Célzott terápiás gyógyszer kiválasztása döntési hálóval

2021. Október 28.

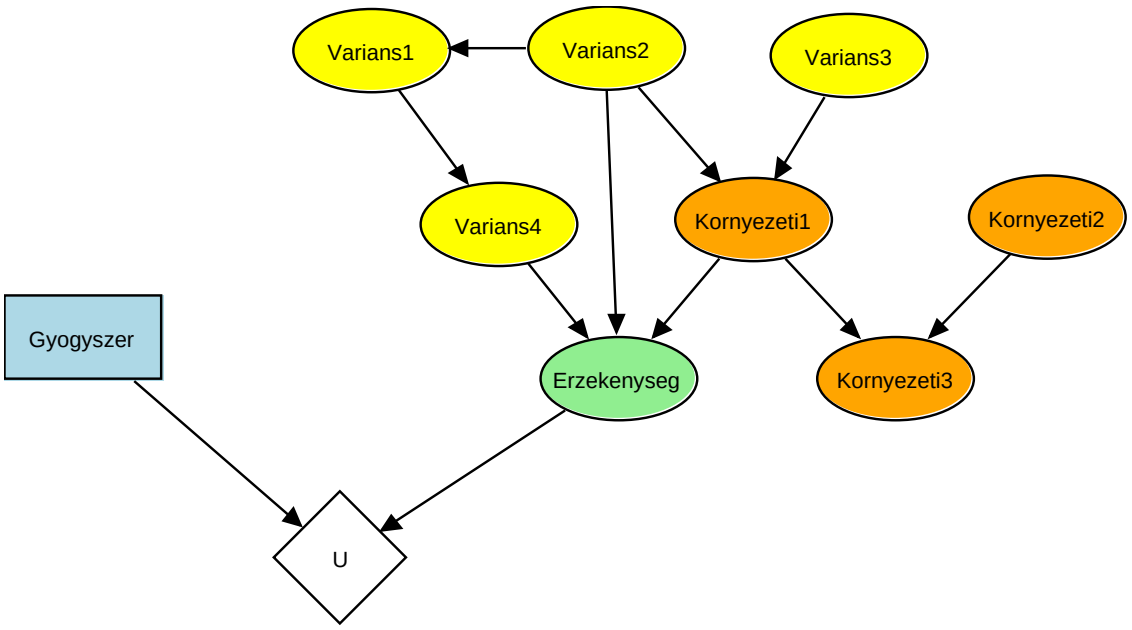
Feladat

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. Az ilyen modellek döntési- és hasznosság-csomóponttal kiegészített változatát döntési hálóknak nevezzük, amelyek a következtetés eredménye alapján lehetővé teszik a lehetséges döntések közül az optimális (legnagyobb várható haszonnal vagy legkisebb várható veszteséggel járó) döntés kiválasztását. A jelen házi feladat témája egy új onkológiai kezeléshez, az úgynevezett célzott terápiához kapcsolódik, amelynek lényege a pácienst jellemző környezeti változók és a daganat genomjában található egyes variánsok alapján annak a megállapítása, hogy a daganat sejtjei mely ismert - és nem feltétlen csak onkológiában használatos - gyógyszerekre lehetnek érzékenyek. Ezek megállapítása alapján akár hétköznapi, vagy teljesen más betegségek kezelésére kifejlesztett gyógyszerek is használhatóak lehetnek egyes daganatos betegségek effektív kezelésére, minimális mellékhatások mellett.

Ebben a házi feladatban a hallgató feladata egy onkológiai gyógyszerérzékenységet modellező Bayes-háló kialakítása, majd annak döntési hálóvá történő kibővítése és az abban való következtetés, valamint a lehetséges döntések közül az optimális döntés kiválasztása. A megvalósítandó modell alapvetően az alábbi komponenseket tartalmazza:

- Diszkrét környezeti változók, amelyek közül nem mindegyiknek ismert az értéke. Ezek sokféle értéket vehetnek föl, és az egyes betegségektől kezdve a szedett gyógyszereken át a táplálkozási és életviteli faktorokig sok releváns dolgot reprezentálhatnak.
- Diszkrét genetikai variánsok, amelyek közül (a génszekvenálás zajos mivolta miatt) nem mindegyiknek ismerjük az értékét.
- Az "Érzékenység" nevű, diszkrét értékű változó, amely a daganat lehetséges gyógyszerérzékenységi típusait reprezentálja.
- A "Gyógyszer" nevű döntési csomópont, amely azokat a lehetséges gyógyszeres kezeléseket reprezentálja, amik közül választanunk kell.
- Az "U" hasznosság-csomópont, amely az "Érzékenység" változó és a "Gyógyszer" döntés lehetséges értékkombinációinak hasznosságát adja meg.

Mivel az egyes daganattípusok releváns környezeti változói és genetikai variánsai mind különbözőek, így a felsorolt komponensekből előálló döntési hálóban található valószínűségi változók száma és struktúrája bemenetenként változó lehet. Emellett ezekhez igazodva bemenetenként különböző lehet a lehetséges gyógyszerek (döntések) száma is. Az így előálló lehetséges döntési hálókra az 1. ábra mutat egy példát.



1. ábra. A feladatban megvalósítandó döntési háló egy lehetséges struktúrája.

A feladat a következő részekre osztható:

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 a megjelölt célváltozóra, adott evidenciák mellett.
5. A célváltozó eloszlásának (lehetséges értékei valószínűségének) visszaadása eredményként.
6. A hasznosság rögzítése a lehetséges döntések és célváltozó értékek lehetséges kombinációira.
7. Az egyes döntések várható hasznosságának kiszámítása.
8. A legnagyobb várható hasznosságú döntés visszaadása eredményként.

Mindezek alapján valósítsa meg a fent részletezett döntési hálót, majd használja a bemenetben leírt következtetések megvalósítására és az optimális döntés meghatározására. Fontos, hogy eredményként az optimális döntés mellett az egyes érzékenység-típusok valószínűségét is vissza kell adni (a "Kimenet" c. részben megadott formátumban), ugyanis ez is kritikus információ lehet a végső döntést meghozó orvosok számára. Részpontszámot csak a helyesen visszaadott eredmények után lehet szerezni.

Bemenet

A feladat bemenete a leírásban részletezett döntési háló struktúrájának, feltételes valószínűségi tábláinak (CPT), a benne ismert evidenciáknak és a döntésekhez tartozó hasznosság-értékeknek a szöveges leírásából áll, az alábbiak szerint:

- Az első sor mindig egy egész számot (a továbbiakban: N_v) tartalmaz, amely a döntési 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
  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`
- A bemenet következő sora a célváltozó indexét tartalmazza, amelytől az egyes döntések hasznossága függ
- Az ezt követő sor a lehetséges döntések számát (n_d) tartalmazza
- Végül a következő $k_t \times n_d$ darab sor a lehetséges célváltozó-érték és döntés kombinációkhoz tartozó hasznosság értékeket tartalmazza, az alábbi formában:
  <t>\t<d>\t<U_{td}>\n
  ahol:
  - k_t: a célváltozó által felvehető diszkrét értékek száma
  - n_d: a lehetséges döntések száma
  - t: a célváltozó értéke
  - d: a döntés indexe

[1] Topologikus sorrend esetén a minden csomópont összes szülője hamarabb szerepel a felsorolásban, mint az adott csomópont.
[2] A változók (csomópontok) felsorolásához hasonlóan itt a szülők mindig topologikus sorrendben szerepelnek.

- U_{td}: a t célváltozó-érték és d döntés kombinációjához tartozó hasznosság (valós szám)
- Példa: Ha a $t = 0$ célváltozó-érték és $d = 2$ döntés kombinációjához $U_{02} = 34.5$ hasznosság tartozik, akkor azt az alábbi sor írja le:
  `0 2 34.5`

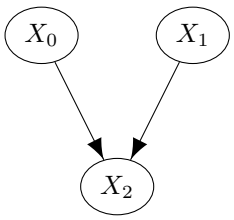
Példa: Vegyünk egy három csomópontból álló döntési hálót, amelyek csomópontjai X_0, X_1 és X_2. A hálóban X_2-nek X_0 és X_1 a szülői, X_0-nak és X_1-nek nincsen szülője. Tegyük fel, hogy X_0 két lehetséges értéket vehet föl (tehát esetében $\boldsymbol{k} = 2$), X_1 lehetséges értékeinek száma pedig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

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

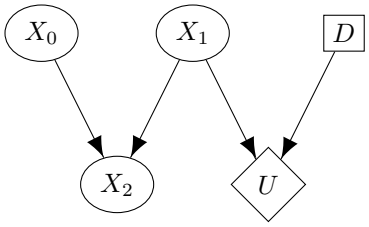
Továbbá a háló struktúrája, és az X_2 változó feltételes valószínűségi táblája az 1. táblázatban látható.



X_0	X_1	$P(X_2 = 0\|X_0, X_1)$	$P(X_2 = 1\|X_0, X_1)$
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. A példához tartozó háló struktúrája és az X_2 változó feltételes valószínűségi táblája.

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 fogja meghatározni a döntések várható hasznosságát, illetve $n_d = 2$ darab lehetséges döntésünk van. Az egyes célváltozó-érték és döntés kombinációkhoz tartozó hasznosságokat a 2. táblázat írja le.



X_1	D	U
0	0	153.2
0	1	−55.7
1	0	50.3
1	1	−125.1
2	0	−15.3
2	1	54.4

2. táblázat. A példához tartozó döntési háló struktúrája és a hasznosság-értékeket tartalmazó táblázat.

Ezek fényében a példához tartozó bemenet az alábbi:

```
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
0 0 153.2
0 1 -55.7
1 0 50.3
1 1 -125.1
2 0 -15.3
2 1 54.4
```

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 célváltozó eloszlását (tehát lehetséges értékeinek valószínűségeit) és az optimális (tehát legnagyobb várható hasznosságú) döntés indexét kell kiíratni. A lehetséges döntések várható hasznossága az adott hasznosság-értékek a célváltozó hozzájuk tartozó értékének valószínűségével súlyozott összege. Például hogyha a célváltozó eloszlása: { $P(X_1 = 0) = 0.257703$; $P(X_1 = 1) = 0.643554$; $P(X_1 = 2) = 0.098743$ }, $n_d = 2$ darab lehetséges döntésünk van és a hasznosság-értékeket a 2. táblázatban láthatóak szerint választjuk meg, akkor a lehetséges döntésekhez tartozó várható hasznosságokat az alábbiak szerint számíthatjuk ki:

$$EU_{d=0} = P(X_1 = 0)U_{00} + P(X_1 = 1)U_{10} + P(X_1 = 2)U_{20} \approx 70.3401$$

$$EU_{d=1} = P(X_1 = 0)U_{01} + P(X_1 = 1)U_{11} + P(X_1 = 2)U_{21} \approx -89,491$$

Ebben az esetben a $d = 0$ döntés az optimális, mivel annak magasabb a várható hasznossága. A célváltozó lehetséges értékeinek valószínűségét legalább 4 tizedesjegy pontossággal, a felvett értékek szerinti sorrendben, mindegyiket egy-egy új sorban (az utolsó sor végére is \n karaktert téve) kell kiíratni, majd ezt követően az optimális döntés indexét kell kiíratni az utolsó sorba, szintén egy "sor vége" (tehát \n) karakterrel bezárólag. Tehát az előző példához tartozó kimenet az alábbi:

```
0.257703
0.643554
0.098743
0
```

A kimenet akkor elfogadható, ha a formátuma helyes, és a várható hasznosságok mindegyike 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 kiértékelő rendszer 1.8.0-s JDK verzióval fogja fordítani és futtatni a beadott kódot.
- 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 kiértékelő rendszer 3.9.1-es verziójú Python interpreterrel fogja futtatni a beadott kódot.
- Python nyelvű megoldás esetén a feladatot megoldó script egyetlen `.py` kiterjesztésű fájlban kerülhet beadásra. Szabadon lehet használni a Python 3 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 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. Ezen felül az elfogadás feltétele az optimális döntés helyes kiíratása is.
- 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, a valószínűségek mindegyike az adott hibahatáron ($\pm 10^{-4}$) belülre esik és a kiírt döntés indexe is megegyezik az optimális döntés indexével.

Hasznos tippek

- A bemenetként adott döntési hálókban a nem ismert változók (tehát nem evidencia változók) száma általában alacsony, így tehát **javasolt az egzakt következtetés megvalósítása** a pontosság érdekében.
- A feladat leírásához csatolva van két bemenet-kimenet pár, egyenként két-két `.txt` fájl formájában, amelyek a sorszám növekedésével együtt egyre nehezebbek, és jól példázzák, hogy a kiértékelés során milyen típusú bemenetek fordulhatnak elő, és hogy azokra milyen választ kell adnia a megoldásként beadott programnak.