se deseja investigar (ao conhecimento que se pretende ‘descobrir’).

Pré-processamento

O pré-processamento, juntamente com a limpeza de dados, pode envolver: remoção de ruído e estratégias para lidar com valores faltantes, podendo também incluir atividades de resolução de entidades.

Valores Faltantes

Há três abordagem principais para lidar com valores faltantes:

Valores Faltantes

Há três abordagem principais para lidar com valores faltantes:

- Descartar os registros com valores faltantes;**

Valores Faltantes

Há três abordagens principais para lidar com valores faltantes:

- Descartar os registros com valores faltantes;**

- Descartar as colunas, atributos ou características com valores faltantes;**

Valores Faltantes

Há três abordagem principais para lidar com valores faltantes:

- Descartar os registros com valores faltantes;

- Descartar as colunas, atributos ou características com valores faltantes;

- Preencher os valores faltantes (*missing value imputation*).

Preenchimento de valores faltantes

Além da tentativa de obter o valor faltante (por exemplo, por enriquecimento dos dados), há três abordagens principais para o preenchimento automático:

Preenchimento de valores faltantes

Além da tentativa de obter o valor faltante (por exemplo, por enriquecimento dos dados), há três abordagens principais para o preenchimento automático:

- Preencher com valor zero/nulo;

Preenchimento de valores faltantes

Além da tentativa de obter o valor faltante (por exemplo, por enriquecimento dos dados), há três abordagens principais para o preenchimento automático:

- Preencher com valor zero/nulo;

- Preencher com média, moda ou mediana;

Preenchimento de valores faltantes

Além da tentativa de obter o valor faltante (por exemplo, por enriquecimento dos dados), há três abordagens principais para o preenchimento automático:

- Preencher com valor zero/nulo;

- Preencher com média, moda ou mediana;

- Preencher utilizando IA (por exemplo, com auxílio de um regressor ou classificador).

Transformação

Transformação dos dados para o formato que será apresentado ao algoritmo de mineração de dados.

Transformação

Transformação dos dados para o formato que será apresentado ao algoritmo de mineração de dados. Pode envolver agregação, normalização, criação de novos atributos, redução de dimensionalidade (seleção de atributos ou projeção) etc.

Transformação

Transformação dos dados para o formato que será apresentado ao algoritmo de mineração de dados. Pode envolver agregação, normalização, criação de novos atributos, redução de dimensionalidade (seleção de atributos ou projeção) etc.

Para dados textuais, envolve a transformação do texto em dados tabulares.

Mineração de Dados (propriamente dita)

Mineração de Dados (propriamente dita)

Identificação de regras de associação

Mineração de Dados (propriamente dita)

Identificação de regras de associação

Criação do modelo de classificação

Mineração de Dados (propriamente dita)

Identificação de regras de associação

Criação do modelo de classificação

Criação do modelo de regressão

Mineração de Dados (propriamente dita)

Identificação de regras de associação

Criação do modelo de classificação

Criação do modelo de regressão

Agrupamento/Clusterização

Mineração de Dados (propriamente dita)

Identificação de regras de associação

Criação do modelo de classificação

Criação do modelo de regressão

Agrupamento/Clusterização

Identificação de Anomalias (*Outliers*)

Análise/Interpretação dos resultados

Análise/Interpretação dos resultados

Análise crítica dos resultados obtidos, considerando não só as métricas de avaliação, mas também o impacto de sua aplicação.

Conjuntos de treinamento, validação e teste

Para se medir o desempenho do modelo que estamos construindo (por exemplo, de classificação ou regressão), precisamos testar o modelo com dados que não foram usados em seu treinamento.

Conjuntos de treinamento, validação e teste

Para se medir o desempenho do modelo que estamos construindo (por exemplo, de classificação ou regressão), precisamos testar o modelo com dados que não foram usados em seu treinamento.

Conjunto de Treinamento;

Conjunto de Validação;

Conjunto de Teste.

Conjuntos ...

Conjuntos ...

Conjunto de Treinamento:
usado para treinar o modelo;

Conjuntos ...

Conjunto de Treinamento:

usado para treinar o modelo;

Conjunto de Validação:

usado para uma validação inicial do modelo e/ou para testar diferentes configurações de hiperparâmetros ou diferentes estratégias de pré-processamento e atributos;

Conjunto de Teste:

usado para efetivamente testar o modelo (nenhuma configuração do modelo deve ter acesso a dados do conjunto de teste).

Conjuntos ...

Existem diferentes abordagem para a separação dos dados nestes conjuntos.

Conjuntos ...

Existem diferentes abordagem para a separação dos dados nestes conjuntos.

Tipicamente é feita uma divisão estratitificada do conjunto de dados;

Conjuntos ...

Existem diferentes abordagem para a separação dos dados nestes conjuntos.

Tipicamente é feita uma divisão estratificada do conjunto de dados;

O conjunto de treinamento e validação podem ficar inicialmente juntos se for realizada uma validação cruzada;

Conjuntos ...

Existem diferentes abordagem para a separação dos dados nestes conjuntos.

Tipicamente é feita uma divisão estratificada do conjunto de dados;

O conjunto de treinamento e validação podem ficar inicialmente juntos se for realizada uma validação cruzada;

Para conjuntos de dados temporais, ao invés de uma divisão 'aleatória estratificada' é comum ser realizada uma divisão temporal.

Validação cruzada

Validação cruzada

O conjunto de treinamento e validação é subdividido em k subconjuntos (por exemplo, de forma 'aleatória estratificada').

Validação cruzada

O conjunto de treinamento e validação é subdividido em k subconjuntos (por exemplo, de forma 'aleatória estratificada').

O seguinte processo é executado k vezes (alternando-se o conjunto de validação):

Validação cruzada

O conjunto de treinamento e validação é subdividido em k subconjuntos (por exemplo, de forma 'aleatória estratificada').

O seguinte processo é executado k vezes (alternando-se o conjunto de validação):

Constrói o modelo com $k-1$ subconjuntos e testa com o subconjunto restante;

Validação cruzada

O conjunto de treinamento e validação é subdividido em k subconjuntos (por exemplo, de forma 'aleatória estratificada').

O seguinte processo é executado k vezes (alternando-se o conjunto de validação):

Constrói o modelo com $k-1$ subconjuntos e testa com o subconjunto restante;

O desempenho é tipicamente medido pela média dos k resultados.

Métricas de Desempenho/Avaliação

Métricas de Desempenho/Avaliação

Existem diversas métricas para avaliar o desempenho de algoritmos de classificação. Muitas delas podem ser obtidas a partir da matriz de confusão.

Métricas de Desempenho/Avaliação

Existem diversas métricas para avaliar o desempenho de algoritmos de classificação. Muitas delas podem ser obtidas a partir da matriz de confusão.

Matriz de Confusão é uma tabela que apresenta o número de elementos que foram classificados em cada classe em contraste com sua classe real.

Matriz de Confusão

		Valor Predito	
		Positivo	Negativo
Valor Real	Positivo	VP	FN
	Negativo	FP	VN

VP - Verdadeiro Positivo

FN - Falso Negativo

FP - Falso Positivo

VN - Verdadeiro Negativo

P - Elementos da classe positiva (VP + FN)

N - Elementos da classe negativa (FP + VN)

Métricas de Desempenho/Avaliação

$$\text{Acurácia (accuracy (ACC))} = \frac{VP + VN}{P + N}$$

$$\text{Precisão (precision)} = \frac{VP}{VP + FP}$$

$$\text{Revocação (recall, sensibilidade)} = \frac{VP}{P}$$

$$\text{Especificidade (specificity)} = \frac{VN}{N}$$

$$\text{Medida F1 (F1 score)} = 2 * \frac{\text{precisao} * \text{revocacao}}{\text{precisao} + \text{revocacao}}$$

Métricas de Desempenho/Avaliação

Para modelos de regressão são utilizadas outras métricas de desempenho.

Métricas de Desempenho/Avaliação

Para modelos de regressão são utilizadas outras métricas de desempenho.

Erro Absoluto Médio (*Mean Absolute Error (MAE)*):

Métricas de Desempenho/Avaliação

Para modelos de regressão são utilizadas outras métricas de desempenho.

Erro Absoluto Médio (*Mean Absolute Error (MAE)*):

$$\frac{\sum_{i=1}^n (|y_i - \hat{y}_i|)}{n}$$

sendo y_i o i-ésimo valor observado de y e $\hat{y}_i$ o valor estimado para y_i .

Métricas de Desempenho/Avaliação

Erro Quadrático Médio (*Mean Square Error (MSE)*):

Métricas de Desempenho/Avaliação

Erro Quadrático Médio (*Mean Square Error (MSE)*):

$$\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}$$

sendo y_i o i-ésimo valor observado de y e $\hat{y}_i$ o valor estimado para y_i .

Métricas de Desempenho/Avaliação

Raiz Quadrada do Erro Quadrático Médio (*Root Mean Absolute Error (RMSE)*):

Métricas de Desempenho/Avaliação

Raiz Quadrada do Erro Quadrático Médio (*Root Mean Absolute Error (RMSE)*):

$$\sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}$$

sendo y_i o i-ésimo valor observado de y e $\hat{y}_i$ o valor estimado para y_i .

Métricas de Desempenho/Avaliação

Métricas de Desempenho/Avaliação

Coeficiente(s) de Correlação.

Métricas de Desempenho/Avaliação

Métricas utilizadas para avaliar ranqueamentos:

Métricas de Desempenho/Avaliação

Métricas utilizadas para avaliar ranqueamentos:

Coeficiente de correlação de Pearson =

$$\frac{\sum_{i=1}^n (x_i - \bar{x}_i) * (y_i - \bar{y}_i)}{\sqrt{\sum_{i=1}^n (x_i - \bar{x}_i)^2 * \sum_{i=1}^n (y_i - \bar{y}_i)^2}}$$

sendo x_i e y_i os i -ésimos valores observados para as variáveis x e y , e $\bar{x}$ e $\bar{y}$, respectivamente, as médias dos valores de x e de y .

Métricas de Desempenho/Avaliação

Métricas utilizadas para avaliar ranqueamentos:

Coeficiente de determinação (R^2) =

$$1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

sendo y_i o valor observado $\bar{y}$ a média das observações e $\hat{y}_i$ o valor estimado/predito para y_i .

Métricas de Desempenho/Avaliação

Métricas utilizadas para avaliar ranqueamentos:

Média da Classificação Recíproca (*Mean reciprocal rank (MRR)*) =

$$\frac{1}{|Q|} * \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

sendo $rank_i$ a posição do primeiro resultado relevante da consulta i e $|Q|$ o número de consultas.

Métricas de Desempenho/Avaliação

Métricas utilizadas para avaliar ranqueamentos:

Ganho Cumulativo Descontado (*DCG*) e
Normalizado (*NDCG*)¹³

$$DCG_p = \sum_{i=1}^p \frac{rel_i}{\log_2(i+1)} \quad NDCG_p = \frac{DCG_p}{\sum_{i=1}^{|REL_p|} \frac{rel_i}{\log_2(i+1)}}$$

sendo rel_i a relevância do resultado na posição i e $|REL_p|$ a lista ordenada de documentos até a posição p .

¹³https://en.wikipedia.org/wiki/Discounted_cumulative_gain

Métricas de Desempenho/Avaliação

Métricas utilizadas para avaliar ranqueamentos¹⁴:

Precisão em k (*Precision at k*): número de itens relevantes entre os k primeiros ranqueados dividido por k: $Precision@k = \frac{Relevantes_k}{k}$

¹⁴Ranqueamentos e/ou sistemas de recomendação.

Métricas de Desempenho/Avaliação

Métricas utilizadas para avaliar ranqueamentos¹⁵:

Revocação em k (*Recall at k*): número de itens relevantes entre os k primeiros ranqueados dividido pelo número total de itens relevantes:

$$Recall@k = \frac{Relevantes_k}{TotalDeRelevantes}$$

¹⁵Ranqueamentos e/ou sistemas de recomendação.

Balanceamento dos Dados

Balanceamento dos Dados

Em diversas aplicações reais, a distribuição das classes no conjunto de dados não é uniforme.

Balanceamento dos Dados

Em diversas aplicações reais, a distribuição das classes no conjunto de dados não é uniforme. Por exemplo, dado um conjunto de usuários arbitrários de uma rede social, provavelmente a maioria não será de *bots*.

Balanceamento dos Dados

Em diversas aplicações reais, a distribuição das classes no conjunto de dados não é uniforme.

Por exemplo, dado um conjunto de usuários arbitrários de uma rede social, provavelmente a maioria não será de *bots*.

E muitos algoritmos de aprendizagem de máquina podem ‘aprender’ a dar preferência para a classe majoritária.

Balanceamento dos Dados

Balanceamento dos Dados

Para evitar isso, podemos balancear o conjunto de treinamento.

Balanceamento dos Dados

Para evitar isso, podemos balancear o conjunto de treinamento.

Isto pode causar uma diminuição em algumas métricas de desempenho (e.g., na acurácia) mas aumentar outras métricas (por exemplo, a revocação da classe minoritária).

Balanceamento dos Dados

Undersampling: remoção dos elementos da classe majoritária até atingir a mesma quantidade de elementos¹⁶ da classe minoritária.

¹⁶no conjunto de treinamento

Balanceamento dos Dados

Undersampling: remoção dos elementos da classe majoritária até atingir a mesma quantidade de elementos¹⁶ da classe minoritária.

Remoção aleatória de elementos.

¹⁶no conjunto de treinamento

Balanceamento dos Dados

Undersampling: remoção dos elementos da classe majoritária até atingir a mesma quantidade de elementos¹⁶ da classe minoritária.

Remoção aleatória de elementos.

Remoção de elementos redundantes.

¹⁶no conjunto de treinamento

Balanceamento dos Dados

Undersampling: remoção dos elementos da classe majoritária até atingir a mesma quantidade de elementos¹⁶ da classe minoritária.

Remoção aleatória de elementos.

Remoção de elementos redundantes.

Substituição de conjuntos de elementos ‘próximos’ por um novo elemento representativo do conjunto.

¹⁶no conjunto de treinamento

Balanceamento dos Dados

Oversampling: adicionamos elementos à classe minoritária até atingir a mesma quantidade de elementos¹⁷ da classe majoritária.

¹⁷no conjunto de treinamento

Balanceamento dos Dados

Oversampling: adicionamos elementos à classe minoritária até atingir a mesma quantidade de elementos¹⁷ da classe majoritária.

Repetição aleatória de elementos.

¹⁷no conjunto de treinamento

Balanceamento dos Dados

Oversampling: adicionamos elementos à classe minoritária até atingir a mesma quantidade de elementos¹⁷ da classe majoritária.

Repetição aleatória de elementos.

Criação de novos elementos de forma sintética.

¹⁷no conjunto de treinamento

Balanceamento dos Dados

Oversampling: adicionamos elementos à classe minoritária até atingir a mesma quantidade de elementos¹⁷ da classe majoritária.

Repetição aleatória de elementos.

Criação de novos elementos de forma sintética.

Repetição/criação com base nos elementos 'mais difíceis' de classificar.

¹⁷no conjunto de treinamento

Referências

- 1 Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic . From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3), 37-54. 1996.
- 2 RUSSEL, S.; NORVIG, P. Inteligência Artificial. Editora Campus, 2013 (3ª edição)

Análise de Redes Sociais

Obtenção e Organização dos Dados

Prof. Luciano Antonio Digiampietri
EACH-USP