

cap1

March 18, 2024

Aprendizagem de Máquinas

Exercício 1 - Lista 1

Por Thiago Barros

Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow

1) Como você definiria o Aprendizado de Máquina?

R: O aprendizado de máquina é a ciência de ensinar as máquinas a resolverem alguns tipos de problemas por meio de dados

2) Você saberia nomear quatro tipos de problemas nos quais o AM se destaca?

R: Problemas com alto grau de complexidade para meios tradicionais. Problemas onde existem mudanças constantes no conjunto de informações (Flutuação dos Dados), Problemas onde existem mudanças nas estruturas das regras para resolução (Conforme o input muda, as regras de resolução mudam, ou seja, novos algoritmos) Problemas que exigem grande volume de dados.

3) O que é um conjunto de treinamento rotulado?

Um conjunto de dados rotulado é aquela cuja a instância é classificada por meio de um rótulo que será usado para prever novas instâncias. Ex: Um conjunto de dados referentes a emails com classificação de spam ou não spam.

4) Quais são as duas tarefas supervisionadas mais comuns?

R: Classificação Regressão

5) Você consegue nomear quatro tarefas comuns sem supervisão?

R:

clustering, visualização, redução da dimensionalidade e aprendizado de regras de associação.

6) Que tipo de algoritmo de Aprendizado de Máquina você utilizaria para permitir que um robô ande em vários terrenos desconhecidos?

R: Aprendizado de Reforço

7) Que tipo de algoritmo você utilizaria para segmentar seus clientes em vários grupos?

R: Clusterização

8. Você enquadraria o problema da detecção de spam como um problema de aprendizado supervisionado ou sem supervisão?

R: se utilizar uma base de dados para treinamento do algoritmo e outra com instâncias para validar o algoritmo, têm-se então que classifica como aprendizado supervisionado, porém, é possível integrar ambos os paradigmas a depender da necessidade.

9. O que é um sistema de aprendizado online?

O sistema de aprendizado online é aquele cujo o processo é incremental e se alimenta sequencialmente as instâncias de dados individualmente ou com pequenos grupos de dados.

10. O que é o out-of-core learning?

É o tipo de aprendizado online onde os algoritmos carregam partes dos dados, executam uma etapa do treinamento nesses dados, repete o processo até que toda a base de dados tenha sido avaliada para treinamento.

11. Que tipo de algoritmo de aprendizado depende de uma medida de similaridade para fazer previsões?

O algoritmo de aprendizado baseado em instância onde o sistema aprende com exemplos por meio de memorização e em seguida generaliza novos casos utilizando uma medida de similaridade. (Ex: o algoritmo usa a similaridade entre emails spam usando uma métrica como número de palavras ou caracteres)

12. Qual a diferença entre um parâmetro de modelo e o hiperparâmetro do algoritmo de aprendizado?

Um parâmetro do modelo é um valor que faz parte do modelo, como por exemplo, o coeficiente angular. O hiperparâmetro é um parâmetro aplicado ao algoritmo de serve para ajustar o aprendizado do algoritmo.

13. O que os algoritmos de aprendizado baseados em modelos procuram? Qual é a estratégia mais comum que eles utilizam para ter sucesso? Como fazem previsões?

O algoritmo de aprendizado baseado em modelos busca obter generalizações além da base de dados estabelecida para a definição do modelo. O modelo tenta aproximar o comportamento de um conjunto de dados que estejam relativamente correlacionados para descrever esse comportamento matematicamente e relaciona-los. A estratégia mais comum consiste em usar uma base de dados separando-a com um conjunto de dados para treinamento, criar o modelo com base na base de treinamento e avaliar o resultado do modelo com a porção de validação.

14. Você pode citar quatro dos principais desafios no Aprendizado de Máquina?

R: 1 - Lidar com Dados Viesados 2 - Lidar com Dados Insuficientes 3 - Lidar com Dados Não representativos 4 - Lidar com Problemas de Generalização (Sobreajuste ou Subajuste)

15. Se o seu modelo é ótimo nos dados de treinamento, mas generaliza mal para novas instâncias, o que está acontecendo? Você pode nomear três soluções possíveis?

O que está acontecendo é que o modelo não apresenta capacidade para fazer predição com valores fora do conjunto de treinamento. Isso é chamado de overfitting ou sobreajuste. Ou se o modelo for simples demais para o processo de explicação do fenômeno real. Existem algumas maneiras de melhorar esse modelo.

Existe algumas maneiras:

Regularizar o modelo: Consiste em restringir os parâmetros do modelo para evitar risco de sobreajuste

Melhorar a base de dados

Testar vários modelos e verificar o que apresenta menor erro de predição

16. O que é um conjunto de testes e por que você o utilizaria?

Dentro do conjunto de dados para criação e treinamento, existe a parcela que será usada para o processo de aprendizagem. O conjunto de testes seria aquele conjunto que modelos definidos serão avaliados para verificar o percentual de erro entre o valor predito e o valor estimado.

17. Qual é o propósito de um conjunto de validação?

O conjunto de validação é aquele que será usado para avaliar o modelo e algoritmo definido como final para aferir a capacidade do modelo ser preditor.

18. O que pode dar errado se você ajustar os hiperparâmetros utilizando o conjunto de teste?

O modelo cai no mesmo problema do sobreajuste e acaba se limitando ao conjunto de teste.

19. O que é validação cruzada e por que você preferiria usá-la ao invés de um conjunto de validação?

Na validação cruzada, o conjunto de treinamento é dividido em subconjuntos complementares e cada modelo é treinado por uma combinação diferente desses conjuntos e validado em relação as partes restantes.

Definido o modelo e hiperparâmetros ideais, um modelo final é treinado com a utilização desses hiperparâmetros no conjunto completo de treinamento e o erro generalizado é medido no conjunto de testes.