

Primeira Aula da Disciplina Aprendizado de Máquina e Mineração de Dados

Nesta disciplina iremos discutir métodos e algoritmos que permitem que sistemas computacionais “aprendam” a resolver um determinado problema com base em experiências passadas. Assim, pode-se notar que os algoritmos construídos na área de aprendizado de máquina não são projetados para resolver um problema que pode ser resolvido de maneira exata (através de algoritmos específicos). Ao invés disso, os algoritmos de aprendizado de máquina são desenvolvidos para identificar padrões presentes em experiências passadas (onde o problema em questão foi resolvido com sucesso) e, a utilizar estes padrões para solucionar novas instâncias do mesmo problema.

Veremos que as experiências passadas podem ser representadas de várias maneiras, sendo que quando são representadas com base em bancos de dados (uma das representações mais comuns) os métodos e algoritmos são chamados de métodos e algoritmos de “mineração de dados”. Assim, pode-se dizer que a mineração de dados é uma forma de aplicação do aprendizado de máquina, na qual, as experiências passadas estão representadas em forma de bancos de dados.

Nesta primeira aula, serão realizados exercícios que permitam-nos fazer uma analogia sobre a forma como os seres humanos “aprendem” a resolver problemas (com base em experiências passadas) e a forma como as máquinas realizam tal aprendizado. Assim, considere que após uma aula de “computação avançada” um colega lhe pergunta se você pode ajuda-lo a escolher qual lente de contato é a mais indicada para ele começar a usar. A sua primeira resposta deve ser: procure um médico, pois não se pode prescrever lentes de contato sem a devida autoridade e conhecimento oftalmológico. Suponha, entretanto, que após seu colega ter sido examinado por um médico especialista, vocês dois tentem (só por brincadeira) identificar (de maneira simplista e superficial) como seria possível identificar qual lente de contato seria adequada para casos aleatórios de pacientes em um consultório oftalmológico. Neste sentido, dando sequência à ideia de prescrição de lentes de contato, vocês encontram na internet um banco de dados contendo os dados de prontuários de 23 pacientes. Os dados são simplificados e para cada um dos 23 pacientes há informações sobre a identificação do paciente (ID), a idade (categorizada em termos oftalmológicos: jovem, pré-presbiópico e prebiópico), o diagnóstico oficial (miópe ou hipermetrope), a presença de astigmatismo (sim ou não), a produção lacrimal do paciente (normal ou reduzida), e por fim, qual foi a lente prescrita (nenhuma, gelatinosa ou rígida).

ID1, jovem, miópe, não, reduzida, nenhuma

ID2, jovem, miópe, não, normal, gelatinosa

ID3, jovem, miópe, sim, reduzida, nenhuma

ID4, jovem, miópe, sim, normal, rígida

ID5, jovem, hipermetrope, não, reduzida, gelatinosa

ID6, jovem, hipermetrope, não, normal, gelatinosa

ID7, jovem, hipermetrope, sim, reduzida, rígida
ID8, jovem, hipermetrope, sim, normal, rígida
ID9, pre-presbiópico, míope, não, reduzida, nenhuma
ID10, pre-presbiópico, míope, não, normal, gelatinosa
ID11, pre-presbiópico, míope, sim, reduzida, nenhuma
ID12, pre-presbiópico, míope, sim, normal, rígida
ID13, pre-presbiópico, hipermetrope, não, reduzida, nenhuma
ID14, pre-presbiópico, hipermetrope, não, normal, gelatinosa
ID15, pre-presbiópico, hipermetrope, sim, reduzida, rígida
ID16, pre-presbiópico, hipermetrope, sim, normal, nenhuma
ID17, presbiópico, míope, não, reduzida, nenhuma
ID18, presbiópico, míope, não, normal, nenhuma
ID19, presbiópico, míope, sim, reduzida, nenhuma
ID20, presbiópico, míope, sim, normal, rígida
ID21, presbiópico, hipermetrope, não, reduzida, nenhuma
ID22, presbiópico, hipermetrope, não, normal, gelatinosa
ID23, presbiópico, hipermetrope, sim, reduzida, nenhuma

Com base nestes dados obtidos na internet, responda:

1) qual lente de contato você prescreveria para os seguintes pacientes (justifique cada uma das suas prescrições):

- a) ID31, pré-presbiópico, hipermetrope, sim, normal, ?
- b) ID32, presbiópico, hipermetrope, sim, normal, ?
- c) ID33, jovem, míope, sim, normal, ?
- d) ID34, presbiópico, hipermetrope, sim, normal, ?

2) qual dos atributos (identificação do paciente (ID), idade, diagnóstico oficial, presença de astigmatismo, produção lacrimal do paciente) é o mais significativos para auxiliar a identificar a melhor lente de contato? Por que? E qual é o menos significativo? Por que?

3) você acha que este banco de dados seria suficiente para que um médico pudesse ser considerado apto para prescrever lentes de contato? Por que? Quais características seriam mais importantes para que se possa considerar um banco de dados “perfeito” para se aprender a resolver um problema (mencione a influência da quantidade de instâncias, quantidade de atributos, o tipo dos atributos (real, inteiro, nominal), etc.)?

4) Você acha que seria possível, em um banco de dados como o mostrado acima, que dois pacientes com as mesmas características apresentassem prescrições diferentes? Por que? Dê exemplos.

5) imagine que o atributo idade fosse um atributo numérico (definindo quantos anos o paciente possui), como você faria para verificar a similaridade entre a idade de dois pacientes? Por que?

5) Considerando que você não é médico, ao invés de tomar como base os padrões presentes em experiências passadas, você teria um algoritmo melhor para prescrever lentes de contato? Justifique sua resposta.

6) Alguns cursos de graduação (o curso de medicina da UFSCar, por exemplo) seguem o processo de “aprendizado baseado em casos”. Supondo que tais cursos não exigissem o estudo teórico (o que não é verdade, pois os alunos precisam também embasar as decisões, em cada caso vivenciado, através de modelos e teorias presentes na literatura). O que você acha deste suposto método? Quantos casos um aluno precisaria vivenciar para se formar?

7) construa um algoritmo (em pseudo-código, ou em alguma linguagem de programação que você tenha familiaridade) capaz de responder a questão 1 desta atividade.

8) construa um algoritmo (em pseudo-código, ou em alguma linguagem de programação que você tenha familiaridade) capaz de responder a questão 2 desta atividade.