Практичне заняття №4

Тема: Побудова СППР на основі мереж Байєса

Мета: володіти методологією розробки мереж довіри Байєса

Теоретична частина

Методика побудови і застосування МБ

При побудові МБ з метою вирішення конкретної задачі необхідно:

1. Виконати аналіз проблеми і дати формалізовану постановку задачі. Сформулювати питання, на яке має бути отримана ймовірнісна відповідь у результаті формування ймовірнісного висновку за допомогою побудованої мережі.

2. Визначити множину даних, що відносяться до змінних задачі, отримати їх експертні оцінки та/або статистичні дані.

3. Поставити у відповідність усім отриманим даним взаємовиключні змінні.

4. Побудувати ациклічний граф, що відображає істотні умови незалежності змінних та існування причинно-наслідкових зв'язків.

5. Визначити апріорні ймовірності та оптимізувати топологію мережі на основі наявної інформації.

6. Виконати навчання мережі і провести формування висновку по відношенню до відповідних станів процесу.

7. Обробити результати: проаналізувати їх і зробити висновки щодо ймовірності очікуваної події.

Розглянемо приклад реалізації сформульованої методики при вирішенні досить простої задачі аналізу причин повернення товарів.Іноземна компанія, яка реалізує свою продукцію в Україні, збирає статистику щодо причин повернення товарів на склад. Розглядаються можливі варіанти: товар пошкоджений у процесі транспортування чи товар бракований. Із статистичних даних відомо, що ймовірність повернення товару через пошкодження при транспортуванні становить 0,05, а ймовірність повернення через брак — 0,3. Необхідно встановити ймовірність того, що повернений товар був бракований.

Використовуючи цю інформацію, будуємо мережу довіри Байєса, де R (Reject) — повернення товару саме через брак, D (Damaged) — повернення через пошкодження товару, RT (Return) — повернення товару на склад (рис. 4).

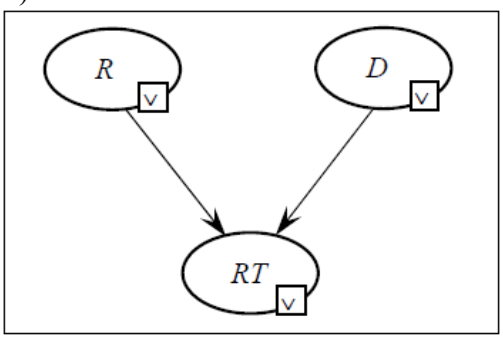


Рис. 4. Приклад мережі довіри Байєса для задачі повернення товарів

Ймовірність повернення товару для даної мережі обчислюється за формулою спільної ймовірності $P(RT) = P(R)\ P(D)\ P(RT \mid R,D)$. Значення ймовірностей вузлів R і D наведені у табл. 1, а значення умовних ймовірностей для вузла RT — у табл. 2.

Таблиця 1. Значення ймовірностей вузлів

Prob.	TRUE	FALSE
$P(D)$	0,05	0,95
$P(R)$	0,3	0,7

Таблиця 2. Значення умовних ймовірностей

$P(RT \mid R,D)$	D = TRUE	D = FALSE
R = TRUE	(0,9; 0,1)	(0,5; 0,5)
R = FALSE	(0,2; 0,8)	(0,1; 0,9)

Використовуючи теорему Байєса, можна обчислити ймовірність повернення товару через брак товару.

$$P(R = \text{TRUE} \mid RT = \text{TRUE}) = \frac{P(RT = \text{TRUE} \mid R = \text{TRUE})\,P(R = \text{TRUE})}{P(RT = \text{TRUE})}.$$

Ймовірність події $P(RT =$ TRUE) можна обчислити, використовуючи усі можливі значення R і D з таблиць 1 і 2.

P(RT = TRUE) = P(D = TRUE)P(R = TRUE)P(RT | D = TRUE, R = TRUE) +
+ P(D = FALSE)P(R = TRUE) P(RT | D = FALSE, R = TRUE) +
+ P(D = TRUE) P(R = FALSE)P(RT | D = TRUE, R = FALSE) +
+ P(D = FALSE)P(R = FALSE)P(RT | D = FALSE, R = FALSE) =

= 0,05 × 0,3× 0,9 + 0,95 × 0,3× 0,5 + 0,05 × 0,7 × 0,2 + 0,95 × 0,7 × 0,1= 0,2295 .

Обчислимо ймовірність того, що товар повернено тому, що він бракований.

$P(RT =$ TRUE $\mid R =$ TRUE$) = P(RT =$ TRUE $\mid R =$ TRUE$, D =$ TRUE$) \times$

$\times\ P(D =$ TRUE$) + P(RT =$ TRUE $\mid R =$ TRUE$, D =$ FALSE$) P(D =$ FALSE$) =$

= 0,9 × 0,05 + 0,5× 0,95 = 0,52 .

Тепер можемо обчислити ймовірність повернення товару через брак (рис. 5).

$$P(R = \text{TRUE} \mid RT = \text{TRUE}) = \frac{P(RT = \text{TRUE} \mid R = \text{TRUE})P(R = \text{TRUE})}{P(RT = \text{TRUE})} =$$

$$= \frac{0,52 \times 0,3}{0,2295} = 0,679739 .$$

Отже, знання про те, що товар було повернуто, спричиняє перегляд усіх ймовірностей появи цієї події. Зміни відбулися в збільшенні ступеня упевненості, що повернення сталося через брак товару: від початкового значення 0,3 до 0,68.

Типовий приклад

Розглянемо простий приклад, близький всім студентам. Щоб здати (Pass) іспит, потрібно підготуватися до нього (Study) або скористатися шпаргалкою (Cheat). Таким чином, є 3 булевих змінних. Хочеться дізнатися ймовірність успішно скласти іспит. У тих випадках, коли відомі ймовірності елементарних (атомарних) подій, можна скористатися методом імовірнісного виведення на основі повного спільного розподілу, яке описується таблицею розмірністю 2x2x2.

	Study		¬Study	
	Cheat	¬Cheat	Cheat	¬Cheat
Pass	0,15	0,4	0,04	0,06
¬Pass	0,01	0,04	0,05	0,25

Сума всіх ймовірностей дорівнює одиниці. У кожній клітці вірогідність настання елементарного події. Ця ймовірність є результуючої, тобто враховує в собі всі фактори. Так, ймовірність успішної здачі іспиту 0,4 враховує ймовірність підготовки до іспиту і ймовірність того,

що студент не скористався шпаргалкою. Ймовірності складних подій легко вирахувати підсумовуванням відповідних рядків або стовпців таблиці. Ймовірність підготовки до іспиту дорівнює сумі клітин лівої половини таблиці і відповідає подій (вчив, не користувався шпаргалкою і здав; вчив, користувався і здав; вчив, користувався і не здав; вчив, не користувався і не здав):

$$P(Study) = 0,15 + 0,4 + 0,01+ 0,04 = 0,6$$

Ймовірність скористатися шпаргалкою дорівнює сумі першого і третього стовпців:

$$P(Cheat) = 0,15 + 0,01 + 0,04 + 0,05 = 0,25$$

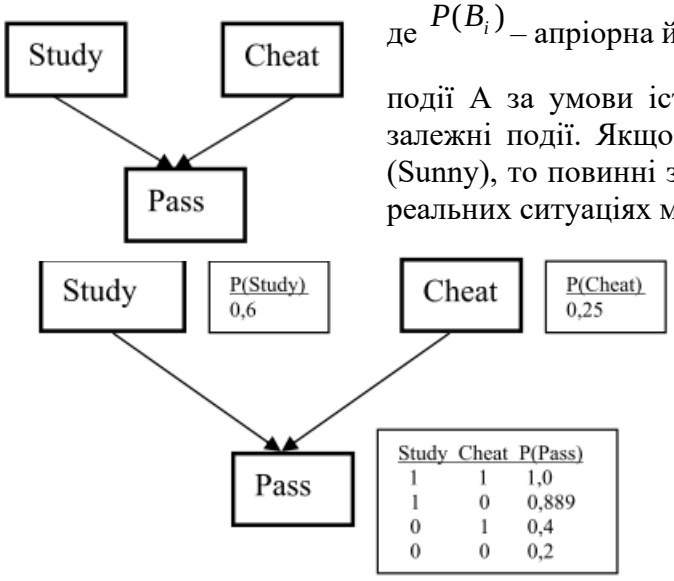
Ймовірність здачі іспиту дорівнює сумі клітин першого рядка (вчив, користувався і здав; вчив, не користувався і здав; не вчив, користувався і здав; не вчив, не користувався і здав):

$$P(Pass) = 0,15 + 0,4 + 0,04 + 0,06 = 0,65$$

Метод ймовірнісного виведення на основі повної спільного розподілу є скоріше доброю ілюстрацією принципу формування ймовірностей, ніж практичним посібником, оскільки ймовірності елементарних подій відомі далеко не завжди. Частіше роблять допущення про рівній ймовірності цих подій. Зазвичай можна оцінити ймовірності істинності окремих змінних. Нехай P(Study) = 0,6 (у 40% випадків будуть більш важливі справи, ніж підготовка до іспиту), а P(Cheat) = 0,25 (один шанс із чотирьох скористатися шпаргалкою). Ці ймовірності називаються апріорними або безумовними. Вони являють собою ступінь впевненості в істинності висловлювання відсутності інших даних.

На перший погляд, вірогідність здачі іспиту дорівнює 0,6 + 0,25 = 0,85. На насправді все складніше. Події Study та Cheat можуть перекриватися, тобто відбуватися спільно. Можна вивчити матеріал і при цьому для страховки скористатися шпаргалкою. Можна також все вивчити, але не здати (професор прискіпався). Можна здати без підготовки і шпаргалок (просто пощастило). Для знаходження шуканої ймовірності необхідно розташовувати умовними ймовірностями, наприклад, P(Pass|Study) – ймовірність здачі іспиту при умови повної підготовки. У загальному випадку ймовірність події A дорівнює

$$P(A) = \sum_i P(A|B_i)\, P(B_i)$$



де $P(B_i)$ – апріорна ймовірність події B_i, $P(A \mid B_i)$ – умовна мовірність події A за умови істинності події B_i. Умовна ймовірність пов'язує залежні події. Якщо ми введемо четверту змінну – сонячну погоду (Sunny), то повинні задавати умовні ймовірності типу P(A | Sunny). У реальних ситуаціях ми можемо зіткнутися з тим, що розмірність задачі буде з-за цього непомірно велика. Для вирішення цієї проблеми використовуються байесовские мережі, які дозволяють встановити залежність змінних і спростити обчислення повного спільного розподілу. Нижче наведена байєсова мережа для розглянутого прикладу. Звернемо увагу, що в нашій моделі змінні Study та Cheat є незалежними. Можлива інша модель, коли користування шпаргалкою обумовлено відсутністю підготовки до іспиту. Вона буде розглянуто пізніше. Кожна вершина мережі відповідає випадкової змінної. Вершини з'єднуються спрямованими ребрами. Якщо стрілка направлена від A до B, то A називається батьківською вершиною вершини B.

Кожна вершина X_i характеризується розподілом умовних ймовірностей

$$P(X_i \mid Parents(X_i)),$$

яке кількісно оцінює вплив на вершину її батьківських вершин. В нашому прикладі вважаємо відомими наступні умовні ймовірності: ймовірність здати іспит при умови підготовки та підстраховки шпаргалками дорівнює одиниці; при підготовку і не користуванні шпаргалкою – 0,889; за умови користування шпаргалкою без підготовки до іспиту – 0,4; а ймовірність здати іспит без підготовки та шпаргалки – 0,2 (просто пощастило). Основний виграш при використанні байєсівських мереж полягає

в те, що ймовірність будь-якого стану тільки на основі батьківських (найближчих) вершин, а не всіх вершин, від яких ця вершина залежить:

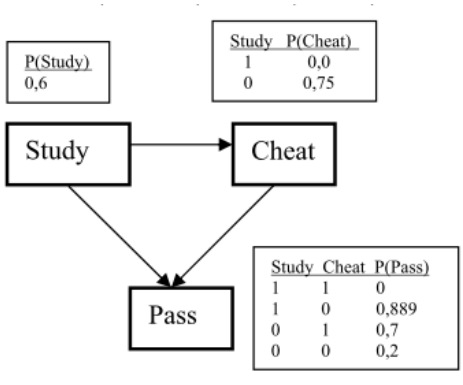
$$P(X_i \mid X_{i-1}, X_{i-2}, \ldots, X_1) = P(X_i \mid Parents(X_i))$$

У нашому прикладі ми маємо вершину Study, яка насправді може бути кінцевим результатом пов'язаних подій: студент відвідував лекції, мав конспект, у нього був доступ до комп'ютера, він мав часу, і т. п. Як результат цих подій ми маємо факт підготовки до іспиту. Подія використання шпаргалки також є наслідком низки подій, ймовірностями яких необхідно розташовувати: наявність часу, технічних засобів, достойне одяг для скритного використання шпаргалки. Для обчислення повного спільного розподілу потрібно знати всі умовні ймовірності, наприклад, ймовірність наявності радіопередавача за умови відвідування лекцій. Нам же для обчислення ймовірності здачі іспиту достатньо знати ймовірності P(Study) і P(Cheat).

Наведене вище правило називається ланцюговим правилом. Слідуючи цьому правилом досить просто обчислити ймовірності подій, послідовно просуваючись у напрямку стрілок. У даному прикладі ми можемо вичислити ймовірність складання іспиту:

P(Pass) = P(Pass | Study, Cheat) * P(Study) * P(Cheat) +
+ P(Pass | Study, Cheat) *P(Study) * P(Cheat) +
+ P(Pass | Study, Cheat) * P(Study) * P(Cheat) +
+P(Pass | Study, Cheat) * P(Study) * P(Cheat) =
= 1,0 * 0,6 * 0,25 + 0,889 * 0,6 * 0,75 + 0,4 * 0,4 * 0,25 + 0,2 * 0,4 * 0,75 = 0,65

У нашому сильно спрощеному прикладі цей виграш в складності обчислень непомітний, оскільки ланцюжок байєсівської мережі не така довга. Другий фактор – незалежність змінних Study та Cheat. Досвідчені студенти можуть помітити деяку неправдоподібність ймовірностей: за умови підготовки до іспиту шпаргалка лише додає ризик бути спійманим і вигнаним з іспиту. Змінимо логіку наступним чином. Будемо вважати, що намагатися скористатися шпаргалкою студент буде, тільки якщо не підготуватися до екзамену. Байєсова мережа зміниться так, як показано нижче. Умовна ймовірність скористатися шпаргалкою дорівнює нулю у випадку підготовки до іспиту і 0,75 при відсутності підготовки. Ймовірність



P(Pass | Study, Cheat) = 0.

Обчислимо P(Cheat) и P(¬Cheat):

P(Cheat) = P(Cheat | ¬Study) * P(¬Study) = 0,75 *(1-0,6) = 0,3

P(¬Cheat) = 1-P(Cheat) = 0,7

Тепер, використовуючи ланцюгове правило, ми можемо обчислити ймовірність успішної складання іспиту:

P(Pass) = P(Pass|Study,¬Cheat) * P(Study) * P(¬Cheat) + P(Pass | ¬Study, Cheat) *
P(¬Study) * P(Cheat) + P(Pass|¬Study, ¬Cheat) *
P(¬Study) * P(¬Cheat) = 0,889 * 0,6 * 0,7 + 0,7 * 0,4 * 0,3 + 0,2 * 0,4 * 0,7 = 0,513

Байесовские мережі дозволяють вирішувати і обернені задачі. Наприклад, відомо, що студент здав іспит успішно. Потрібно знайти ймовірність того, що він був до іспиту підготовлений. Для цього треба скористатися **правилом Байєса:**

$$P(A \mid B) = P(B \mid A) * P(A) / P(B)$$

У нашому випадку ймовірність здачі іспиту, який був успішно зданий, P(Pass) = 0,513; ймовірність здачі іспиту при підготовці до нього P(Pass | Study, Cheat) = 0,889; ймовірність підготовки і користування шпаргалкою P(Study, Cheat) = 0,6 * 0,7 = 0,42, отже,

P(Study, Cheat | Pass) = P(Pass | Study, Cheat) * P(Study, Cheat) / P(Pass) =0,889* 0,6 * 0,7 / 0,513 = 0,727

Знайдемо тепер ймовірність того, що іспит складено з допомогою шпаргалки:

P(¬Study, Cheat | Pass) = P(Pass | ¬Study, Cheat) * P(¬Study, Cheat) / P(Pass) =
= 0,7 * 0,4* 0,3 * / 0,513 = 0,164

І, нарешті, ймовірність того, що причиною здачі іспиту було чисте везіння:

P(Study,Cheat | Pass) = P(Pass | Study,Cheat) * P(Study, Cheat) / P(Pass)
= 0,2 * 0,4 * 0,7 / 0,513 = 0,109

Таким чином, байесовские мережі забезпечують декомпозицію складних завдань і при цьому позбавляють від необхідності задавати безліч умовних ймовірностей.

Приклад побудови найпростішої байєсівської мережі довіри.

Розглядаємо невелику яблучну плантацію «яблучного Джека». Одного разу Джек виявив, що його прекрасне яблучне дерево позбулося листя. Тепер він хоче з'ясувати, чому це сталося. Він знає, що листя часто опадає, якщо: дерево засихає в результаті нестачі вологи; або дерево хворіє. Дана ситуація може бути змодельована байєсівської мережею довіри, що містить 3 вершини: «Хворіє», «всохло» і «облетіло».

Рис.1. Приклад байєсівської мережі довіри з трьома подіями.

У даномунайпростішомувипадку розглянемо ситуацію, при якій кожна вершина може приймати всього лише два можливих станів і, як слідство знаходиться в одному з них, а саме:

Вершина (подія) БСД	Стан 1	Стан 2
"Хворіє"	«Хворіє»	«Ні»
"Засохло"	«Всохло»	«Ні»
"Облетіло"	«Так»	«Ні»

Вершина "Хворіє" говорить про те, що дерево захворіло, будучи в стані «хворіє», в іншому випадку вона знаходиться в стані «ні». Аналогічно для інших двох вершин. Розглянута байєсівську мережа довіри, моделює той факт, що є причинно-наслідковий залежність від події "Хворіє" до події "облетіло" і від події "засохло" до події "облетіло". Це відображено стрілками на байєсівської мережі довіри.

Коли є причинно-наслідковий залежність від вершини А до іншої вершині В, то ми очікуємо, що якщо А перебуває в деякому певному стані, це впливає на стан В. Слід бути уважним, коли моделюється залежність в байєсівських мережах довіри. Іноді зовсім не очевидно, який напрямок повинна мати стрільця.Наприклад, в розглянутому прикладі, ми говоримо, що є залежність від "Хворіє" до "облетіло", тому що коли дерево хворіє, це може викликати опадання його листя. Обпадання листя є наслідком хвороби, а не хвороба - наслідком опадання листя.На наведеному вище малюнку дано графічне представлення байєсівської мережі довіри. Однак, це тільки якісне уявлення байєсівської мережі довіри. Перед тим, як назвати це повністю байєсівської мережею довіри необхідно визначити кількісне уявлення, тобто безліч таблиць умовних ймовірностей:

Апріорна ймовірність p ("Хворіє")

Хворіє = «хворіє»	Хворіє = «ні»
0,1	0,9

Апріорна ймовірність p ("засохло")

Засохло = «всохло»	Засохло = «ні»
0,1	0,9

Таблиця умовних ймовірностей p ("облетіло" | "Хворіє", "засохло")

	Засохло = «всохло»	Засохло = «ні»

	Хворіє = «хворіє»	Хворіє = «ні»	Хворіє = «хворіє»	Хворіє = «ні»
Облетіло = «так»	0,95	0,85	0,90	0,02
Облетіло = «ні»	0,05	0,15	0,10	0,98

Наведені таблиці ілюструють три вершини байєсівської мережі довіри. Зауважимо, що всі три таблиці показують імовірність перебування деякої вершини в певному стані, обумовленим станом її батьківських вершин. Але так як вершини Хворіє і засохло не мають батьківських вершин, то їх ймовірності є маргінальними, тобто не залежать (не обумовлені) ні від чого.На даному прикладі ми розглянули, що і як описується дуже простий байєсівської мережею довіри. Сучаснiпрограмні засоби(такі як MSBN, Hugin тощо) забезпечують інструментарій для побудови таких мереж, а також можливість використання байєсівських мереж довіри для введення нових свідоцтв та отримання рішення (висновку) за рахунок перерахунку нових ймовірностей у всіх вершинах,відповіднихнововведеним свідченнями.У нашому прикладі хай відомо, що дерево скинуло листя. Це свідчення вводиться вибором стану «так» у вершині "облетіло". Після цього можна дізнатися ймовірності того, що дерево всохло. Для наведених вище вихідних даних, результати виведення шляхом поширення вірогідності по БСД будуть:p ("Хворіє" = «хворіє» | "облетіло" = «так») = 0,47; p ("засохло" = «всохло» | "облетіло" = «так») = 0,49.**Завдання до виконання:**

1. Побудуйте мережу Байєса із використанням комп’ютерної програми Netica (або інших програм). Тему вибрати самостійно. [14]
2. Завантажте готові приклади мереж Байєса:

Chest Clinic – мережа для діагностування пацієнта (в літературі ця мережа нерідко зустрічається під назвою Asia)

Car diagnostic – мережа для визначення ймовірності того, що автомобіль заводиться.

Готові приклади мереж Chest Clinic та Car diagnostic можна знайти в папці **Data**.

3. Для мережі Chest Clinic задайте наступні інстанційованні значення

Smoking = smoker

Visit to Asia = visit

та обчисліть значення ймовірностей вершин X Ray Result та Dyspnea за допомогою програми Netica та вручну.

4. Результати виконання лабораторної роботи необхідно оформити у вигляді звіту, який повинен містити:

– структуру мережі Байєса призначеної оцінки кредитоспроможності клієнта, Chest Clinic та Car diagnostic;

– для кожної мережі змоделювати 2-3 ситуації наявності інстанційованих значень вершин в мережі.

Результати слід представити у вигляді таблиці. Наприклад, для мережі оцінки кредитоспроможності клієнта, таблиця результатів моделювання може мати вигляд:

№	Інстанційовані вершини	Результат моделювання
1	TypeOfContract = Empl_Full_Time Age = more_36	P(Result = bad) = 6,57% P(Result = Good) = 93,43%
2	ContactPerson = no MaritalStatus = Widowed	P(Result = bad) = 7,79% P(Result = Good) = 92,21%

Результати обчислення вручну та із використанням програми **для** мережі **Chest Clinic** представити у вигляді таблиці:

Вершина	Результати Netica	Результати, одержані вручну
X Ray Result = abnormal		
X Ray Result = normal		
Dyspnea = present		
Dyspnea = absent		

5. Висновки.