Vírusfertőzésre való érzékenység modellezése Bayes-hálóval

2020. Október 28.

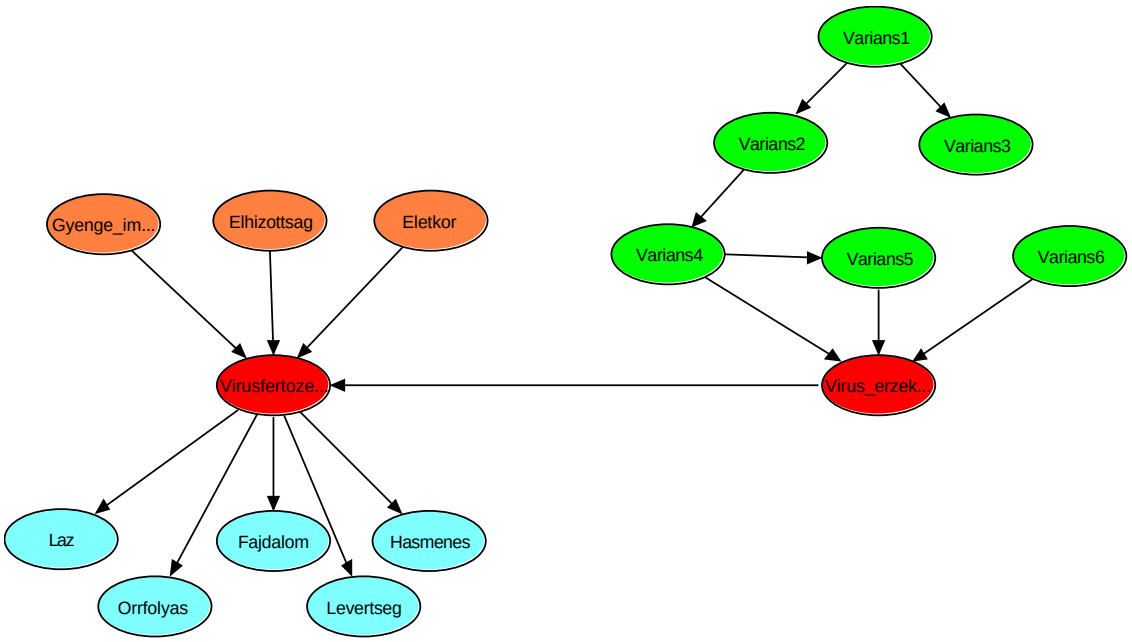
Feladat

Orvosbiológiai területen gyakran találkozhatunk olyan problémákkal, ahol az összefüggések bizonytalanok az élő szervezetek komplexitása valamint a vizsgálati módszerek és a megfigyelhetőség korlátai miatt. Ilyenkor az egyik lehetséges eszköz, melyet ilyen jellegű tudás reprezentációjára használhatunk, a valószínűségi gráfos modellek osztálya. Ezek a modellek egyfelől lehetővé teszik a bizonytalan tudás ábrázolását, másrészt az így felépített összefüggések rendszerében való következtetést.

Ebben a házi feladatban a hallgató feladata egy vírusfertőzésre való érzékenységet vizsgáló Bayes-háló kialakítása és az abban való következtetés. A vírusfertőzésre való érzékenységhez és egy potenciális vírusfertőzés súlyosságához kötődő függőségi kapcsolatokat az alábbi modell írja le:

- Potenciális vírusfertőzés súlyosságát leíró változó (*Virus_sulyossag*), amely három értéket vehet fel: {*enyhe, közepes, súlyos*}.
- A vírusfertőzés súlyosságát három előzményváltozó befolyásolja:
  - gyenge immunrendszer (*Gyenge_immun*:{*nem,igen*})
  - elhízottság (*Elhizott*:{*nem,igen*})
  - életkor (*Eletkor*:{*nem kitett,kitett*})
- A vírusfertőzés súlyosságának következményeként többféle tünet állhat elő:
  - láz (*Laz*:{*nem,igen*})
  - fájdalom (*Fajdalom*:{*nem,igen*})
  - hasmenés (*Hasmenes*:{*nem,igen*})
  - orrfolyás (*Orrfolyas*:{*nem,igen*})
  - levertség (*Levertseg*:{*nem,igen*}).
- A vírusfertőzés súlyosságát egy további faktor befolyásolja, melyet a vírusfertőzésre való érzékenység csomóponttal jelölünk: (*Virus_erzekeny*:{*nem,igen*}).
- A vírusfertőzésre való érzékenységet számos genetikai tényező befolyásolja, ezek hatását két vagy háromértékű változókkal modellezzük. Tehát a *Variáns-1*, *Variáns-2*,..., *Variáns-k* nevű csomópontok azt jelölik, hogy egy adott genetikai tényező milyen mértékben tekinthető rizikósnak a vírusérzékenység szempontjából {*alacsony*, *magas*} vagy {*alacsony*, *közepes*, *magas*}. A Variáns csomópontok száma és struktúrája bemenetenként változó lehet.

A modellben szereplő változók közötti függőségi kapcsolatokat az 1. ábrán látható gráfstuktúra reprezentálja. A vírusfertőzés súlyosságához kapcsolódó csomópontok alkotta részgráf struktúrája kötött, a vírusfertőzésre való érzékenységet reprezentáló csomóponthoz kapcsolódó genetikai variánsok alkotta részgráf viszont **egyedi minden feladatnál** (eltérő számú csomópontot és éleket tartalmaz). A paraméterezést lokális feltételes valószínűségi táblák formájában a bemenet adja meg.
A feladat a következő részekre osztható:



1. ábra. Vírusfertőzés súlyosságát modellező Bayes-háló struktúrája

1. A Bayes-háló struktúrájának (aciklikus irányított gráf) kialakítása a bemenetben megadott szülő–gyermek függőségi viszonyok alapján.
2. A Bayes-háló paraméterezésének meghatározása lokális feltételes valószínűségi táblák segítségével a bemenet alapján.
3. Evidencia változók értékének rögzítése a bemenet alapján.
4. Egzakt következtetés megvalósítása egy kijelölt célváltozóra, adott evidenciák mellett.
5. Célváltozó eloszlásának (lehetséges értékei valószínűségének) visszaadása eredményként.

Mindezek alapján valósítsa meg a specifikációban megadott Bayes-hálót, majd alkalmazza azt a bemenetben leírt következtetések megvalósítására. Részpontszámot csak a helyesen visszaadott eredmények után lehet szerezni.

Bemenet

A feladat bemenete a leírásban részletezett Bayes-háló struktúrájának, feltételes valószínűségi tábláinak (CPT) és a benne ismert evidenciáknak szöveges leírásából, illetve a célváltozó indexéből áll, az alábbiak szerint:

- Az első sor mindig egy egész számot (a továbbiakban: N_v) tartalmaz, amely a Bayes-háló csomópontjainak számát jelöli.
- A következő N_v darab sor a háló változóinak (csomópontjainak) leírását tartalmazza topologikus sorrendben[1], ahol minden sor (csomópont) az alábbi sémát követi:
  <$\boldsymbol{k}$> \t <$\boldsymbol{n}_{Pa}$> \t <I_1> \t <I_2> ... \t <$I_{\boldsymbol{n}_{Pa}}$>
  \t <v_{11}>,<v_{12}>...,<$v_{1\boldsymbol{n}_{Pa}}$>:<p_{11}>,<p_{12}>...,<p_{1k}>
  ...
  \t <v_{c1}>,<v_{c2}>...,<$v_{c\boldsymbol{n}_{Pa}}$>:<p_{c1}>,<p_{c2}>...,<p_{ck}> \n

[1] Topologikus sorrend esetén a minden csomópont összes szülője hamarabb szerepel a felsorolásban, mint az adott csomópont.

ahol:

  - $\boldsymbol{k}$: az adott változó által felvehető diszkrét értékek száma (pl. bináris csomópontnál $\boldsymbol{k} = 2$)
  - \t: tabulátor (tab) karakter
  - $\boldsymbol{n}_{Pa}$: az adott változó szülőinek száma
  - I_i: az adott változó i-edik szülőjének[2] indexe 0-val kezdődő indexelést használva, a változók (sorok) kiírási sorrendje szerint
  - v_{ij}: a j-edik szülő által a szülők összes lehetséges értékkombinációi közül az i-edik kombinációban felvett érték
  - p_{ij}: a sor által definiált változó által felvehető összes lehetséges érték közül a j-edik érték valószínűsége feltéve, hogy a szülők a lehetséges értékkombinációik közül az i-edik kombinációt veszik fel
  - \n: a sor végét jelző "new line" karakter

- A változók (csomópontok) leírását követően a következő sor egy egész számot (a továbbiakban: N_e) tartalmaz, amely az evidencia-változók számát jelöli (tehát azt, hogy az eddig ismertetett változók közül mennyinek ismerjük az értékét).
- Az evidencia-változók számát leíró sort követően jön az azok által fölvett érték leírása, soronként egy-egy változó index-érték párral az alábbi módon:
  <V_e>\t<v>\n
  ahol:
  - V_e: a sor által leírt evidencia változó indexe (0-ról indulva, a változók korábbi felsorolásának sorrendje szerint, akárcsak a szülőknél)
  - v: a sor által leírt evidencia változó által felvett érték
  - Példa: Ha a $V_e = 1$-es változó (tehát X_1) evidenciaként a $v = 2$ értéket veszi föl, akkor azt az alábbi sor írja le:
    `1 2`
- Végül pedig a kimenet utolsó sora a célváltozó indexét tartalmazza, amelynek az eloszlását ki kell számítani a megadott evidencia függvényében.

Példa: Vegyünk egy három csomópontból álló Bayes-hálót, amelyek csomópontjai X_0, X_1 és X_2. A hálóban X_2-nek X_0 és X_1 a szülői, X_0-nak és X_1-nek nincsen szülője. Tegyük fel, hogy X_0 két lehetséges értéket vehet föl (tehát esetében $\boldsymbol{k} = 2$), X_1 lehetséges értékeinek száma pedig három ($\boldsymbol{k} = 3$). Legyenek X_0 értékeinek valószínűségei:

$$P(X_0 = 0) = 0.352; P(X_0 = 1) = 0.648$$

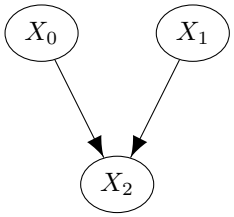
X_1 értékeinek valószínűségei pedig:

$$P(X_1 = 0) = 0.01; P(X_1 = 1) = 0.39; P(X_1 = 2) = 0.6$$

Továbbá a háló struktúrája, és az X_2 változó feltételes valószínűségi táblája a 1. táblázatban látható.
Ezeken felül tételezzük fel, hogy $X_2 = 1$ (tehát X_2 értéke ismert, vagyis X_2 része az evidenciának), és a célváltozónk pedig X_1, vagyis a feladatban az X_1 változó lehetséges értékeinek valószínűségeit kell kiszámolni. Ezek fényében a példához tartozó bemenet az alábbi[3]:

```
3
2  0          0.352,0.648
3  0          0.01,0.39,0.6
2  2  0  1    0,0:0.3,0.7  0,1: 0.5,0.5  0,2: 0.4,0.6  1,0: 0.8,0.2  1,1: 0.2,0.8  1,2: 0.7,0.3
1
2  1
1
```

[2] A változók (csomópontok) felsorolásához hasonlóan itt a szülők mindig topologikus sorrendben szerepelnek.
[3] Az átláthatóság kedvéért az eredeti tabulátoros elválasztás helyett táblázatos formában.



X_0	X_1	$P(X_2 = 0 \mid X_0, X_1)$	$P(X_2 = 1 \mid X_0, X_1)$
0	0	0.3	0.7
0	1	0.5	0.5
0	2	0.4	0.6
1	0	0.8	0.2
1	1	0.2	0.8
1	2	0.7	0.3

1. táblázat. X_2 változó feltételes valószínűségi táblája

Itt érdemes észrevenni, hogy az X_0 és X_1 csomópontok szülőinek száma 0, így a szülők indexeinek listája is üres, illetve a lehetséges értékeik valószínűségei is egy szimpla felsorolásban szerepelnek, mindig a 0-s értéktől kezdődően, a $(k-1)$-gyes értékkel bezárólag.

Kimenet

Kimenetként a bemenet utolsó sorában az indexével megjelölt célváltozó összes lehetséges értékének valószínűségét kell kiíratni, legalább 4 tizedesjegy pontossággal, a felvett értékek szerinti sorrendben, mindegyiket egy-egy új sorban (az utolsó sor végére is \n karaktert téve). Tehát például ha a célváltozó 3 darab lehetséges értéket vehet föl, amelyek valószínűségei: $\{P(0) = 0.257703; P(1) = 0.643554; P(2) = 0.098743\}$ akkor a kimenet az alábbi:

0.257703
0.643554
0.098743

A kimenet akkor elfogadható, ha a formátuma helyes, és a valószínűségek mindegyike (vagy együttesen: a célváltozó eloszlása) legfeljebb 10^{-4}-nel tér el az elvárt értéktől.

Fontos tudnivalók

- A megoldás forráskódja nem tartalmazhat ékezetes vagy nem ASCII[0:127] karaktert.
- Java nyelvű megoldás esetén a beadott forráskódnak tartalmaznia kell egy `Main` osztályt, azon belül egy `main()` függvényt. Külső csomagokat nem lehet használni.
- Python nyelvű megoldás esetén a feladatot megoldó script egyetlen `.py` kiterjesztésű fájlban kerülhet beadásra. Szabadon lehet használni a Python3 nyelv beépített könyvtárait (pl.: `math`, `functools`, stb...), azokon kívül viszont semmilyen egyéb, külső könyvtárat (pl.: `numpy`) nem lehet használni.
- A változók indexelése a felsorolásuk szerinti sorrend alapján történik 0-val kezdődően, tehát az elsőként listázott változó indexe 0, a következő változóé 1, és így tovább. Ez a sorrend egyben topologikus sorrend is.
- A bemenetként adott Bayes-háló - a bemenetben szereplő sorrend szerinti - utolsó 9 darab változója minden bemenetnél azonos.
- A megoldás csak akkor elfogadható, hogyha a célváltozó minden lehetséges értékének valószínűsége kiírásra kerül külön sorban és a megfelelő sorrendben, ezen értékek mindegyike legalább 4 tizedes jegyig meg van adva, és egyenként legfeljebb 10^{-4}-nel térnek el a helyes megoldástól. Ez a tolerancia első sorban a kerekítésből és a lebegőpontos számábrázolásból történő esetleges hibák kiküszöbölése végett került bevezetésre.
- A feltöltött megoldás megengedett futásideje CPU-időben bemenetenként 30 másodperc. Időtúllépés esetén a rendszer automatikusan leállítja a kód futását.

- A feltöltött megoldás összesen legfeljebb 400 MB memóriát allokálhat. Ezen érték túllépése esetén a rendszer automatikusan leállítja a kód futását.

Értékelés

A megoldást több különböző bemeneten értékeljük ki, a végleges pontszám pedig az alapján kerül kiszámításra, hogy ezek közül hány bemenetre adott helyes eredményt. Egy kimenetért pontosan akkor jár pont, hogyha elfogadható, tehát ha a formátuma helyes és a valószínűségek mindegyike az adott hibahatáron ($\pm 10^{-4}$) belülre esik.

Hasznos tippek

- A bemenetként adott Bayes-hálókban a nem ismert változók (tehát nem evidencia változók) száma általában alacsony, így tehát **javasolt az egzakt következtetés megvalósítása** a pontosság érdekében.
- Előfordulhat olyan bemenet, ahol az ($N_v - 10$)-es indexű változó (`Virus_erzekenyseg`) is szerepel az evidenciák között. Ebben az esetben fontos észrevenni, hogy mivel ezen csomópont minden felmenője genetikai variáns, és a genetikai variánsoknak nincs másik gyereke (egyéb generikai variánsokon kívül), így ebben az esetben (tehát ha ezen változó evidenciaként szerepel) a háló többi része független a genetikai variánsoktól. Ebben az esetben a következtetésnél érdemes elhagyni a hálóból a genetikai variánsokat (tehát azon csomópontokat, amelyek indexe kisebb, mint $N_v - 10$), jelentősen javítva ezáltal a következtetés futásidején.
- A feladat leírásához csatolva van három bemenet-kimenet pár, egyenként két-két `.txt` fájl formájában, amelyek jól példázzák, hogy a kiértékelés során milyen típusú bemenetek fordulhatnak elő:
  - Az `input1.txt` egy egyszerűbb bemenetet tartalmaz.
  - Az `input2.txt` egy olyan bemenetet tartalmaz, ahol habár a változók száma meglehetősen nagy, ezen változók többségének ismert az értéke (tehát szerepelnek az evidenciában), így az ismeretlen változók száma és az egzakt következtetés futásideje is kezelhető marad.
  - Az `input3.txt` egy olyan bemenetet tartalmaz, ahol a genetikai variánsok között sok az ismeretlen változó, így a teljes hálón normál esetben az egzakt következtetés nem lenne praktikus. Fontos észrevenni azonban, hogy szerepel az evidencia-változók között a `Virus\_erzekenyseg` változó értéke, így megtehetjük, hogy elhagyjuk a nála kisebb index-szel rendelkező genetikai variánsokat, tehát azon változókat amelyek indexe kisebb, mint ($N_v - 10$).