

Міністерство освіти і науки, молоді та спорту України

Національний університет “Львівська політехніка”

Кафедра автоматизованих систем управління

Методичні вказівки

до лабораторної роботи №1

“ Data Mining – класифікація та регресія.

Методи побудови правила класифікації”

з дисципліни

“Інтелектуальний аналіз даних”

для студентів базового напрямку підготовки по спеціальності

“ Комп’ютерні науки ” (шифр 122)

Львів-2019

Методичні вказівки до лабораторної роботи № 1 “Data Mining – класифікація та регресія. Методи побудови правила класифікації” з дисципліни *“*Інтелектуальний аналіз даних*”* для студентів спеціальності - шифр 122 *“Комп’ютерні науки”*/ Укл. доц. *Ковівчак Я.В.*, Львів: Національний університет “Львівська політехніка”, 2019.

Методичні вказівки обговорено та схвалено на засіданні кафедри АСУ Протокол № \_\_\_\_\_\_\_\_\_\_\_ від «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_2019 р.

Завідувач кафедрою АСУ \_\_\_\_\_\_\_\_\_\_\_\_\_\_ ***Цмоць І. Г.***

Методичні вказівки обговорено та схвалено на засіданні методичної комісії базового напрямку підготовки

Протокол № \_\_\_\_\_\_\_\_\_\_\_ від «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_2019 р.Протокол № \_\_\_\_\_\_\_\_\_\_\_ від «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_2019 р.

**Лабораторна робота № 1**

**Методи побудови правила класифікації.**

***Мета:*** Оволодіти методами побудови правил класифікації та навчитись застосовувати ці знання на практиці.

***Теоретична частина:***

*Data Mining* – сукупність методів, алгоритмів та засобів видобування інформації з сирих, неопрацьованих даних.

**Завдання класифікації**

Класифікація є найбільш простою і одночасно найбільш часто розв'язуваною задачею Data Mining. Зважаючи на поширеність задач класифікації необхідне чітке розуміння суті цього поняття.

*Класифікація* – одна з основних задач Data Mining, яка полягає у віднесенні об’єктів до наперед визначених класів на основі значень їх характеристик або визначення значення залежної (класової) змінної на основі значень незалежних змінних (характеристик об’єктів).

Класифікація вимагає дотримання наступних правил:

* В кожному акті поділу необхідно застосовувати тільки одну ознаку;
* Розподіл має бути пропорційним, тобто загальний обсяг видових понять повинен дорівнювати обсягу діленого родового поняття;
* Члени поділу повинні взаємно виключати один одного, їх обсяги не повинні перехрещуватися;
* Розподіл має бути послідовним.

Розрізняють:

* допоміжну (штучну) класифікацію, яка проводиться за зовнішніми ознаками і служить для надання множині предметів (процесів, явищ) потрібного порядку;
* природну класифікацію, яка проводиться по істотних ознаках, що характеризують внутрішню спільність предметів і явищ. Вона є результатом і важливим засобом наукових досліджень, тому передбачає і закріплює результати вивчення закономірностей об'єктів, що класифікуються.

В залежності від обраних ознак, їх поєднання та процедури поділу понять класифікація може бути:

* простою - поділ родового поняття тільки за ознакою і тільки один раз до розкриття усіх видів. Прикладом такої класифікації є дихотомія, при якій членами поділу бувають тільки два поняття, кожне з яких є суперечним іншому (тобто дотримується принцип: "А і не А");
* складною - застосовується для поділу одного поняття з різних підстав і синтезу таких простих розподілів в єдине ціле. Прикладом такої класифікації є періодична система хімічних елементів.

Під класифікацією будемо розуміти віднесення об'єктів (спостережень, подій) до одного з наперед відомих класів.

*Класифікація* - це закономірність, що дозволяє робити висновок щодо визначення характеристик конкретної групи. Таким чином, для проведення класифікації повинні бути присутніми ознаки, що характеризують групу, до якої належить та чи інша подія або об'єкт (зазвичай при цьому на підставі аналізу вже класифікованих подій формулюються якісь правила).

Класифікація відноситься до стратегії навчання з вчителем (supervised learning), яке також іменують контрольованим або керованим навчанням.

Завданням класифікації часто називають пророкування категоріальної залежної змінної (тобто залежної змінної, яка є категорією) на основі вибірки неперервних і / або категоріальних змінних.

Наприклад, можна передбачити, хто з клієнтів фірми є потенційним покупцем певного товару, а хто - ні, хто скористається послугою фірми, а хто - ні, і т.д. Цей тип завдань відноситься до завдань бінарної класифікації, в них залежна змінна може приймати тільки два значення (наприклад, так чи ні, 0 або 1).

Інший варіант класифікації виникає, якщо залежна змінна може приймати значення з деякої множини визначених класів. Наприклад, коли необхідно передбачити, яку марку автомобіля захоче купити клієнт. У цих випадках розглядається безліч класів для залежної змінної.

Класифікація може бути одномірною (за однією ознакою) та багатовимірної (з двома і більше ознаками).

Багатовимірна класифікація була розроблена біологами при вирішенні проблем дискримінації для класифікування організмів. Однією з перших робіт, присвячених цьому напряму, вважають роботу Р. Фішера (1930 р.), в якій організми розділялися на підвиди залежно від результатів вимірювань їх фізичних параметрів. Біологія була і залишається найбільш популярним і зручним середовищем для розробки багатовимірних методів класифікації.

В задачі класифікації потрібно визначити значення залежної змінної об’єкту на основі значень інших змінних, що характеризують даний об’єкт. Формально задачу класифікації можна описати наступним чином. Маємо множину об’єктів:

*I={i1, i2, … ,ij, … , in},* (1.1)

де *ij* – об’єкт, що досліджується. Прикладом таких об’єктів може бути інформація про проведення ігор при різних погодніх умовах (табл. 1.1).

*Таблиця 1.1*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Спостереження** | **Температура** | **Вологість** | **Вітер** | **Гра** |
| Сонце | Жарко | Висока | Нема | Ні |
| Сонце | Жарко | Висока | Є | Ні |
| Хмарно | Жарко | Висока | Нема | Так |
| Дощ | Норма | Висока | Нема | Так |
| Дощ | Холодно | Норма | Нема | Так |
| Дощ | Холодно | Норма | Є | Ні |
| Хмарно | Холодно | Норма | Є | Так |
| Сонце | Норма | Висока | Нема | Ні |
| Сонце | Холодно | Норма | Нема | Так |
| Дощ | Норма | Норма | Нема | Так |
| Сонце | Норма | Норма | Є | Так |
| Хмарно | Норма | Висока | Є | Так |
| Хмарно | Жарко | Норма | Нема | Так |
| Дощ | Норма | Висока | Є | Ні |

Кожен об’єкт характеризується набором змінних:

*Ij={x1, x2, … , xh, … , xm, y},* (1.2)

де *xh* – незалежні змінні, значення яких відомі та на основі яких визначається значення залежної змінної *y*

В Data Mining часто набір незалежних змінних позначають у вигляді вектора:

*X={x1, x2, … , xh, … , xm}.* (1.3)

Кожна змінна xh може приймати значення з деякої множини:

*Ch={ch1, ch2, … }.* (1.4)

Якщо значеннями змінної є елементи скінченної множини, то говорять, що вона має категоріальний тип. Наприклад, змінна *спостереження* приймає значення на множині значень *{сонце, хмарно, дощ}.*

Якщо множина значень *C={c1, c2, … , cr, … , ck}* змінної *y* скінченна, то задача називається задачею класифікації.

***Представлення результатів***

Правила класифікації складаються з двох частин: умови та висновку – *якщо {значення}, то {висновок}.*

Умовою є перевірка одної чи декількох незалежних змінних. Перевірки декількох змінних можуть бути об’єднані за допомогою операції «і» , «або» чи «не». Висновком буде значення залежної змінної або розподіл її ймовірності по класах, наприклад:

*якщо {спостереження = сонце, температура = жарко} то {гра = ні};*

*якщо {спостереження = хмарно, температура = холодно} то {гра = так}.*

Основною перевагою правил є легкість їх сприйняття та запис природною мовою. Ще одна перевага – їх відносна незалежність. В набір правил легко додати нове правило без необхідності змінювати вже існуючі. Відносність незалежності правил зв’язана з тим, що можливо вони будуть суперечити один одному. Якщо змінні, що характеризують деякий об’єкт, задовільняють умовні частини правил з різними висновками, то виникає невизначеність зі значенням його залежної змінної. Наприклад, нехай у нас є правила:

*якщо {спостереження = сонце } то {гра = ні};*

*якщо {спостереження = хмарно, температура = холодно} то {гра = так}.*

У них об’єкти, що задовільняють умови другого правила, задовільняють умови і першого. Та все-таки висновок різний. Іншими словами, у відповідності з цими правилами при однакових обставинах можемо отримати суперечливі вказівки, що недопустимо.

***Алгоритм побудови 1-правил***

Розглянемо найпростіший алгоритм формування елементарних правил для класифікафії об’єкту. Він будує правила по значенню одної незалежної змінної, тому в літературі його часто називають «1-правило» (1-rule) або коротко 1R-алгоритм.

Ідея алгоритму дуже проста. Для будь-якого можливого значення кожної незалежної змінної формується правило, яке класифікує об’єкти із навчальної вибірки. При цьому в заключній частині правила вказується значення залежної змінної, яке найбільш часто зустрічається в об’єктів із вибраним значенням незалежної змінної. В даному випадку помилкою правила буде кількість об’єктів, які мають те ж значення біжучої змінної, але не відносяться до вибраного класу.

Таким чином, для кожної змінної буде отримано набір правил (для кожного значення). Оцінивши ступінь помилки кожного набору, ви­би­ра­єть­ся змінна, для якої побудовані правила з найменшою помилкою.

Для прикладу, представленого в табл. 1.1, в результаті будуть отримані правила та їх оцінки, приведені в табл. 1.2.

*Таблиця 1.2*

|  |  |
| --- | --- |
| **Правило** | **Помилка** |
| Якщо (спостереження = сонце) тоді (гра = ні) | 2/5 |
| Якщо (спостереження = хмарно) тоді (гра = так) | 0/4 |
| Якщо (спостереження = дощ) тоді (гра = так) | 2/5 |
| Якщо (температура = жарко) тоді (гра = ні) \* | 2/4 |
| Якщо (температура =норма) тоді (гра = так) | 2/6 |
| Якщо (температура = холодно) тоді (гра =так) | 1/4 |
| Якщо (вологість = висока) тоді (гра = ні) | 3/7 |
| Якщо (вологість = норма) тоді (гра = так) | 1/7 |
| Якщо (вітер = немає) тоді (гра = так) | 2/8 |
| Якщо (вітер = є) тоді (гра = ні) \* | 3/6 |

Якщо в навчальній вибірці зустрічаються об’єкти з пропущеними значеннями незалежних змінних, то алгоритм 1R підраховує такі об’єкти для кожного можливого значення змінної.

Іншою проблемою для цього алгоритму є чисельні значення змінних. Очевидно, якщо змінна має дійсний тип, то кількість допустимих значень може бути незкінченною. Для вирішення цієї проблеми всю область значень такої змінної розбивають на інтервали таким чином, щоб кожен з них співвідносився з відповідним класом в навчальній вибірці. В результаті буде отриманий набір дискретних значень, з якими може працювати даний алгоритм.

Допустимо, що дані змінної «температура», приведені у прикладі, мають наступні числові значення і відповідні їм значення залежної змінної:

*4 5 8 9 10 11 12 12 15 15 20 21 23 25*

*так |ні| так так так |ні ні| так так так |ні| так так |ні*

В цьому випадку діапазон значень можна було би розбити на інтервали наступним чином:

*{до 4,5; 4,5-7,5; 7,5-10,5; 10,5-12; 12-17,5; 17,5-20,5; 20,5-24; більше 24}*

Більш серйозна проблема даного алгоритму – це надчутливість (overfitting). Справа в тому, що алгоритм буде обирати змінні, які приймають найбільшу кількість можливих значень, так як для них помилка буде найменшою. Наприклад, для змінної, що є ключем помилка дорівнюватиме нулю. Тим не менш, для таких змінних правила будуть абсолютно бездіяльними, тому при формуванні навчальної вибірки для цього алгоритму важливо правильно обрати набір незалежних змінних.

На завершення слід зауважити, що алгоритм 1R, незважаючи на свою простоту, в багатьох випадках на практиці виявляється досить ефективним. Це пояснюється тим, що багато об’єктів дійсно можна класифікувати лише за одним атрибутом. Крім того, небагаточисленність сформованих правил дозволяє легко зрозуміти і використати отримані результати.

**Методи класифікації**

Для класифікації використовуються різні методи.

Основні з них:

1. Байєсова (“наївна”) класифікація;
2. Класифікація за допомогою штучних нейронних мереж;
3. Класифікація методом опорних векторів;
4. Статистичні методи, зокрема, лінійна регресія;
5. Класифікація за допомогою методу найближчого сусіда;
6. Класифікація CBR-методом;
7. Класифікація за допомогою генетичних алгоритмів.
8. Класифікація за допомогою дерев рішень

**Метод "найближчого сусіда"**

Метод "найближчого сусіда" ("nearest neighbour") відноситься до класу методів, робота яких грунтується на зберіганні даних в пам'яті для порівняння з новими елементами. При появі нового запису для прогнозування знаходяться відхилення між цим записом та подібними наборами даних, і найбільш подібний (або ближній сусід) ідентифікується.

При розгляді нового клієнта банку, його атрибути порівнюються з усіма існуючими клієнтами даного банку (дохід, вік і т.д.). Безліч "найближчих сусідів" потенційного клієнта банку вибирається на підставі найближчого значення доходу, віку і т.д. При такому підході використовується термін "k-найближчий сусід" ("k-nearest neighbour"). Термін означає, що вибирається k "верхніх" (найближчих) сусідів для їх розгляду в якості безлічі "найближчих сусідів". Оскільки не завжди зручно зберігати всі дані, іноді зберігається тільки безліч "типових" випадків.

У такому випадку використовуваний метод називають міркуванням за аналогією (Case Based Reasoning, CBR), міркуванням на основі аналогічних випадків, міркуванням за прецедентами.

*Прецедент* - це опис ситуації в поєднанні з детальним зазначенням дій, що вживаються в даній ситуації.

Підхід, заснований на прецедентах, умовно можна поділити на наступні етапи:

• збір детальної інформації про поставлену задачу;

• зіставлення цієї інформації з деталями прецедентів, що зберігаються в базі, для виявлення аналогічних випадків;

• вибір прецеденту, найбільш близького до поточної проблеми, з бази прецедентів;

• адаптація вибраного рішення до поточної проблеми, якщо це необхідно;

• перевірка коректності кожного отриманого рішення;

• занесення детальної інформації про новий прецедент в базу прецедентів.

Таким чином, висновок, заснований на прецедентах, являє собою такий метод аналізу даних, який робить висновки щодо даної ситуації за наслідками пошуку аналогій, що зберігаються в базі прецедентів.

Даний метод за своєю суттю відноситься до категорії "навчання без вчителя", тобто є "самонавчальною" технологією, завдяки чому робочі характеристики кожної бази прецедентів з плином часу і накопиченням прикладів поліпшуються. Розробка баз прецедентів по конкретній предметній області відбувається природною для людини мовою, та може бути виконана найбільш досвідченими співробітниками компанії - експертами або аналітиками, які працюють в даній предметній області.

Однак це не означає, що CBR-системи самостійно можуть приймати рішення. Даний метод лише пропонує можливі варіанти вирішення і вказує на найбільш оптимальний.

*Переваги методу:*

* Простота використання отриманих результатів.
* Рішення є універсальними (можливе їх використання для інших випадків).
* Метою пошуку є не гарантовано вірне рішення, а найкраще з можливих.

*Недоліки методу:*

* Даний метод не створює будь-яких моделей чи правил, узагальнюючих попередній досвід. При обранні рішення він грунтується на всьому масиві доступних історичних даних, тому неможливо точно сказати, на якій підставі формуються відповіді.

*1. Приклад застосування методу*

Навчальна вибірка складається з атрибутів (незалежних змінних) ім’я, вік, стать людини та залежної змінної (мітки класу) фанат (вказує фанатом яких музичних виконавців чи колективів є людина).

*Таблиця 1.3*

|  |  |  |  |
| --- | --- | --- | --- |
| **Ім’я** | **Вік** | **Стать** | **Фанат** |
| Bill | 32 | M | Rolling Stones |
| Henry | 40 | M | Neither |
| Mary | 16 | F | Taylor Swift |
| Tiffany | 14 | F | Taylor Swift |
| Michael | 55 | M | Neither |
| Carlos | 40 | M | Taylor Swift |
| Ashely | 20 | F | Neither |
| Robert | 15 | M | Taylor Swift |
| Sally | 55 | F | Rolling Stones |
| John | 15 | M | Rolling Stones |

Припустимо, ми хочемо дізнатись до якого класу буде належати Michelle, яка є жінкою (Стать) та має 5 років (Вік).

Обираємо число найближчих сусідів, які братимуть участь в класифікації за загальноприйнятою формулою k = , де N – об’єм навчальної вибірки. В нашому випадку .



Ми хочемо передбачити, чи є особа, яка не входить у вищенаведену таблицю фанатом Taylor Swift, Rolling Stones або не є фанатом жодного з вищезазначених виконавців. Для класифікації використаємо атрибути вік та стать.

Такі дані, як вік описуються числовим значення, в той час як стать ми описуємо словом. Нам необхідно перетворити дискретне значення в числове. Тоді ми можемо вказати, що Male = 0, а Female = 1. Тепер всі наші дані є числовими.

Відстань між об’єктами розрахуємо за Евклідовою нормою.

Відстань між об’єктами з координатами () і () дорівнює:



Для нашого випадку, можна трактувати, як:



Отримаємо таблицю з додатковим показником «Відстань»:

*Таблиця 1.4*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Ім’я** | **Вік** | **Стать** | **Відстань** | **Фанат** |
| Bill | 32 | M | 27.02 | Rolling Stones |
| Henry | 40 | M | 35.01 | Neither |
| Mary | 16 | F | 11.00 | Taylor Swift |
| Tiffany | 14 | F | 9.00 | Taylor Swift |
| Michael | 55 | M | 50.01 | Neither |
| Carlos | 40 | M | 35.01 | Taylor Swift |
| Ashely | 20 | F | 15.00 | Neither |
| Robert | 15 | M | 10.00 | Taylor Swift |
| Sally | 55 | F | 50.00 | Rolling Stones |
| John | 15 | M | 10.05 | Rolling Stones |

Так, як k = 3, то найближчими сусідами до Michelle є:

* Tiffany, яка має відстань - 9
* Robert, який має відстань - 10
* John, який має відстань - 10.05

Тепер розглянемо до яких класів належать “найближчі сусіди”:

* Tiffany , яка має відстань 9 ---> Taylor Swift
* Robert , який має відстань 10 ---> Taylor Swift
* John, який має відстань 10.05 ---> Rolling Stones

Оскільки клас Taylor Swift зустрічається двічі, то метод найближчого сусіда передбачає, що Michelle є фанатом Taylor Swift.

**Важливо**

Якщо ми оберемо значення k занадто малим, то є небезпека, що єдиним найближчим об'єктом виявиться «викид», який дасть невірне рішення. Здавалося б, збільшуючи значення параметра k, ми знижуємо ймовірність випадкового потрапляння на такі «викиди» в якості найближчих сусідів досліджуваного об'єкта. Але тут виникає інша небезпека. Щоб зрозуміти в чому вона полягає, розглянемо випадок, коли k дорівнює загальній кількості об'єктів N. Зрозуміло, що тоді «переможе» найпопулярніший (модальний) клас, і відстань до досліджуваного об'єкта не буде грати взагалі ніякої ролі.

Проблему вибору оптимального значення параметра k називають «bias-variance tradeoff», тобто «Компроміс між « викидами» і дисперсією». **На практиці найчастіше вважають k = []**. В нашому прикладі k = 3 та результатом класифікації є те, що Michelle є фанатом Taylor Swift.



У тому випадку, якщо ми впевнені в «чистоті» вибірки, ми можемо обирати k меншим. Існує також прийом під назвою «weighted voting» (тобто буквально «зважене голосування»), при якому ближчі сусіди досліджуваного об'єкта мають більшу вагу, ніж більш віддалені.

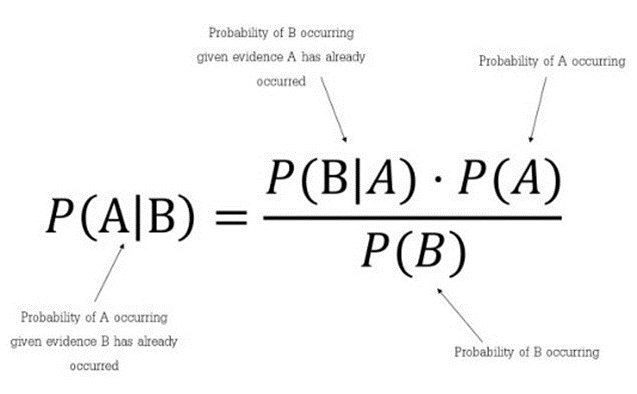
**Байєсова (“наївна”) класифікація;**

“*Наївна” Байєсова класифікація* - це алгоритм класифікації, заснований на теоремі Байєса з припущенням про незалежність ознак. Отже, іншими словами, цей метод класифікації передбачає, що наявність якої-небудь ознаки в класі не пов'язано з наявністю будь-якої іншої ознаки.  У зв'язку з таким припущенням алгоритм називається «наївним». Моделі на основі алгоритму досить прості і вкрай корисні при роботі з дуже великим обсягом даних. При своїй простоті “наївна” Байєсова класифікація здатна перевершити навіть деякі складні алгоритми класифікації.

*Теорема Байєса* - це проста математична формула, яка використовується для обчислення умовних ймовірностей.

*Умовна ймовірність* - це ймовірність настання однієї події за умови, що інша подія (за припущенням, допущенню, підтверджених або непідтверджених доказом твердженням) вже відбулася.

Теорема Баєса задається наступною формулою:



Теорема показує, як часто відбувається подія A при настанні події B, позначається як P (A | B) і має другу назву «апостеріорна ймовірність». При цьому потрібно знати:

1. Як часто відбувається подія B при настанні події A, що позначається у формулі як P (B | A);

2. Наскільки ймовірним є те, що A не залежить від інших подій, що позначається у формулі як P (A);

3. Наскільки ймовірним є те, що B не залежить від інших подій. У формулі вона позначається як P (B).

*Властивості наївної класифікації:*

* Використання всіх змінних і визначення всіх залежностей між ними;
* Наявність двох припущень щодо змінних;
* Всі змінні є однаково важливими;
* Всі змінні є статистично незалежними, тобто значення однієї змінної нічого не говорить про значення іншої.

**Кроки вирішення задач шляхом використання алгоритму “наївної” Байєсової класифікації:**

**Крок 1.** Перетворити набір даних в частотну таблицю (frequency table).

**Крок 2.** Створити таблицю правдоподібності (likelihood table), розрахувавши відповідні ймовірності.

**Крок 3.** За допомогою теореми Байєса розрахувати апостеріорну ймовірність для кожного класу при заданих умовах. Клас з найбільшою апостеріорною ймовірністю буде результатом розв’язку задачі.

*Переваги методу:*

* в моделі визначаються залежності між усіма змінними, це дозволяє легко обробляти ситуації, в яких значення деяких змінних невідомі;
* байєсовський метод дозволяє природним чином поєднувати закономірності, виведені з даних, і, наприклад, експертні знання, отримані в явному вигляді;

*Недоліки методу*

* перемножувати умовні ймовірності коректно лише тоді, коли всі вхідні змінні дійсно статистично незалежні; хоча часто даний метод показує досить гарні результати при недотриманні умови статистичної незалежності, але теоретично така ситуація повинна оброблятися більш складними методами, заснованими на навчанні байєсівських мереж;
* на результати класифікації впливають тільки індивідуальні значення вхідних змінних, комбінований вплив пар або трійок значень різних атрибутів тут не враховується. Це могло б покращити якість класифікаційної моделі з точки зору її прогнозуючої точності, проте, збільшило б кількість варіантів, які необхідно перевірити.

*2. Приклад застосування методу*

**Навчальна вибірка:**

*Таблиця 1.5*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **День** | **Прогноз** | **Температура** | **Вологість** | **Вітер** | **Грати в м’яч** |
| D1 | Дощ | Гаряче | Висока | Слабкий | Ні |
| D2 | Дощ | Гаряче | Висока | Сильний | Ні |
| D3 | Похмуро | Гаряче | Висока | Слабкий | Так |
| D4 | Сонячно | Тепло | Висока | Слабкий | Так |
| D5 | Сонячно | Прохолодно | Нормальна | Слабкий | Так |
| D6 | Сонячно | Прохолодно | Нормальна | Сильний | Ні |
| D7 | Похмуро | Прохолодно | Нормальна | Сильний | Так |
| D8 | Дощ | Тепло | Висока | Слабкий | Ні |
| D9 | Дощ | Прохолодно | Нормальна | Слабкий | Так |
| D10 | Сонячно | Тепло | Нормальна | Слабкий | Так |
| D11 | Дощ | Тепло | Нормальна | Сильний | Так |
| D12 | Похмуро | Тепло | Висока | Сильний | Так |
| D13 | Похмуро | Гаряче | Нормальна | Слабкий | Так |
| D14 | Сонячно | Тепло | Висока | Сильний | Ні |

Необхідно з’ясувати чи варто при заданих погодніх умовах (характеристиках) грати в м’яч, тобто визначити клас (так або ні).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Прогноз** | **Температура** | **Вологість** | **Вітер** | **Грати в м’яч** |
| Дощ | Прохолодно | Висока | Сильний | ? |

**Крок 1.** Перетворюємо набір даних в частотні таблиці (frequency tables).

***Частотні таблиці***

*Таблиця 1.6*

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Грати в м’яч** | |
| **Так** | **Ні** |
| **Прогноз** | Сонячно | 3 | 2 |
| Похмуро | 4 | 0 |
| Дощ | 2 | 3 |

*Таблиця 1.7*

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Грати в м’яч** | |
| **Так** | **Ні** |
| **Темп.** | Гаряче | 2 | 2 |
| Тепло | 4 | 2 |
| Прохолодно | 3 | 1 |

*Таблиця 1.8*

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Грати в м’яч** | |
| **Так** | **Ні** |
| **Вологість** | Висока | 3 | 4 |
| Нормальна | 6 | 1 |

*Таблиця 1.9*

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Грати в м’яч** | |
| **Так** | **Ні** |
| **Вітер** | Слабкий | 6 | 2 |
| Сильний | 3 | 3 |

**Крок 2.** Створимо таблицю правдоподібності (likelihood table), розрахувавши відповідні ймовірності.

***Таблиці правдоподібності***

*Таблиця 1.10*

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Грати в м’яч** | |
| **Так** | **Ні** |
| **Прогноз** | Сонячно | 3/9 | 2/5 |
| Похмуро | 4/9 | 0/5 |
| Дощ | 2/9 | 3/5 |

*Таблиця 1.11*

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Грати в м’яч** | |
| **Так** | **Ні** |
| **Темп.** | Гаряче | 2/9 | 2/5 |
| Тепло | 4/9 | 2/5 |
| Прохолодно | 3/9 | 1/5 |

*Таблиця 1.12*

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Грати в м’яч** | |
| **Так** | **Ні** |
| **Вологість** | Висока | 3/9 | 4/5 |
| Нормальна | 6/9 | 1/5 |

*Таблиця 1.13*

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Гра в м’яч** | |
| **Так** | **Ні** |
| **Вітер** | Слабкий | 6/9 | 2/5 |
| Сильний | 3/9 | 3/5 |

**Крок 3.** За допомогою теореми Байєса розраховуємо апостеріорну ймовірність для кожного класу при заданих умовах. Клас з найбільшою апостеріорною ймовірністю буде результатом розв’язку задачі.



0.8 > 0.2, отже грати в м’яч у таку погоду не варто (клас Ні).

**Метод опорних векторів (SVM)**

*Метод опорних векторів* (Support Vector Machine - SVM) відноситься до групи граничних методів. Він визначає класи за допомогою меж областей.

За допомогою даного методу вирішуються завдання бінарної класифікації. Мета методу опорних векторів - знайти площину, що розділяє дві множини об'єктів. В основі методу лежить поняття площин рішень. Площина рішень розділяє об'єкти з різною класовою приналежністю.

*Приклад*

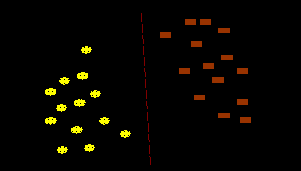


Рис. 1. Поділ класів прямою лінією

На Рис.1 наведено приклад, в якому беруть участь об'єкти двох типів. лінія яка розділяє задає границю, праворуч від якої - всі об'єкти типу brown (коричневий), а зліва - типу yellow (жовтий). Новий об'єкт, що потрапляє направо, класифікується як об'єкт класу brown або - як об'єкт класу yellow, якщо він розташувався по ліву сторону від розділяючої прямої. У даному випадку кожен об'єкт характеризується двома вимірами.

Вирішення задачі бінарної класифікації за допомогою методу опорних векторів полягає в пошуку деякої лінійної функції, яка правильно розділяє набір даних на два класи.

*Задача 1*

Розглянемо задачу класифікації, де число класів дорівнює двом.

Задачу можна сформулювати як пошук функції f(x), яка приймає значення менше нуля для векторів одного класу і більше нуля - для іншого. В якості вихідних даних для вирішення поставленого завдання, тобто пошуку класифікуючої функції f(x), дано тренувальний набір векторів простору, для яких відома їх приналежність до одного з класів. Сімейство класифікуючих функцій можна описати через функцію f(x). Гіперплощина визначена вектором а й значенням b, тобто f (x) = ax + b.

Вирішення даної задачі  проілюстровано на Рис.3.

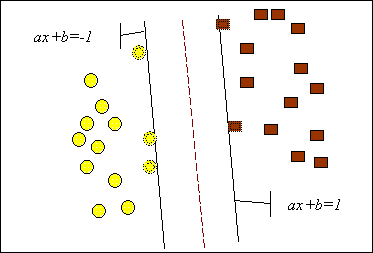


Рис. 3. Лінійний SVM.

В результаті рішення задачі, тобто побудови SVM-моделі, знайдена функція приймає значення менше нуля для векторів одного класу і більше нуля - для векторів іншого класу.

Для кожного нового об'єкта негативне або позитивне значення визначає його приналежність до одного з класів.

Найкращою функцією класифікації є функція, для якої очікуваний ризик мінімальний. Поняття очікуваного ризику в даному випадку означає очікуваний рівень помилки класифікації.

Безпосередньо оцінити очікуваний рівень помилки побудованої моделі неможливо, це можна зробити за допомогою поняття емпіричного ризику. Однак слід враховувати, що мінімізація останнього не завжди призводить до мінімізації очікуваного ризику. Цю обставину слід пам'ятати при роботі з відносно невеликими наборами тренувальних даних.

*Емпіричний ризик* - рівень помилки класифікації на тренувальному наборі.

Таким чином, в результаті вирішення задачі методом опорних векторів для лінійно роздільних даних ми отримуємо функцію класифікації, яка мінімізує верхню оцінку очікуваного ризику.

Однією з проблем, пов'язаних з вирішенням завдань класифікації розглянутим методом, є та обставина, що не завжди можна легко знайти лінійну межу між двома класами.

У таких випадках один з варіантів вирішення проблеми - збільшення розмірності, тобто перенесення даних з площини в тривимірний простір, де можливо побудувати таку площину, яка ідеально розділить безліч зразків на два класи. Опорними векторами в цьому випадку будуть служити об'єкти з обох класів, які є екстремальними.

Таким чином, за допомогою додавання так званого оператора ядра і додаткових розмірностей, знаходяться межі між класами у вигляді гіперплощин.

Однак слід пам'ятати: складність побудови SVM-моделі полягає в тому, що чим вище розмірність простору, тим складніше з ним працювати. Один з варіантів роботи з даними високої розмірності - це попереднє застосування будь-якого методу пониження розмірності даних для виявлення найбільш істотних компонент, а потім використання методу опорних векторів.

Як і будь-який інший метод, метод SVM має свої сильні і слабкі сторони, які слід враховувати при виборі даного методу.

*Переваги методу:*

* на відміну від більшості інших методів вимагає досить невеликого набору даних
* при правильній роботі моделі, побудованої на тестовій множині, цілком можливе застосування даного методу на реальних даних.

*Недоліки методу:*

* для класифікації використовується не вся множина зразків, а лише їх невелика частина, яка знаходиться на границях

*Метод опорних векторів дозволяє:*

* отримати функцію класифікації з мінімальною верхньої оцінкою очікуваного ризику (рівня помилки класифікації);
* використовувати лінійний класифікатор для роботи з нелінійно роздільними даними, поєднуючи простоту з ефективністю.

*3. Приклад застосування методу*

**Математичні формули**

g(x) = w \* x + w0

g(x) ≥ 1, ∀x ∈ A

g(x) ≤ -1, ∀x ∈ B

Де

w - нормальний вектор до поділяючої гіперплощини

w0 – допоміжний параметр

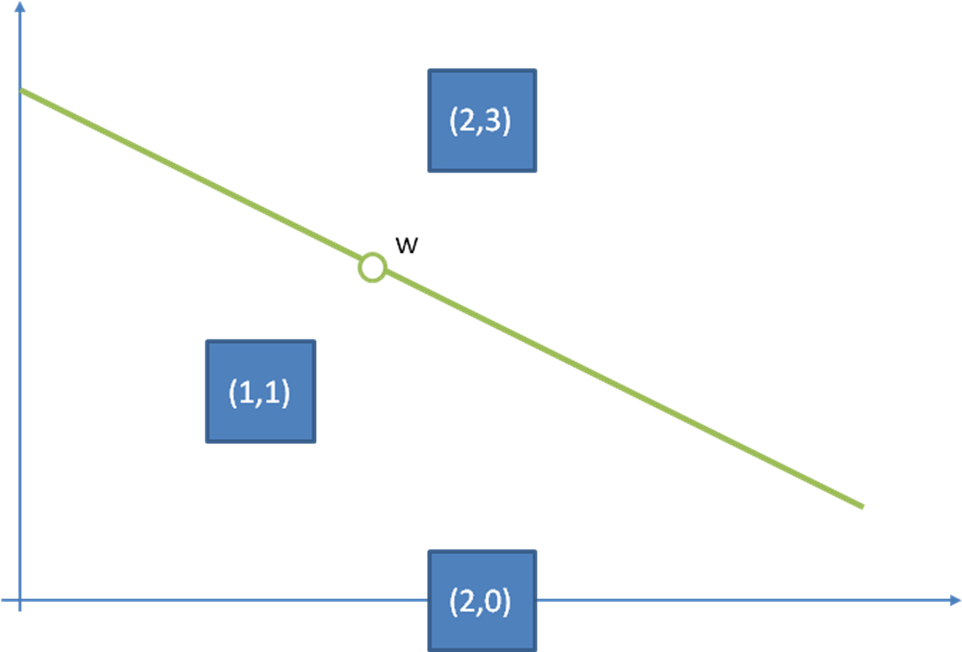
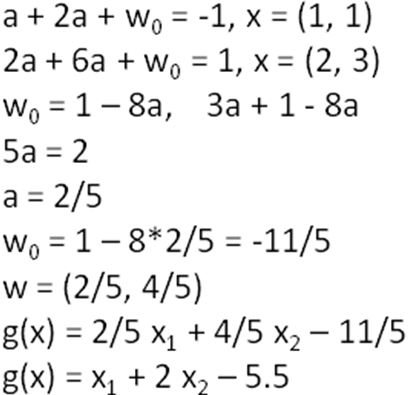


Рис. 4. Пошук класифікатора

**Контрольні запитання:**

1. Які є методи побудови правил класифікації.
2. Що таке Data Mining?
3. Якою буває класифікація?
4. Дотримання яких правил вимагає класифікація?
5. Чому не можна в кожному акті поділу застосовувати кілька ознак?
6. Яким має бути розподіл для класифікації?
7. Що таке крос-валідація і для чого вона потрібна?
8. Які є методи класифікації, які вимагають крос-валідації?
9. Що таке прецедент?
10. Чому Байєсовська класифікація називається також «наївною»?
11. Що таке емпіричний ризик та коли він трапляється?
12. Що таке «перенавчання»?
13. Де використовується коефіцієнт слабкості *learnrate*?
14. В чому полягає суть методу «найближчого сусіда»?
15. Де застосовується Байєсівська класифікація і чому?

Навчальне видання

**“**Інтелектуальний аналіз даних**”**

Методичні вказівки до лабораторної роботи № 4 “Data Mining – класифікація та регресія. Методи побудови правила класифікації ” з дисципліни *“* Інтелектуальний аналіз даних *”* для студентів спеціальності 0804 *“Комп’ютерні науки”*

Укладач:

**доц. Ковівчак Ярослав Васильович**

Комп’ютерний набір, верстку та редагування

здійснили ст. гр. КН-30, каф. АСУ, Никитенко А.Ю., Хомицький Р.Б., Труняк М.С.