Aprendizagem de Regras para Classificação de Patologias da Coluna

O aprendizado de máquina é um ramo da inteligência artificial que explora métodos para aprender a partir de dados. A finalidade da aprendizagem é tomar decisões guiadas pelos dados ao invés de seguir instruções explicitamente programadas. Este trabalho tem como objetivo aprender um conjunto de regras para diagnóstico de patologias da coluna vertebral. Os arquivos deste trabalho contém dados extraídos de pacientes, a partir de radiografias panorâmicas. Uma parte dos dados são de pacientes que não possuem patologias na coluna, chamados de normais. A outra parte dos dados são pacientes com alguma patologia da coluna verteral, como hérnias de disco e espondilolistese. Cada um dos pacientes é descrito pelas informações:

- ângulo de incidência pélvica (PI)
- ângulo de versão pélvica (PT)
- ângulo de lordose (LA)
- inclinação sacral (SS)
- raio pélvico (RP)
- grau de deslizamento (GS)

Esses dados estão disponíveis em vários arquivos. Cada arquivo possui algumas das informações acima de vários pacientes. Para cada uma dessas informações, o arquivo contém várias colunas com intervalos de valores sobre essa informação. Cada coluna dessas, vamos chamar de atributo. Por exemplo, para a informação PI temos as seguintes colunas:

```
PI <= 42.09; PI <= 48.12; PI <= 54.92; PI <= 63.52; PI <= 70.62; PI <= 80.61
```

Os nomes dos arquivos são da forma `column_bin_<f>a_<n>p.csv`, em que `<f>` é a quantidade de atributos dos pacientes e `<n>` é a quantidade de pacientes no arquivo. Abaixo, temos o exemplo do arquivo `column_bin_5a_3p.csv`. A primeira linha do arquivo representa os atributos dos pacientes. Observe que a segunda linha representa um paciente com PI (ângulo de incidência pélvica) maior que 70.62, PI menor ou igual a 80.61, GS (grau de deslizamento) maior que 37.89, GS menor ou igual a 57.55 e com patologia da coluna vertebral (P = 1). Veja que a terceira linha representa um outro paciente com outros valores para os atributos e também com a patologia. Já a última linha representa um outro paciente com outros valores para os atributos e sem patologia (P = 0).

```
PI <= 42.09; PI <= 70.62; PI <= 80.61; GS <= 37.89; GS <= 57.55; P
0          ; 0          ; 1          ; 0          ; 1          ; 1
0          ; 0          ; 0          ; 0          ; 0          ; 1
0          ; 0          ; 0          ; 1          ; 1          ; 0
```

Por exemplo, olhando os dados do arquivo acima, veja que um paciente com PI entre 70.62 e 80.61 pode ser diagnosticado com uma patalogia da coluna vertebral. Também observe que um paciente com GS maior que 57.55 pode ser diagnosticado com patalogia. Essas classificações do diagnóstico podem ser representadas pelo conjunto de regras:

$$\{[PI \leq 80.61, PI > 70.62] \Rightarrow P,$$
$$[GS > 57.55] \Rightarrow P\}$$

O conjunto de regras acima possui duas regras. **Observe atentamente a notação usada para o conjunto de regras**. **É importante considerar que essas regras NÃO são fórmulas da lógica proposicional**. Por exemplo, veja que **o símbolo de aplicação de regra** $\Rightarrow$ não é a implicação $\rightarrow$ das fórmulas da lógica proposicional.

Veja que o primeiro atributo da primeira regra **se aplica** ao primeiro paciente, pois o atributo $PI \leq 80.61$ do primeiro paciente é 1. Ao mesmo tempo, veja que $PI > 70.62$ da primeira regra também se aplica ao primeiro paciente, pois o atributo $PI \leq 70.62$ do primeiro paciente é 0. Veja que como **os dois atributos** da primeira regra **se aplicam ao primeiro paciente**, então **a primeira regra se aplica** ao primeiro paciente. Como **uma das regras se aplica ao primeiro paciente**, então **esse paciente é classificado com a patologia da coluna vertebral**. Nesse caso, também dizemos que **o conjunto de regras se aplica ao primeiro paciente**.

Observe agora que a primeira regra **não** se aplica ao segundo paciente, pois o atributo $PI \leq 80.61$ do segundo paciente é 0. Já a segunda regra se aplica ao segundo paciente, pois o atributo $GS \leq 57.55$ do segundo paciente é 0. Como **alguma das regras se aplica ao segundo paciente,** então dizemos que **o conjunto de regras se aplica ao segundo paciente e que ele é classificado com a patologia**.

Veja também que **nenhuma das regras acima se aplica ao paciente da última linha que não tem patologia**. Observe que o atributo $PI \leq 80.61$ do último paciente é 0, logo, essa regra não se aplica a esse paciente. Além disso, o atributo $GS \leq 57.55$ do último paciente é 1 e, portanto, a segunda regra também não se aplica ao último paciente. Como nenhuma das regras se aplica ao último paciente, **esse paciente é classificado como sem a presença da patologia**.

Para um outro exemplo, veja que o conjunto de regras abaixo com apenas uma regra também se aplica ao dois pacientes com patologia e não se aplica ao paciente sem patologia.

$$\{[GS > 37.89] \Rightarrow P\}$$

Exposto isso, este tema de trabalho compreende o desenvolvimento de uma aplicação para obter um conjunto de regras para classificação de indivíduos em: normal (sem patologia), ou com patologia. O conjunto de regras deve se aplicar à todos os pacientes com a patologia e não pode se aplicar a nenhum dos pacientes sem patologia. **Observe que VOCÊ NÃO VAI CONSTRUIR O CONJUNTO DE REGRAS EXPLICITAMENTE**. Neste projeto, **você vai modelar o problema de aprendizagem de regras com satisfatibilidade da lógica proposicional**. Ou seja, **você vai construir uma fórmula da lógica proposicional que representa as restrições da aprendizagem de regras e, em seguida, vai verificar se essa fórmula é satisfatível**. **O conjunto de regras será construído a partir de uma valoração que satisfaz a fórmula que modela a aprendizagem de regras**. A seguir, vamos entrar em detalhes em **quais são as restrições necessárias** para a aprendizagem de regras e **como definir essas restrições como fórmulas da lógica proposicional**.

Para obter o conjunto de regras, sua aplicação terá como parâmetros um número natural m maior que zero para representar a quantidade de regras e um arquivo `column_bin_<f>a_<n>p.csv`. A partir dessa entrada, sua aplicação deverá determinar se existe um conjunto com m regras para classificar todos os pacientes do arquivo de forma correta. Caso exista, sua aplicação deverá apresentar o conjunto de regras. Caso não exista, sua aplicação deverá indicar essa não existência. **Veja que, a partir de uma entrada m e de um arquivo, temos que construir uma fórmula da lógica proposicional que é satisfatível se e somente se existir um conjunto com m regras capaz de classificar os pacientes**. Portanto, se a fórmula for **insatisfatível**, então **não** existe um conjunto com m regras que classifique corretamente todos os pacientes. Além disso, **se a fórmula for satisfatível, então a valoração que satisfaz a fórmula deve ter as informações necessárias para construir o conjunto de regras**.

Dessa forma, você deve modelar o problema usando satisfatibilidade da lógica proposicional. Para ajudar na modelagem, você pode usar as dicas a seguir. Use variáveis atômicas $x_{a,i,le}$, $x_{a,i,gt}$, $x_{a,i,s}$ e $c_{i,j}$ tal que a é um atributo, i é um natural com $1 \leq i \leq m$ representando a i-ésima regra e j é um natural com $1 \leq j \leq n$ em que n é a quantidade de pacientes. Por exemplo, $x_{GS\leq 37.89,2,le}$, $x_{GS\leq 37.89,2,gt}$, $x_{GS\leq 37.89,3,s}$, $x_{PI\leq 42.09,3,gt}$, $x_{PI\leq 70.62,1,le}$, $c_{1,20}$ e $c_{3,11}$ são exemplos de atômicas para o parâmetro $m = 3$. **ATENÇÃO: perceba que $x_{GS\leq 37.89,2,le}$ e $GS \leq 37.89$ são coisas distintas. O primeiro é uma fórmula atômica da lógica proposicional e o segundo é um atributo do conjunto de dados**.

Na sua modelagem, as restrições representadas pela fórmula que será construída forçarão que as atômicas devam ter os seguintes significados:

- $v(x_{a,i,le}) = T$ se e somente se o atributo a ocorre com $\leq$ na regra i.

- $v(x_{a,i,gt}) = T$ se e somente se o atributo a ocorre com $>$ na regra i.
- $v(x_{a,i,s}) = T$ exatamente quando o atributo a não aparece na regra i.
- $v(c_{i,j}) = T$ exatamente quando a regra i cobre o paciente j (ou seja, a regra i se aplica ao paciente j).

Observe que se $v(x_{GS\leq 37.89,2,le}) = F$, $v(x_{GS\leq 37.89,2,gt}) = T$ e $v(x_{GS\leq 37.89,2,s}) = F$, então a regra 2 das m regras possui $GS > 37.89$ como elemento da regra.
A partir de um arquivo `column_bin_<f>a_<n>p.csv` e do natural m, você deve construir restrições da aprendizagem de regras usando uma fórmula da lógica proposicional com as atômicas definidas acima. Você pode construir sua fórmula a partir das restrições descritas em linguagem natural a seguir. **Observe que essas restrições vão especificar propriedades que o conjunto de regras que queremos obter deve cumprir**.

1. Para cada atributo e cada regra, temos **exatamente** uma das três possibilidades: o atributo aparece com $\leq$ na regra, o atributo aparece com $>$ na regra, ou o atributo não aparece na regra.
2. Cada regra deve ter algum atributo aparecendo nela.
3. Para cada paciente sem patologia e cada regra, algum atributo do paciente não pode ser aplicado à regra.
4. Para cada paciente com patologia, cada regra e cada atributo, se o atributo do paciente não se aplicar ao da regra, então a regra não cobre esse paciente.
5. Cada paciente com patologia deve ser coberto por alguma das regras.

Para entender melhor a modelagem do prolema com as restrições acima, vamos usar um exemplo. **Vamos omitir do exemplo apenas a restrição 1 para que você pense como essa restrição deve ser representada como uma fórmula da lógica proposicional**. Para o nosso exemplo, vamos assumir um arquivo que só tenha três pacientes ($n = 3$) e cinco atributos como definido abaixo:

```
PI <= 42.09; PI <= 70.62; PI <= 80.61; GS <= 37.89; GS <= 57.55; P
0          ; 0          ; 1          ; 0          ; 1          ; 1
0          ; 0          ; 0          ; 0          ; 0          ; 1
0          ; 0          ; 0          ; 1          ; 1          ; 0
```

Além disso, vamos supor que o parâmetro m seja quatro ($m = 4$), ou seja, que estamos verificando se é possível classificar corretamente todos os pacientes do conjunto de dados com quatro regras.

Dessa forma, a restrição 2 fica da seguinte maneira:

$$(\neg x_{PI\leq 42.09,1,s} \vee \neg x_{PI\leq 70.62,1,s} \vee \neg x_{PI\leq 80.61,1,s} \vee \neg x_{GS\leq 37.89,1,s} \vee \neg x_{GS\leq 57.55,1,s})$$
$$(\neg x_{PI\leq 42.09,2,s} \vee \neg x_{PI\leq 70.62,2,s} \vee \neg x_{PI\leq 80.61,2,s} \vee \neg x_{GS\leq 37.89,2,s} \vee \neg x_{GS\leq 57.55,2,s})$$
$$(\neg x_{PI\leq 42.09,3,s} \vee \neg x_{PI\leq 70.62,3,s} \vee \neg x_{PI\leq 80.61,3,s} \vee \neg x_{GS\leq 37.89,3,s} \vee \neg x_{GS\leq 57.55,3,s})$$
$$(\neg x_{PI\leq 42.09,4,s} \vee \neg x_{PI\leq 70.62,4,s} \vee \neg x_{PI\leq 80.61,4,s} \vee \neg x_{GS\leq 37.89,4,s} \vee \neg x_{GS\leq 57.55,4,s})$$

A restrição 3 fica:

$$(x_{PI\leq 42.09,1,le} \vee x_{PI\leq 70.62,1,le} \vee x_{PI\leq 80.61,1,le} \vee x_{GS\leq 37.89,1,gt} \vee x_{GS\leq 57.55,1,gt})$$
$$(x_{PI\leq 42.09,2,le} \vee x_{PI\leq 70.62,2,le} \vee x_{PI\leq 80.61,2,le} \vee x_{GS\leq 37.89,2,gt} \vee x_{GS\leq 57.55,2,le})$$
$$(x_{PI\leq 42.09,3,le} \vee x_{PI\leq 70.62,3,le} \vee x_{PI\leq 80.61,3,le} \vee x_{GS\leq 37.89,3,gt} \vee x_{GS\leq 57.55,3,gt})$$
$$(x_{PI\leq 42.09,4,le} \vee x_{PI\leq 70.62,4,le} \vee x_{PI\leq 80.61,4,le} \vee x_{GS\leq 37.89,4,gt} \vee x_{GS\leq 57.55,4,gt})$$

A restrição 4 fica:

$(x_{PI\leq 42.09,1,le} \rightarrow \neg c_{1,1}) \wedge (x_{PI\leq 70.62,1,le} \rightarrow \neg c_{1,1}) \wedge (x_{PI\leq 80.61,1,gt} \rightarrow \neg c_{1,1}) \wedge (x_{GS\leq 37.89,1,le} \rightarrow \neg c_{1,1}) \wedge (x_{GS\leq 57.55,1,gt} \rightarrow \neg c_{1,1})$
$(x_{PI\leq 42.09,1,le} \rightarrow \neg c_{1,2}) \wedge (x_{PI\leq 70.62,1,le} \rightarrow \neg c_{1,2}) \wedge (x_{PI\leq 80.61,1,le} \rightarrow \neg c_{1,2}) \wedge (x_{GS\leq 37.89,1,le} \rightarrow \neg c_{1,2}) \wedge (x_{GS\leq 57.55,1,le} \rightarrow \neg c_{1,2})$
$(x_{PI\leq 42.09,2,le} \rightarrow \neg c_{2,1}) \wedge (x_{PI\leq 70.62,2,le} \rightarrow \neg c_{2,1}) \wedge (x_{PI\leq 80.61,2,gt} \rightarrow \neg c_{2,1}) \wedge (x_{GS\leq 37.89,2,le} \rightarrow \neg c_{2,1}) \wedge (x_{GS\leq 57.55,2,gt} \rightarrow \neg c_{2,1})$
$(x_{PI\leq 42.09,2,le} \rightarrow \neg c_{2,2}) \wedge (x_{PI\leq 70.62,2,le} \rightarrow \neg c_{2,2}) \wedge (x_{PI\leq 80.61,2,le} \rightarrow \neg c_{2,2}) \wedge (x_{GS\leq 37.89,2,le} \rightarrow \neg c_{2,2}) \wedge (x_{GS\leq 57.55,2,le} \rightarrow \neg c_{2,2})$
$(x_{PI\leq 42.09,3,le} \rightarrow \neg c_{3,1}) \wedge (x_{PI\leq 70.62,3,le} \rightarrow \neg c_{3,1}) \wedge (x_{PI\leq 80.61,3,gt} \rightarrow \neg c_{3,1}) \wedge (x_{GS\leq 37.89,3,le} \rightarrow \neg c_{3,1}) \wedge (x_{GS\leq 57.55,3,gt} \rightarrow \neg c_{3,1})$
$(x_{PI\leq 42.09,3,le} \rightarrow \neg c_{3,2}) \wedge (x_{PI\leq 70.62,3,le} \rightarrow \neg c_{3,2}) \wedge (x_{PI\leq 80.61,3,le} \rightarrow \neg c_{3,2}) \wedge (x_{GS\leq 37.89,3,le} \rightarrow \neg c_{3,2}) \wedge (x_{GS\leq 57.55,3,le} \rightarrow \neg c_{3,2})$
$(x_{PI\leq 42.09,4,le} \rightarrow \neg c_{4,1}) \wedge (x_{PI\leq 70.62,4,le} \rightarrow \neg c_{4,1}) \wedge (x_{PI\leq 80.61,4,gt} \rightarrow \neg c_{4,1}) \wedge (x_{GS\leq 37.89,4,le} \rightarrow \neg c_{4,1}) \wedge (x_{GS\leq 57.55,4,gt} \rightarrow \neg c_{4,1})$
$(x_{PI\leq 42.09,4,le} \rightarrow \neg c_{4,2}) \wedge (x_{PI\leq 70.62,4,le} \rightarrow \neg c_{4,2}) \wedge (x_{PI\leq 80.61,4,le} \rightarrow \neg c_{4,2}) \wedge (x_{GS\leq 37.89,4,le} \rightarrow \neg c_{4,2}) \wedge (x_{GS\leq 57.55,4,le} \rightarrow \neg c_{4,2})$

E a restrição 5 fica da seguinte forma:

$$(c_{1,1} \vee c_{2,1} \vee c_{3,1} \vee c_{4,1})$$
$$(c_{1,2} \vee c_{2,2} \vee c_{3,2} \vee c_{4,2})$$

A partir dessas restrições e das demais, você deve verificar se a conjunção dessas fórmulas é satisfatível. Se não for satisfatível, então não existe um conjunto de 4 regras para classificar corretamente todos os pacientes. Caso seja satisfatível, então existe um conjunto de 4 regras para classificar todos os pacientes de forma correta. E, a partir da valoração que satisfaz a fórmula, é possível obter esse conjunto de regras.

Para mostrar como obter o conjunto de regras a partir da valoração que satisfaz a fórmula, vamos considerar o exemplo abaixo com apenas três atributos, três pacientes e duas regras ($m = 2$).

```
PI <= 42.09; PI <= 70.62; GS <= 37.89; P
0          ; 0          ; 0          ; 1
0          ; 1          ; 0          ; 1
0          ; 0          ; 1          ; 0
```

Imagine que a fórmula que representa as restrições da aprendizagem de regras tenha sido construída e que a seguinte valoração deixa essa fórmula verdadeira:

$\{(x_{PI\leq 42.09,1,le}, F), (x_{PI\leq 42.09,1,gt}, T), (x_{PI\leq 42.09,1,s}, F), (x_{PI\leq 42.09,2,le}, F), (x_{PI\leq 42.09,2,gt}, F), (x_{PI\leq 42.09,2,s}, T)$
$(x_{PI\leq 70.62,1,le}, T), (x_{PI\leq 70.62,1,gt}, F), (x_{PI\leq 70.62,1,s}, F), (x_{PI\leq 70.62,2,le}, F), (x_{PI\leq 70.62,2,gt}, F), (x_{PI\leq 70.62,2,s}, T)$
$(x_{GS\leq 37.89,1,le}, F), (x_{GS\leq 37.89,1,gt}, F), (x_{GS\leq 37.89,1,s}, T), (x_{GS\leq 37.89,2,le}, F), (x_{GS\leq 37.89,2,gt}, T), (x_{GS\leq 37.89,2,s}, F)$
$(c_{1,1}, F), (c_{1,2}, T), (c_{2,1}, F), (c_{2,2}, T)\}$

Veja que como $x_{PI\leq 42.09,1,gt}$ é verdadeiro nessa valoração, então $PI > 42.09$ deve aparecer na primeira regra. Além disso, como $x_{PI\leq 42.09,2,s}$ é verdadeiro nessa valoração, então esse atributo não vai aparecer (nem como $PI \leq 42.09$ e nem como $PI > 42.09$) na segunda regra. Já para o atributo $PI \leq 70.62$, observe que $x_{PI\leq 70.62,1,le}$ é verdadeiro na valoração e, portanto, deve aparecer na primeira regra. Além do mais, como $x_{PI\leq 70.62,2,s}$ é verdadeiro nessa valoração, então esse atributo não deve aparecer na segunda regra. Por fim, como $x_{GS\leq 37.89,1,s}$ e $x_{GS\leq 37.89,2,gt}$ são verdadeiros nessa valoração, então esse atributo não deve aparecer na primeira regra e deve aparecer como $GS > 37.89$ na segunda regra. Portanto, o conjunto de regras obtido a partir da valoração acima é:

$$\{[PI > 42.09, PI \leq 70.62] \Rightarrow P,$$
$$[GS > 37.89] \Rightarrow P\}$$

Depois de obter as regras e para testar a sua modelagem, você deve verificar se o conjunto de regras obtido classifica corretamente todos os indivíduos do arquivo. Para cada paciente no arquivo, você deve usar as regras obtidas e atribuir um diagnóstico de acordo. Em seguida, você pode verificar se o seu programa atribuiu o diagnóstico correto, ou seja, aquele que está registrado no arquivo.