Processamento de Linguagem Natural

Aprendizagem de Máquina e PLN Estatístico

Marlo Souza[1]
[1]Universidade Federal da Bahia - Brasil

25 de julho de 2024

Sumário

Aprendizado de Máquina

Aprendizado de Representação

Pensando problemas de PLN como problemas de AM

Aprendizado de Máquina

Aprendizado de Representação

Pensando problemas de PLN como problemas de AM

O ciclo do PLN

Aprendizado de Máquina

Aprendizado de Representação

Pensando problemas de PLN como problemas de AM

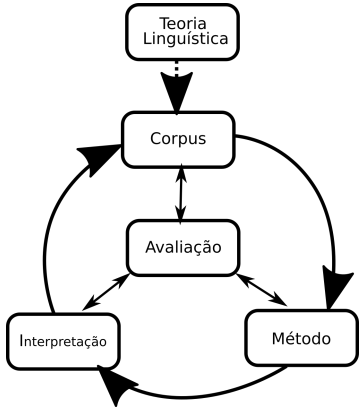


Figura: O Ciclo de Desenvolvimento em PLN

Dados e Representação

Aprendizado de Máquina

Aprendizado de Representação

Pensando problemas de PLN como problemas de AM

É preciso lembrar que dados são, por definição, representações da realidade codificadas por humano, logo não existem “dados crus”, pois dados são sempre o resultado de um processo de informação (coleta, codificação e armazenamento)

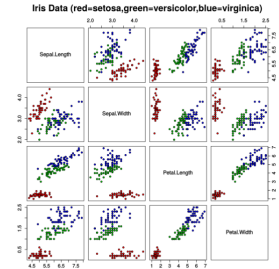
Dados e Representação

Aprendizado de Máquina

Aprendizado de Representação

Pensando problemas de PLN como problemas de AM

É preciso lembrar que dados são, por definição, representações da realidade codificadas por humano, logo não existem “dados crus”, pois dados são sempre o resultado de um processo de informação (coleta, codificação e armazenamento)



Aprendizado de Máquina 5

Aprendizado de Representação

Pensando problemas de PLN como problemas de AM

Aprendizado de Máquina

O que é AM

Aprendizado de Máquina 6
Aprendizado de Representação
Pensando problemas de PLN como problemas de AM

- Área que investiga métodos para que um sistema inteligente possa "aprender"a partir de dados;

O que é AM

Aprendizado de Máquina 6
Aprendizado de Representação
Pensando problemas de PLN como problemas de AM

- Área que investiga métodos para que um sistema inteligente possa "aprender"a partir de dados;
- "Aprender"= Reconhecer padrões de co-ocorrência;

O que é AM

Aprendizado de Máquina 6
Aprendizado de Representação
Pensando problemas de PLN como problemas de AM

- Área que investiga métodos para que um sistema inteligente possa "aprender"a partir de dados;
- "Aprender"= Reconhecer padrões de co-ocorrência;
- Envolve conhecimentos de I.A., Ciência da Computação, Estatística, Psicologia e Ciência Cognitiva.

O que é AM

Aprendizado de Máquina 7
Aprendizado de Representação
Pensando problemas de PLN como problemas de AM

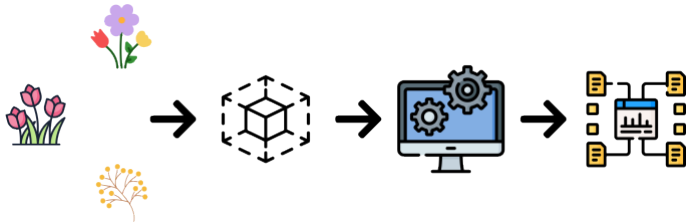


Figura: Fluxo do Aprendizado de Máquina

O que se aprende no Aprendizado de Máquina?

Aprendizado de Máquina 8
Aprendizado de Representação
Pensando problemas de PLN como problemas de AM

Sabendo que dados são sempre codificações (e portanto interpretações) da realidade, é preciso considerar que:

- Algoritmos de AM tem por finalidade gerar uma descrição do **espaço de representação** em conformidade com as observações (i.e. os dados da amostra)

O que se aprende no Aprendizado de Máquina?

Aprendizado de Máquina 8

Aprendizado de Representação

Pensando problemas de PLN como problemas de AM

Sabendo que dados são sempre codificações (e portanto interpretações) da realidade, é preciso considerar que:

- Algoritmos de AM tem por finalidade gerar uma descrição do **espaço de representação** em conformidade com as observações (i.e. os dados da amostra)
- Um modelo é, portanto, determinado pelo **viés de representação** do fenômeno e pelo **viés de aprendizado** do algoritmo utilizado.

O que se aprende no Aprendizado de Máquina?

Sabendo que dados são sempre codificações (e portanto interpretações) da realidade, é preciso considerar que:

- Algoritmos de AM tem por finalidade gerar uma descrição do **espaço de representação** em conformidade com as observações (i.e. os dados da amostra)
- Um modelo é, portanto, determinado pelo **viés de representação** do fenômeno e pelo **viés de aprendizado** do algoritmo utilizado.
- Um modelo aprendido de dados é, então, nada mais que uma função $f : Rep \mapsto Rot$ entre o espaço de representação e o espaço de rótulos.

Uma tipologia de métodos em AM

Os métodos de AM podem ser separados nas seguintes classes:

- Aprendizado supervisionado: o método recebe dados marcados com o resultado esperado da função-objetivo a ser aprendida;

Uma tipologia de métodos em AM

Os métodos de AM podem ser separados nas seguintes classes:

- Aprendizado supervisionado: o método recebe dados marcados com o resultado esperado da função-objetivo a ser aprendida;
- Aprendizado não supervisionado: o método recebe dados sem qualquer marcação da função-objetivo;

Uma tipologia de métodos em AM

Os métodos de AM podem ser separados nas seguintes classes:

- Aprendizado supervisionado: o método recebe dados marcados com o resultado esperado da função-objetivo a ser aprendida;
- Aprendizado não supervisionado: o método recebe dados sem qualquer marcação da função-objetivo;
- Aprendizado semi-supervisionada: o método recebe dados em que parte está marcada e parte não;

Uma tipologia de métodos em AM

Os métodos de AM podem ser separados nas seguintes classes:

- Aprendizado supervisionado: o método recebe dados marcados com o resultado esperado da função-objetivo a ser aprendida;
- Aprendizado não supervisionado: o método recebe dados sem qualquer marcação da função-objetivo;
- Aprendizado semi-supervisionada: o método recebe dados em que parte está marcada e parte não;
- Aprendizado por instrução: o de supervisão fraca, em que o modelo receberá uma avaliação relativa a uma ou mais decisões tomadas durante sua execução;

Aprendizagem Supervisionada: Classificação e Regressão

Aprendizado de Máquina 10
Aprendizado de Representação
Pensando problemas de PLN como problemas de AM

- Classificação: dado uma sequência de pontos $X = \langle v_1, v_2, \cdots, v_n \rangle$ de um espaço $\mathbb{V}$ e uma sequência de marcações $y = \langle c_1, c_2, \cdots, c_n \rangle$ de um conjunto discreto de classes C, determinar uma função em $\mathbb{V}$ que separe *otimamente* os exemplos de classes diferentes.

Aprendizagem Supervisionada: Classificação e Regressão

Aprendizado de Máquina 10
Aprendizado de Representação
Pensando problemas de PLN como problemas de AM

- Classificação: dado uma sequência de pontos $X = \langle v_1, v_2, \cdots, v_n \rangle$ de um espaço $\mathbb{V}$ e uma sequência de marcações $y = \langle c_1, c_2, \cdots, c_n \rangle$ de um conjunto discreto de classes C, determinar uma função em $\mathbb{V}$ que separe *otimamente* os exemplos de classes diferentes.Normalmente $\mathbb{V}$ é um espaço vetorial e a função aprendida determina superficies sobre $\mathbb{V}$ estabelecendo as bordas das clases;

Aprendizagem Supervisionada: Classificação e Regressão

- Classificação: dado uma sequência de pontos $X = \langle v_1, v_2, \cdots, v_n \rangle$ de um espaço $\mathbb{V}$ e uma sequência de marcações $y = \langle c_1, c_2, \cdots, c_n \rangle$ de um conjunto discreto de classes C, determinar uma função em $\mathbb{V}$ que separe *otimamente* os exemplos de classes diferentes.Normalmente $\mathbb{V}$ é um espaço vetorial e a função aprendida determina superficies sobre $\mathbb{V}$ estabelecendo as bordas das clases;
- Regressão: dado um conjunto de pontos $X = \langle v_1, v_2, \cdots, v_n \rangle$ de um espaço $\mathbb{V}$ e um conjunto de valores contínuos $y = \langle r_1, r_2, \cdots, r_n \rangle$, determinar uma função $f : \mathbb{V} \rightarrow \mathbb{R}$ que aproxime *otimamente* os pontos conhecidos da função;

Aprendizagem Supervisionada: Classificação e Regressão

Aspectos importantes a serem considerados:

- Grau de generalização: navalha de Ockham e viés de treinamento

Aprendizagem Supervisionada: Classificação e Regressão

Aspectos importantes a serem considerados:

- Grau de generalização: navalha de Ockham e viés de treinamento
- Os dados podem ser ruidosos: acurácia e sobreajuste (*overfitting*)

Aprendizado de Máquina
Aprendizado de Representação 12
Pensando problemas de PLN como problemas de AM

Aprendizado de Representação

Conjunto de Características

Aprendizado de Máquina
Aprendizado de Representação 13
Pensando problemas de PLN como problemas de AM

O conjunto de características define o espaço de representação dos dados que guiará o método de aprendizagem. As características devem

- Ser descritivas e relevantes ao fenômeno discutido

Conjunto de Características

O conjunto de características define o espaço de representação dos dados que guiará o método de aprendizagem. As características devem

- Ser descritivas e relevantes ao fenômeno discutido
- Ser suficientemente descritivas para se conseguir identificar o fenômeno

Conjunto de Características

O conjunto de características define o espaço de representação dos dados que guiará o método de aprendizagem. As características devem

- Ser descritivas e relevantes ao fenômeno discutido
- Ser suficientemente descritivas para se conseguir identificar o fenômeno

NÂO É FÁCIL DEFINIR UM BOM CONJUNTO DE CARACTERÍSTICAS $\Rightarrow$ Engenharia de features

Representações também podem ser aprendidas

Aprendizado de Máquina
Aprendizado de Representação 14
Pensando problemas de PLN como problemas de AM

Explorando a estrutura latente dos dados, podemos aprender melhores representações deles

Representações também podem ser aprendidas

Aprendizado de Máquina
Aprendizado de Representação 14
Pensando problemas de PLN como problemas de AM

Explorando a estrutura latente dos dados, podemos aprender melhores representações deles

- A partir de uma representação inicial (maior compromisso aos dados iniciais)

Representações também podem ser aprendidas

Aprendizado de Máquina
Aprendizado de Representação 14
Pensando problemas de PLN como problemas de AM

Explorando a estrutura latente dos dados, podemos aprender melhores representações deles

- A partir de uma representação inicial (maior compromisso aos dados iniciais)
- Utilizar os dados anotados para aprender um conjunto de features que seja ideal para o problema

Representações também podem ser aprendidas

Explorando a estrutura latente dos dados, podemos aprender melhores representações deles

- A partir de uma representação inicial (maior compromisso aos dados iniciais)
- Utilizar os dados anotados para aprender um conjunto de features que seja ideal para o problema
- Features aprendidas podem ser indicativas das particularidades do corpus
- Features aprendidas não são claramente compreendidas.

Métodos de Aprendizagem de Representações

Aprendizado de Máquina
Aprendizado de Representação 15
Pensando problemas de PLN como problemas de AM

Aprendizagem de Representações é um tópico antigo, mas que ganhou renovada importância no contexto da aplicação de redes neurais ao processamento de textos, uma vez que redes neurais possuem vieses de representação flexíveis (dependentes das funções de ativação, largura e profundidade da rede).

Métodos de Aprendizagem de Representações

Aprendizado de Máquina
Aprendizado de Representação 15
Pensando problemas de PLN como problemas de AM

Aprendizagem de Representações é um tópico antigo, mas que ganhou renovada importância no contexto da aplicação de redes neurais ao processamento de textos, uma vez que redes neurais possuem vieses de representação flexíveis (dependentes das funções de ativação, largura e profundidade da rede).
Exemplos:

- PCA - Análise de Componentes Principais

Métodos de Aprendizagem de Representações

Aprendizagem de Representações é um tópico antigo, mas que ganhou renovada importância no contexto da aplicação de redes neurais ao processamento de textos, uma vez que redes neurais possuem vieses de representação flexíveis (dependentes das funções de ativação, largura e profundidade da rede).
Exemplos:

- PCA - Análise de Componentes Principais
- LDA - Análise Latente de Dirichilet

Métodos de Aprendizagem de Representações

Aprendizado de Máquina
Aprendizado de Representação 15
Pensando problemas de PLN como problemas de AM

Aprendizagem de Representações é um tópico antigo, mas que ganhou renovada importância no contexto da aplicação de redes neurais ao processamento de textos, uma vez que redes neurais possuem vieses de representação flexíveis (dependentes das funções de ativação, largura e profundidade da rede).
Exemplos:

- PCA - Análise de Componentes Principais
- LDA - Análise Latente de Dirichilet
- Word Embeddings

Métodos de Aprendizagem de Representações

Aprendizagem de Representações é um tópico antigo, mas que ganhou renovada importância no contexto da aplicação de redes neurais ao processamento de textos, uma vez que redes neurais possuem vieses de representação flexíveis (dependentes das funções de ativação, largura e profundidade da rede).
Exemplos:

- PCA - Análise de Componentes Principais
- LDA - Análise Latente de Dirichilet
- Word Embeddings
- Modelos de Linguagem e Arquitetura Encoder-Decoder

Word Embeddings

Aprendizado de Máquina
Aprendizado de Representação 16
Pensando problemas de PLN como problemas de AM

Word Embeddings são modelos de representação semântica de palavras (i.e. um modelo de semântica lexical) em que um vocabulário é mapeado em um espaço vetorial de representação aprendido automaticamente.

Word Embeddings

Word Embeddings são modelos de representação semântica de palavras (i.e. um modelo de semântica lexical) em que um vocabulário é mapeado em um espaço vetorial de representação aprendido automaticamente.
Espera-se que a geometria do espaço de representação codifique aspectos semânticos da língua.

Word Embeddings

Word Embeddings são modelos de representação semântica de palavras (i.e. um modelo de semântica lexical) em que um vocabulário é mapeado em um espaço vetorial de representação aprendido automaticamente.
Espera-se que a geometria do espaço de representação codifique aspectos semânticos da língua.Assim, palavras são mapeadas a vetores nesse espaço e direções no espaço a traços semânticos.

Word Embeddings

Word Embeddings são modelos de representação semântica de palavras (i.e. um modelo de semântica lexical) em que um vocabulário é mapeado em um espaço vetorial de representação aprendido automaticamente.
Espera-se que a geometria do espaço de representação codifique aspectos semânticos da língua.Assim, palavras são mapeadas a vetores nesse espaço e direções no espaço a traços semânticos.

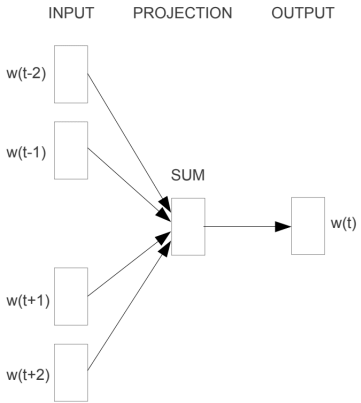
- O significado de uma palavra está em seu uso (Wittgenstein, 1953)

Word Embeddings

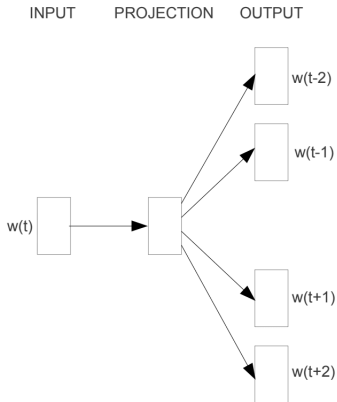
Word Embeddings são modelos de representação semântica de palavras (i.e. um modelo de semântica lexical) em que um vocabulário é mapeado em um espaço vetorial de representação aprendido automaticamente.
Espera-se que a geometria do espaço de representação codifique aspectos semânticos da língua.Assim, palavras são mapeadas a vetores nesse espaço e direções no espaço a traços semânticos.

- O significado de uma palavra está em seu uso (Wittgenstein, 1953)
- Se A e B são usadas em contextos quase idênticos, possuem significados similares (Harris, 1954)

Word2Vec



CBOW



Skip-gram

Aprendizado de Máquina

Aprendizado de Representação

Pensando problemas de PLN como problemas de AM 18

Pensando problemas de PLN como problemas de AM

Como usar AM em PLN?

- Definição do problema como um problema de classificação/regressão

Como usar AM em PLN?

- Definição do problema como um problema de classificação/regressão
- Descrição dos dados: conjunto de características (*features*)

Como usar AM em PLN?

- Definição do problema como um problema de classificação/regressão
- Descrição dos dados: conjunto de características (*features*)
- Definição do método de treino/viés