Лекція 2

Класифікація у векторному просторі. Метод Роккіо [1]

У цій лекції для представлення ознак при класифікації текстів ми використовуємо векторну модель. У ній кожна ознака розглядається як вектор, що складається з дійсних чисел, як правило — з ваг *tf–idf* кожного терміна, де вага $tf\text{-}idf = tf \times idf$, *tf* — частота терміну в документі, *idf* — обернена частота терміну в документі, яка дорівнює $\log\frac{N}{df}$, де N — загальна кількість документів у колекції, а *df* — частота документів, у яких зустрічається термін. Таким чином, простір ознак $\mathbb{X}$, тобто область визначення функції класифікації γ, збігається з простором $\mathbb{R}^{|V|}$, де V — кількість термінів у колекції. У цій главі описуються методи класифікації, що оперують векторами дійсних чисел.

В основі використання моделі векторного простору для класифікації лежить *гіпотеза компактності*.

> **Гіпотеза компактності.** Ознаки, що належать тому самому класу, утворюють компактну область, причому області, що відповідають різним класам, не перетинаються.

Існує багато задач класифікації текстів, зокрема задачі, у яких класи відрізняються вживанням слів. Наприклад, ознаки в класі *China*, швидше за все, мають великі значення на осях, що відповідають термінам Chinese, Beijing і Mao, у той час как ознаки з класу *UK* — великі значення на осях, що відповідають термінам London, British і Queen. Отже, ознаки з двох класів утворять різні неперервні області. Між цими областями можна провести межі і класифікувати нові ознаки. Саме це є темою даної глави.

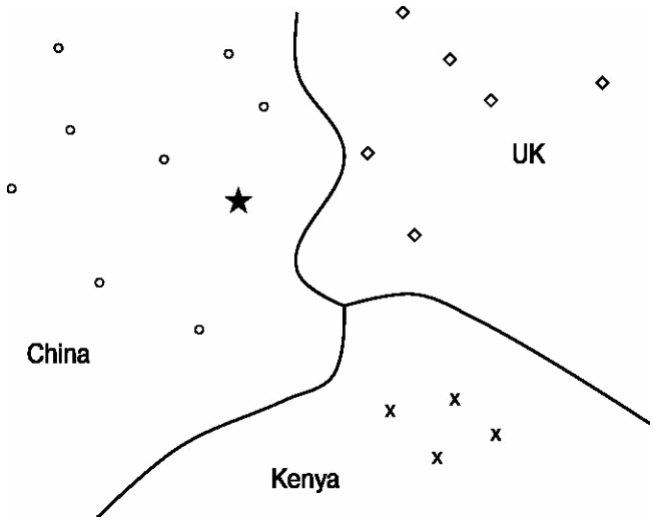

Рис. 2.1 Класифікація на три класи

Чи заповнює множина ознак неперервну область, залежить від конкретного вибору представлення ознаки: типу зважування, списку стоп-слів і т.д. Для того щоб переконатися, что представлення ознаки грає дуже важливу роль, розглянемо два класи (ознак): *написані групою авторів* і *написані окремою людиною*. Висока частота займенника першої особи *I* (*я*), очевидно, є ознакою другого класу. Однак ця інформація, швидше за все, буде вилучена з представлення ознаки, якщо використовується список стоп-слів. Якщо представлення ознаки обране невдало, то гіпотеза компактності не буде виконуватися і класифікація у векторному просторі стане неможливою.

У даному випадку можна повторити ті ж міркування, що привели нас до зважених представлень, зокрема — до нормалізованого по довжині представлення tf–idf. Наприклад, термін, що п'ять разів зустрічається в ознаці, повинний мати більшу вагу, ніж термін, що зустрічається тільки один раз, але приписувати такому терміну в п'ять разів більшу вагу означає приписувати йому занадто велике значення. У векторній моделі класифікації не слід застосовувати незважені і ненормалізовані частоти.

У цій і наступній лекціях розглядаються дві моделі векторної класифікації: Роккіо (Rocchio) і kNN (k nearest neighours — k найближчих сусідів). Класифікація Роккио (раздел 2.2) розділяє векторний простір на області, що оточують центроїди, чи *прототипи*, по одному для кожного класу. Ці центроїди являють собою центри мас всіх ознак у класі. Класифікація Роккіо проста в реалізації й ефективна по швидкості роботи, але неточна, якщо класи далекі від сфер із приблизно однаковими радіусами.

Метод kNN, чи класифікація по *k* найближчих сусідах (k nearest neighbor), описана в наступній лекції, відносить тестову ознаку до класу, якому належать *k* його найближчих сусідів. Метод kNN не вимагає явного

навчання і допускає використання навчальної множини в процесі класифікації без попередньої обробки. Він має велику часову складність порівняно з іншими методами класифікації ознак. Якщо навчальна множина велика, то метод kNN краще справляється з несферичними й іншими складними класами, ніж метод Роккио.

Багато класифікаторів текстів можна розглядати як лінійні класифікатори, тобто класифікатори, засновані на простій лінійній комбінації ознак. Такі класифікатори розбивають простір ознак на області за допомогою поділяючих гіперплощин (decision hyperplanes). Через конфлікт між зсувом і дисперсією, який буде розглянутий пізніше, більш складні нелінійні моделі не завжди краще лінійних. Нелінійні моделі мають багато параметрів, які варто підігнати на обмеженому обсязі даних для навчання, і при цьому для невеликих і зашумленних наборів даних зростає імовірність помилок.

Застосовуючи бінарні класифікатори для рішення задач з декількома класами, ми інтерпретуємо їх або як задачі *однозначної класифікації* (one-of), тобто ознака повинна бути віднесена только до одному з декількох взаємно виключних класів, або как задачі *багатозначної класифікації* (any-of), тобто ознака може бути приписана будь-якій кількості класів. Бінарні класифікатори розв'язують задачу багатозначної класифікації, а их комбінації можна використовувати для розв'язування задач однозначної класифікації.

2.1. Представлення ознак і міри близькості у векторному просторі

Будемо представляти ознаки у виді векторів із простору $\mathbb{R}^{|V|}$. Для того щоб проілюструвати властивості векторів ознак у задачах векторної класифікації, представимо їх у виді точок на площині. Вектори ознак є нормалізованими по довжині одиничними векторами, координати яких лежать на поверхні гіперсфери. Ми можемо розглядати двовимірні площини як проекції поверхні гіперсфери на площину. Відстані між точками на поверхні сфери і між точками на її проекції приблизно збігаються, якщо області на поверхні малі, а проекція обрана корректно (вправа 2.1).

Рішення векторних класифікаторів засновані на понятті відстані, наприклад на обчисленні найближчих сусідів у методі kNN. У цій главі як основну міру близькості обрана евклидово відстань. Між косинусною мірою подібності і відстанню для векторів, нормалізованих по довжині, існує пряма відповідність. У векторній класифікації дуже рідко має значення, як виражається близькість двох ознак — через міру подібності чи відстань.

Однак, крім ознак, у теорії векторної класифікації велику роль грають центроїди, чи усереднені вектори. Центроїди не є нормалізованими за довжиною. Для таких векторів скалярний добуток, косинусна міра подібності й евклидова відстань, у принципі, поводяться по-різному (вправа 2.4). Косінусна міра подібності дорівнює

$$sim(d_1,d_2)=\frac{\left(\vec{V}(d_1),\vec{V}(d_2)\right)}{\left\|\vec{V}(d_1)\right\|\left\|\vec{V}(d_2)\right\|},$$

де $\vec{V}(d_1)$ і $\vec{V}(d_2)$ — вектори ознак документів d_1 і d_2, $\left(\vec{V}(d_1),\vec{V}(d_2)\right)$ — скалярний добуток між цими векторами, $\left\|\vec{V}(d_1)\right\|$ і $\left\|\vec{V}(d_2)\right\|$ — норми векторів $\vec{V}(d_1)$ і $\vec{V}(d_2)$. Позначимо як $\vec{v}(d_1)$ і $\vec{v}(d_2)$ нормалізовані вектори, тобто $\vec{v}(d_1)=\frac{\vec{V}(d_1)}{\left\|\vec{V}(d_1)\right\|}$ і $\vec{v}(d_2)=\frac{\vec{V}(d_2)}{\left\|\vec{V}(d_2)\right\|}$. У такому випадку

$$sim(d_1,d_2)=\left(\vec{v}(d_1),\vec{v}(d_2)\right).$$

В основному при визначенні подібності між ознаками і центроїдом нас будуть цікавити невеликі області, причому чим менше область, тим більше схожими стають властивості всіх трьох мір близькості.

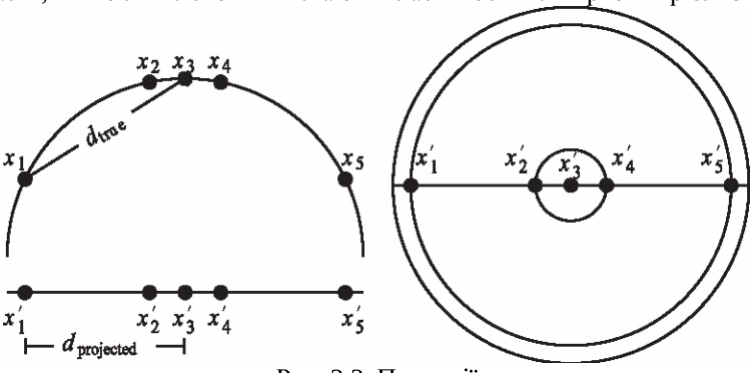

Рис. 2.2. Проекції

Вправа 2.1. Для невеликих областей відстані на поверхні гіперсфери добре апроксимуються відстанями між їхніми проекціями (див. рис. 2.2), оскільки для невеликих кутів $\alpha \approx \sin\alpha$. При якій величині кута обертання $\alpha/\sin\alpha$ дорівнює 1) 1,01, 2) 1,05 і 3) 1,1?

2.2. Метод Роккіо

На рис. 2.1 показано три класи на двовимірній площині: *China*, *UK* і *Kenya*. Ознаки відзначені кружечками, ромбиками і хрестиками. *Поділяючі границі* (decision boundaries) на малюнку обрані таким чином, щоб розділити три класи, але в іншому їхні властивості довільні. Для класифікації нової ознаки, відзначеної на рисунку зірочкою, ми визначаємо область, у яку вона потрапила, а потім клас, що відповідає цій області (у даному випадку це клас *China*). Векторна класифікація зводиться до розробки алгоритму, що обчислює "гарні" межі, де термін "гарні" означає високу точність класифікації на даних, не використаних у ході навчання.

Для векторної класифікації ознак необхідно визначити межі між класами, оскільки саме вони визначають результат класифікації. Ймовірно, найбільш відомим методом визначення цих меж є *метод Роккіо* (Rocchio classification), у якому для ідентифікації меж використовуються центроїди. Центроїд класу обчислюється як усереднений вектор, чи центр мас членів класу.

$$\vec{\mu}(c) = \frac{1}{|D_c|} \sum_{d \in D_c} \vec{v}(d) \qquad (2.1)$$

Тут D_c — множина ознак із простору $\mathbb{D}$, що належать класу c: $D_c = \{d: <d, c> \in \mathbb{D}\}$, $\vec{v}(d)$ — нормалізований вектор ознаки d.

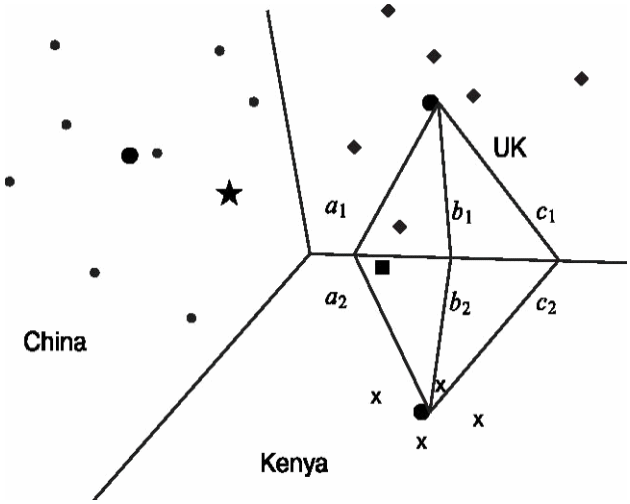

Рис. 2.3. Класифікація Роккио

Межа між двома класами в методі Роккио являє собою множину точок, рівновіддалених від двох центроїдів. Наприклад, $|a_1| = |a_2|$, $|b_1| = |b_2|$ і $|c_1| = |c_2|$. Ця множина точок завжди утворює лінію. Узагальненням цієї лінії в M-вимірному просторі є гіперплощина, що являє собою множину точок $\vec{x}$, що задовольняють умові

$$\vec{w}^T \vec{x} = b \ . \qquad (2.2)$$

Тут $\vec{w}$ — M- мірний *вектор нормалі* (normal vector) до гіперплощини, а b — константа. Це визначення гіперплощин охоплює лінії (будь-яка лінія на двовимірній площині описується рівнянням $w_1x_1 + w_2x_2 = b$) і двовимірні площини (будь-яка площина в тривимірному просторі визначається рівнянням $w_1x_1 + w_2x_2 + w_3x_3 = b$). Лінія розділяє площину на дві напівплощини, площина розділяє тривимірний простір на два підпростори і гіперплощина розділяє простір більш високої розмірності на два підпростори.

Таким чином, межами областей класів у методі Роккио є гіперплощини. Правило класифікації в алгоритмі Роккио полягає у визначенні області, у яку потрапляє точка. Це еквівалентно пошуку центроїда $\vec{\mu}(c)$, до якого точка лежить ближче, ніж до інших центроідів, і приписуванню цієї точці до класу c. Як приклад розглянемо зірочку, яка належить області *China*, тому алгоритм Роккио відносить її до класу *China*. Псевдокод алгоритму Роккио представлений нижче.

TrainRocchio($\mathbb{C}$, $\mathbb{D}$)

1 **for each** $c_j \in \mathbb{C}$

2 **do** $D_j \leftarrow \{d: \langle d, c_j\rangle \in \mathbb{D}\}$

3 $\vec{\mu}_j \leftarrow \frac{1}{|D_j|}\sum_{d\in D_j}\vec{v}(d)$

4 **return** $\{\vec{\mu}_1, \vec{\mu}_2, ..., \vec{\mu}_j\}$

ApplyRocchio($\mathbb{C}$, $\mathbb{D}$)

1 **return** $\arg\min_j |\vec{\mu}_j - \vec{v}(d)|$

Рис. 2.4. Класифікація Роккио. Навчання і тестування

Приклад 2.1. У табл. 2.1. приведено векторні представлення з вагами tf–idf п'яти ознак, отриманих за допомогою формули $(1 + \log_{10}\mathrm{tf}_{t,d})\log_{10}(4/\mathrm{df}_t)$, якщо $\mathrm{tf}_{t,d} > 0$. Центроїдами класів є два вектори: $\vec{\mu}_c = \frac{1}{3}\left(\vec{d}_1 + \vec{d}_2 + \vec{d}_3\right)$ і $\vec{\mu}_{\bar{c}} = \frac{1}{1}\vec{d}_4$. Відстані тестової ознаки від центроїдів дорівнюють $|\vec{\mu}_c - \vec{d}_5| \approx 1{,}15$ і $|\vec{\mu}_{\bar{c}} - \vec{d}_5| \approx 0{,}0$. Отже, алгоритм Роккіо віднесе ознаку $\vec{d}_5$ до класу *c*.

Поділяюча гіперплощина в цьому випадку описується наступними параметрами.

$\vec{w} \approx (0 \quad -0{,}71 \quad -0{,}71 \quad 1/3 \quad 1/3 \quad 1/3)^T$,

$b = -1/3$

Обчислення вектора $\vec{w}$ і константи *b* описано у вправі 2.15 (див. наступну лекцію). Легко перевірити, що ця гіперплощина є шуканою поділяючою границею: $\vec{w}^T\vec{d}_1 \approx 0\cdot 0 + (-0{,}71)\cdot 0 + (-0{,}71)\cdot 0 + \frac{1}{3}\cdot 0 + \frac{1}{3}\cdot 1{,}0 + \frac{1}{3}\cdot 0 = \frac{1}{3} > b$ (аналогічно $\vec{w}^T\vec{d}_i > b$ для $2 \le i \le 3$) і $\vec{w}^T\vec{d}_4 = -1 < b$. Таким чином, ознаки з класу *c* лежать вище напівплощини ($\vec{w}^T\vec{d} > b$), а ознаки з класу $\bar{c}$ — нижче напівплощини ($\vec{w}^T\vec{d} < b$).

Таблиця 2.1. Вектори і центроїди класів

	Ваги термінів					
Вектор	Chinese	Japan	Tokyo	Macao	Beijing	Shanghai
$\vec{d}_1$	0	0	0	0	1,0	0
$\vec{d}_2$	0	0	0	0	0	1,0
$\vec{d}_3$	0	0	0	1,0	0	0
$\vec{d}_4$	0	0,71	0,71	0	0	0
$\vec{d}_5$	0	0,71	0,71	0	0	0
$\vec{\mu}_C$	0	0	0	0,33	0,33	0,33
$\vec{\mu}_{\bar{C}}$	0	0,71	0,71	0	0	0

Критерієм класифікації є евклидова відстань. Як альтернативу можна використовувати косинусну міру подібності.

$$\textit{Документ } d \textit{ належить до класу } c = \arg\max_{c'} \cos\left(\vec{\mu}(c'), \vec{v}(d)\right).$$

Как зазначено в розділі 2.1, два критерії класифікації іноді приводять до різних результатів. Тут ми навели варіант алгоритму Роккио, заснований на використанні евклидової відстані, оскільки вона підкреслює тісний зв'язок з методом кластеризации *K* середніх (K-means clustering), що буде описаний пізніше.

Кромі відповідності гіпотезі компактності, класи в класифікації Роккио повинні мати приблизну форму сфер із приблизно однаковими радіусами. На рис. 2.3 чорний квадрат безпосередньо під межею між класами *UK* і *Kenya* більше підходить для класу *UK*, оскільки клас *UK* має більш широкий розкид, ніж клас *Kenya*. Однак алгоритм Роккіо відносить його до класу *Kenya*, тому що при класифікації він ігнорує особливості розподілу точок у класі і використовує лише відстані до центроїду.

Припущення про сферичність на рис. 2.5 також не виконується. Клас “*a*” неможливо добре описати за допомогою єдиного прототипу, оскільки він складається з двох кластерів. Алгоритм Роккіо часто невірно розпізнає *мультимодальні класи* такого типу. Прикладом мультимодального класу в задачах текстової класифікації є країни на зразок Бірми, що з 1989 року стала називатися Мьянмой. Два кластери до і після зміни назви не обов'язково розташовані близько один від одного у векторному просторі.

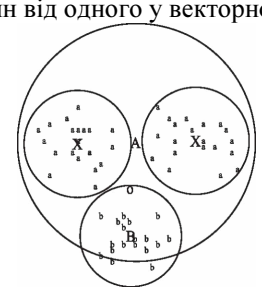

Рис. 2.5. Мультимодальний клас “a” складається з двох різних кластерів (невеликі кола у верхній частині малюнка з центрами в точках X). Класифікація Роккио неправильно віднесе термін “o” до класу термінів “a”, тому що він ближче до центроїду A класу “a”, а не до центроїду B класу “b”

Ще одним прикладом, у якому класи рідко являють собою сфери з однаковими радіусами, є задача бінарної класифікації. Більшість бінарних класифікаторів розрізняють клас на зразок *China*, що займає невелику область у просторі, і його доповнення, розкидане по всьому просторі. Припущення про рівність радіусів привело б до великої кількості хибно-позитивних рішень. Отже, для більшості задач бінарної класифікації необхідно уточнити вирішальне правило.

Документ d належить класу c тоді і лише тоді, коли $\left|\vec{\mu}(c)-\vec{v}(d)\right|<\left|\vec{\mu}(\overline{c})-\vec{v}(d)\right|-b$

Тут b — позитивна константа. Центроїд “негативних” ознак можна не використовувати взагалі, тому вирішальне правило можна спростити до $\left|\vec{\mu}(c)-\vec{v}(d)\right|<b'$, де b' — позитивна константа.

У табл. 2.2. наведено оцінки часової складності класифікації Роккіо. Складність підсумовування всіх ознак при обчисленні вектора суми дорівнює $\Theta(|\mathbb{D}|L_{ave})$, а не $\Theta(|\mathbb{D}||V|)$, оскільки на суму впливають тільки ненульові елементи. Складність розподілу кожної суми векторів на розмір класу при обчисленні центроїду дорівнює $\Theta(|V|)$. У цілому час навчання лінійно залежить від розміру колекції.

Таблиця 2.2. Оцінки часової складності навчання і тестування для класифікації Роккіо (L_{ave} — середня кількість лексем в ознаці, L_a і M_a — кількість лексем і типів у тестовій ознаці відповідно; часова складність обчислення евклидової відстані між центроїдами класів і ознак дорівнює $\Theta(|\mathbb{C}|M_a)$)

Операція	*Часова складність*
Навчання	$\Theta(|\mathbb{D}|L_{ave}+|\mathbb{C}||V|)$
Тестування	$\Theta(L_a+|\mathbb{C}|M_a)=\Theta(|\mathbb{C}|M_a)$

Вправа 2.2. Покажіть, что класифікація Роккіо може приписати ознаці мітку, що відрізняється від его мітки в навчальному класі.

У наступній лекції ми опишемо інший метод векторної класифікації kNN, що краще працює з несферичними і незв'язними класами, а також з класами, що маютьінші “неправильності”.

Література

1. Маннинг К.Д., Рагхаван П., Шютце Х. Введение в информационный поиск.— М.: Изд-во “Вильямс”, 2011.