Lista de Exercícios 01

(prazo final para entrega: até 23:59 de 25/02/2019 - segunda-feira)

Aprendizado de Máquina
Aluno(a): Marianna de Pinho Severo
Matrícula: 374856

1. **O que é Aprendizado de Máquina e para que serve?**
   O aprendizado de máquina é uma sub-área da inteligência artificial que se preocupa com o desenvolvimento de algoritmos que conseguem aprender sozinhos, a partir de uma grande quantidade de dados, da aplicação de modelos matemáticos e de técnicas de ciência da computação. Ele possui aplicação em inúmeras áreas, como economia, robótica, descoberta de doenças, etc. e novas aplicações têm sido criadas a cada dia. Os algoritmos desenvolvidos são capazes de encontrar e aprender padrões em conjuntos de dados, realizar previsões, classificar dados e muitas outras atividades.

2. **O que são dados rotulados (labels) e para que servem?**
   Quando um conjunto de dados é recebido como entrada de um algoritmo e o objetivo é descobrir a saída para cada um desses dados, dizemos que dados rotulados são aqueles cujas saídas (ou destinos) já conhecemos. Geralmente, deles são escolhidos conjuntos de atributos que são utilizados na identificação de padrões que conduzem ao aprendizado. Esses dados são usados para o treinamento de algoritmos de aprendizado supervisionado.

3. **Quais os problemas mais comuns de aprendizado supervisionado?**
   Os problemas mais comuns são de Classificação e de Regressão.

4. **Quais os problemas mais comuns de aprendizado não-supervisionado?**
   Os problemas mais comuns são de Clusterização, de Associação, de Visualização e de Redução de dimensionalidade.

5. **Que tipo de algoritmo de Machine Learning (ML) você usaria para:**

   a) Permitir um robô andar em diversos tipos de terreno?
   Utilizaria aprendizado supervisionado, pois o robô poderia ser treinado para andar em tipos de terrenos pré-determinados, cujos dados de entrada e saída já seriam determinados.

   b) Segmentar clientes em múltiplos grupos?
   Utilizaria aprendizado não supervisionado, pois ele poderia encontrar semelhanças não conhecidas entre conjuntos de clientes, organizando-os em grupos de acordo com essas semelhanças.

6. **O que é um sistema de aprendizado online? Quais suas vantagens e desvantagens?**

É um sistema em que o algoritmo está sempre aprendendo, conforme interage com novos conjuntos de dados, ou seja, ele não se limita ao aprendizado obtido durante o treinamento. Algumas de suas vantagens consistem em eles serem capazes de utilizar os conhecimentos adquiridos a qualquer momento para processamento de novos dados; devido à maneira como são construídos, eles conseguem manter um baixo e constante tempo requerido para o processamento de novos dados; eles utilizam memória de maneira mais eficiente; e eles possuem uma maior capacidade de se adequarem a mudanças. Como algumas desvantagens, estão a complexidade de implementação e a maior quantidade de processamento a ser realizada, porque eles estão frequentemente aprendendo. Nem todo modelo funciona bem fazendo isso, de se ajustar frequentemente.

7. **O que significa aprendizado *out-of-core*?**
   O aprendizado *out-of-core* consiste em uma série de algoritmos que conseguem aprender a partir de conjuntos de dados que precisam ser passados aos poucos, pois não cabem completamente na memória do computador. Assim, o algoritmo não precisa processar todos os dados de uma vez, mas pode ir aprendendo conforme consome grupos desses dados.

8. **Que tipo de aprendizado usa medidas de similaridade para fazer predições?**
   O aprendizado não supervisionado, usando a técnica de clusterização.

9. **Em um modelo de ML, qual a diferença entre parâmetros e hiper-parâmetros?**
   Parâmetros são configurações do modelos cujos valores podem ser obtidos a partir dos dados, sendo configurações internas ao modelo. Seus valores, geralmente, não são atribuídos manualmente.

   Os hiperparâmetros, por outro lado, são configurações externas ao modelo. Eles surgem do fato de, algumas vezes, não ser possível obter os valores dos parâmetros a partir dos dados, então os valores são atribuídos de maneira manual aos hiperparâmetros, por meio de inúmeras técnicas, como tentativa e erro, heurísticas, entre outras. Eles são, inclusive, utilizados para o ajuste dos parâmetros.

10. **O que são e qual a diferença entre *Underfitting* e *Overfitting?* O que fazer para solucionar cada um desses problemas?**
    Overfitting e *Underfitting* são problemas em algoritmos de aprendizado de máquina que diminuem o seu desempenho. O *overfitting* ocorre quando o modelo se torna tão complexo, considerando tantos parâmetros, que ele leva em consideração até os ruídos presentes nos dados. Assim, dizemos que o modelo apresenta um ótimo desempenho para os dados de treinamento, mas não generaliza bem para os dados de teste, apresentando elevada variância. O *underfitting*, por sua vez, ocorre quando o modelo não é complexo o suficiente para capturar o padrão nos dados de treinamento e, consequentemente, apresenta desempenho ruim para os dados de teste. Dizemos que ele possui uma elevada *bias*.

    Para lidar como *overfitting*, duas principais técnicas podem ser utilizadas: a reamostragem, para estimar a acurácia do modelo, e a utilização de um conjunto de

validação. Já para o *underfitting*, uma das formas de lidar é a utilização de regressão polinomial.

11. Para que servem: conjunto de treino, conjunto de validação e conjunto de teste?

O conjunto de treino é utilizado para a construção (treinamento) do modelo, fornecendo os dados a partir dos quais o modelo aprenderá padrões a serem empregados em novos conjuntos de dados. O conjunto de validação fornece dados para serem testados ainda durante o treinamento, ajudando no ajuste dos parâmetros e, consequentemente, no aprendizado. Por fim, o conjunto de teste é utilizado para testar o modelo final, fornecendo dados ainda não vistos pelo modelo.

12. O que significa validação cruzada e qual a vantagem de usá-la ao invés de usar um conjunto de validação?

Existem algumas técnicas de validação cruzada e, entre elas, uma das mais conhecidas é a k-fold. A validação cruzada consiste na divisão do conjunto de dados em vários subconjuntos; um desses subconjuntos é utilizado para validação do modelo, enquanto o restante é empregado para treinamento; e esse processo se repete até que todos os subconjuntos tenham sido usados para validação.

Diversas vezes, o conjunto de dados para treinamento é escasso. Então, poder utilizar todo o conjunto de dados para treinar, mesmo que em estágios diferentes do treinamento, permite um melhor aproveitamento do que reduzir ainda mais o subconjunto de treinamento, como ocorre na validação.