Лекція 6

Лінійні і нелінійні класифікатори [1]

У цьому розділі ми покажемо, що два методи навчання — наївний байєсівський і Роккіо — є прикладами лінійних класифікаторів, ймовірно, найбільш важливої групи класифікаторів текстів, і протиставимо їм нелінійні класифікатори. Для простоти розглянемо лише бінарні класифікатори і будемо називати *лінійним класифікатором* бінарний класифікатор, що приймає рішення про приналежність документа класу за допомогою порівняння лінійної комбінації ознак з визначеним граничним значенням.

На площині лінійним класифікатором є лінія (див. рис. 6.1). Загальний функціональний вигляд лінійних класифікаторів такий: $w_1x_1 + w_2x_2 = b$. Правило класифікації за допомогою лінійного класифікатора полягає в тім, що документ відноситься до класу c, якщо $w_1x_1 + w_2x_2 > b$, і до класу $\overline{c}$, якщо $w_1x_1 + w_2x_2 \leq b$. Тут $(x_1, x_2)^T$ — двовимірне векторне представлення документа, а $(w_1, w_2)^T$ — вектор параметрів, що разом з числом b визначають поділяючу межу.

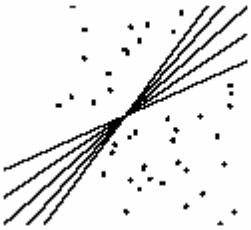

Рис. 6.1. Два лінійно роздільних класи можна відокремити один від одного за допомогою нескінченно великої кількості гіперплощин

Двовимірний лінійний класифікатор можна узагальнити для просторів більш високої розмірності, визначивши гіперплощину за допомогою формули

$$\vec{w}^T\vec{x} = b \tag{6.1}$$

Критерій рішення має наступний вид: якщо $\vec{w}^T\vec{x} > b$, то клас c, а якщо $\vec{w}^T\vec{x} \leq b$, то клас $\overline{c}$. Гіперплощина, використовувана в лінійному класифікаторі, називається *поділяючою гіперплощиною*.

Відповідний алгоритм лінійної класифікації в M-вимірному просторі показаний на рис. 6.2. На перший погляд, лінійна класифікація виглядає тривіальною. Однак навчання лінійного класифікатора пов'язане із значними складностями, наприклад, при визначенні параметрів $\vec{w}$ і $\vec{b}$ за навчальною множиною. Якщо як показник якості методу навчання використовувати ефективність навченого лінійного класифікатора на нових даних, то одні методи обчислюють набагато більш точні параметри, ніж інші.

ApplyLinearClassifier($\vec{w}$, $\vec{b}$, $\vec{x}$)

1 $score \leftarrow \sum_{i=1}^{M} w_i x_i$
2 **if** $score > b$
3 **then return** 1
4 **else return** 0

Рис. 6.2. Алгоритм лінійної класифікації

Тепер покажемо, що алгоритм Роккіо і наївний байєсівський метод є лінійними класифікаторами. Помітимо, що в алгоритмі Роккіо вектор $\vec{x}$ лежить на поділяючій межі, якщо він лежить на однаковій відстані від центроїдів двох класів.

$$\left|\vec{\mu}(c_1) - \vec{x}\right| = \left|\vec{\mu}(c_2) - \vec{x}\right| \tag{6.2}$$

Прості алгебраїчні обчислення показують, що це відповідає лінійному класифікатору з вектором нормалі $\vec{w} = \vec{\mu}(c_1) - \vec{\mu}(c_2)$ і $b = 0{,}5\left(\left|\vec{\mu}(c_1)\right|^2 - \left|\vec{\mu}(c_2)\right|^2\right)$.

Лінійність наївного байесівського методу випливає з його вирішального правила, відповідно до якого вибирається клас c з найбільшим значенням $\hat{P}(c|d)$.

$$\hat{P}(c|d) \propto \hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(t_k|c)$$

Тут n_d — кількість лексем у документі, що входять у лексикон. Позначивши доповнення через $\overline{c}$, одержимо логарифм відношення шансів.

$$\log\frac{\hat{P}(c|d)}{\hat{P}(\overline{c}|d)} = \log\frac{\hat{P}(c)}{\hat{P}(\overline{c})} + \sum_{1\le k\le n_d}\log\frac{\hat{P}(t_k|c)}{\hat{P}(t_k|\overline{c})} \tag{6.3}$$

Ми вибираємо клас *c*, якщо відношення імовірностей більше одиниці, що еквівалентно, якщо логарифм відношення імовірностей (шансу) більше нуля. Легко бачити, що рівність (6.3) — це варіант рівності (6.1), якщо $w_i = \log\left[\hat{P}(t_i|c)/\hat{P}(t_i|\overline{c})\right]$, x_i — кількість входжень терміна t_i у документ *d* і $b = -\log\left[\hat{P}(c)/\hat{P}(\overline{c})\right]$. Тут індекс *i*, що змінюється від 1 до *M*, нумерує терміни лексикона (а не координати в документі *d* на відміну від індексу *k*, а $\vec{x}$ і $\vec{w}$ — *M*- мірні вектори). Таким чином, у просторі логарифмів наївний байєсівський метод є лінійним класифікатором.

Приклад 6.1. У табл. 6.1 описаний лінійний класифікатор для категорії *interest* у колекції Reuters-21578. Документ $\vec{d}_1$ "rate discount dlrs world" відноситься до класу *interest*, оскільки $\vec{w}^T\vec{d}_1 = 0{,}67\cdot 1 + (-0{,}71)\cdot 1 + (-0{,}35)\cdot 1 = 0{,}07 > 0 = b$. Документ $\vec{d}_2$ "prime dlrs" відноситься до додаткового класу (не *interest*), оскільки $\vec{w}^T\vec{d}_2 = -0{,}01 \le b$. Для простоти ми використовуємо бінарне векторне представлення документа: якщо термін зустрічається в документі, вектор містить одиницю, а якщо немає — нуль.

Таблиця 6.1. Лінійний класифікатор. Терміни t_i і параметри w_i лінійного класифікатора для класу *interest* (у змісті interest rate — процентна ставка) у колекції Reuters-21578. Граничне значення $b = 0$. Терміни типу dlr і world мають негативні ваги, оскільки є індикаторами конкуруючого класу *currency* (валюта)

t_i	w_i	d_{1i}	d_{2i}	t_i	w_i	d_{1i}	d_{2i}
prime	0,70	0	1	dlrs	–0,71	1	1
rate	0,67	1	0	world	–0,35	1	0
interest	0,63	0	0	sees	–0,33	0	0
rates	0,60	0	0	year	–0,25	0	0
discount	0,46	1	0	group	–0,24	0	0
bundesbank	0,43	0	0	dlr	–0,24	0	0

Лінійна задача за визначенням характеризується тим, що базові розподіли $P(d|c)$ і $P(d|\overline{c})$ двох класів розділені лінією. Ця лінія називається *межею між класами*. Межа між класами є "справжньою" і відрізняється від поділяючої межі, яку метод навчання обчислює як її наближення.

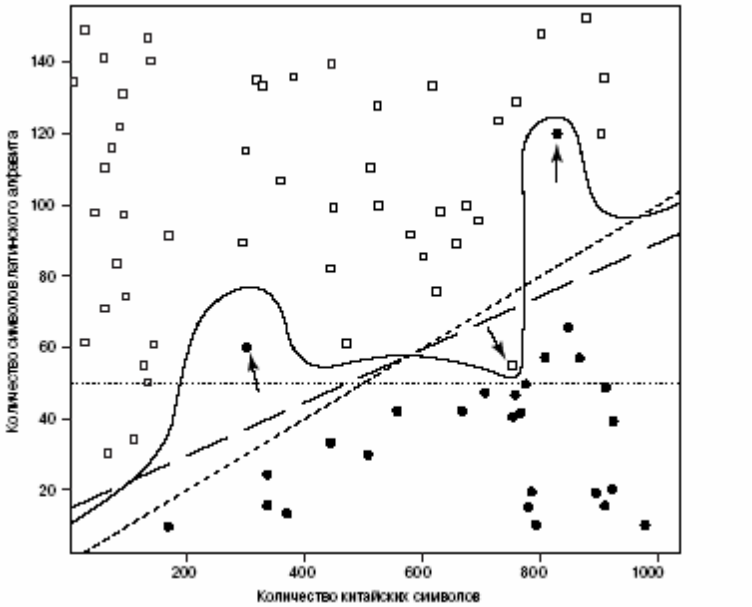

Рис. 6.3. Лінійна задача із шумом. У цьому гіпотетичному сценарії класифікації веб-сторінок чорними кружечками відзначені веб-сторінки винятково китайською мовою, а квадратиками — веб-сторінки на суміші китайської й англійської мов. Ці два класи розділені прямою лінією (пунктирною), хоча існують три шумових документи (відзначені стрілочками)

Як це зазвичай і буває в класифікації текстів, існують *шумові документи*. Ці документи недостатньо добре відповідають загальному розподілу класів. Раніше ми назвали шумовою ознакою таку ознаку, включення якої в представлення документа в середньому підвищує помилку класифікації. Аналогічно шумовим називається такий документ, включення якого в навчальну множину призводить до збільшення помилки класифікації. Інтуїтивно зрозуміло, що базовий розподіл розділяє простір представлення документів в основному на області з однорідними мітками класів. Документ, що не відповідає домінантному класу у визначеній області, є шумовим.

Шумові документи — це одна з причин, з яких навчання лінійного класифікатора являє собою непросту задачу. Якщо при виборі поділяючої гіперплощини приділити занадто велику увагу шумовим документам, то класифікатор стане неточним на нових даних. І що ще більш важливо, зазвичай важко визначити, які документи є шумовими і, отже, потенційно можуть знизити точність класифікації.

Якщо існує гіперплощина, що ідеально точно розділяє два класи, то такі класи називаються *лінійно роздільними*. Насправді, якщо існує лінійна раздільність, то кількість лінійних роздільників нескінченна.

Існує ще одна проблема, зв'язана з навчанням лінійного класифікатора. Якщо задача є лінійно роздільною, то необхідно сформулювати критерій для вибору поділяючої гіперплощини серед багатьох гіперплощин, що ідеально розділяють навчальні дані. У принципі, одні з таких гіперплощин будуть добре розділяти нові дані, а інші — ні.

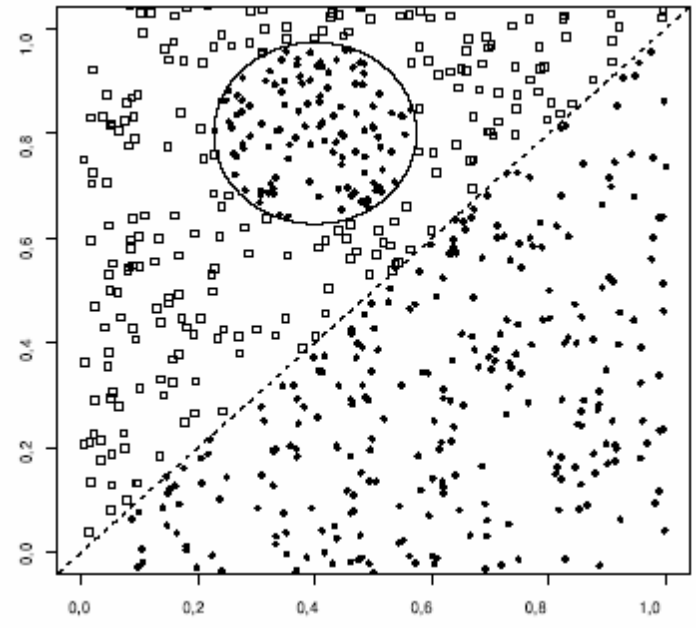

Рис. 6.4. Нелінійна задача

Прикладом нелінійного класифікатора є метод kNN. Нелінійність класифікатора kNN стає інтуїтивно ясною, якщо подивитися на рис. 6.4. Поділяюча межа в методі kNN складається з локально лінійних сегментів, але в цілому вона має складну форму, що не збігається ані з лінією на площині, ані з гіперплощиною в просторах більш високої розмірності.

Розглянемо приклад нелінійної задачі, у якій між розподілами $P(d|c)$ і $P(d|\overline{c})$ немає гарного лінійного поділу, оскільки, наприклад, у лівій напівплощині є “круглий” анклав. Лінійні класифікатори невірно класифікують цей анклав, у той час як метод kNN вирішує такі задачі з високою точністю, якщо навчальна множина достатньо велика.

Якщо проблема носить нелінійний характер, а її межі між класами неможливо добре апроксимувати за допомогою гіперплощин, то нелінійні класифікатори звичайно виявляються точніше лінійних. Якщо ж задача є лінійною, то краще використовувати більш простий лінійний класифікатор.

Вправа 6.1. Доведіть, що кількість лінійних роздільників двох класів або нескінченна, або дорівнює нулю.

Класифікація з декількома класами

Бінарні лінійні класифікатори можна розширити на варіант $J > 2$ класів. Вибір методу класифікації в цьому випадку залежить від того, чи є класи взаємовиключними.

Класифікація для класів, що не є взаємовиключними, називається *багатозначною*. У цьому випадку документ може належати декільком класам одночасно, одному класу чи не належати жодному класу. Рішення щодо одного класу не виключає рішень щодо інших класів. Іноді говорять, що класи *не залежать* друг від друга, але це неправильно; класи рідко є статистично незалежними. У термінах формальної постановки задачі класифікації можна сказати, що в задачі

багатозначної класифікації відбувається навчання J різних класифікаторів γ_j, причому кожний із класифікаторів повертає або мітку класу c_j, або мітку класу $\overline{c}_j$, тобто $\gamma_j(d) \in \{c_j, \overline{c}_j\}$.

Рішення задачі багатозначної класифікації за допомогою лінійних класифікаторів очевидне.

1. Будуємо класифікатори для кожного класу, при цьому навчальна множина складається з набору документів, що належать класу (позитивні мітки), і його доповненню (негативні мітки).
2. Маючи тестовий документ, застосовуємо до нього кожний класифікатор окремо. Рішення одного класифікатора не впливає на рішення іншого.

Ще одним різновидом класифікації з декількома класами є *однозначна класифікація*. У цьому випадку класи не перетинаються. Кожен документ повинний належати тільки одному з класів. Однозначна класифікація називається також *мультиноміальною*, *політомічною*, *багатокласовою* чи *класифікацією з однією міткою*. З формальної точки зору в цій класифікації існує єдина функція класифікації γ, областю значень якої є простір $\mathbb{C}$, тобто $\gamma(d) \in \{c_1, ..., c\}$. Наприклад, метод kNN є (нелінійним) однозначним класифікатором.

Істинно однозначні задачі класифікації текстів зустрічаються рідше, ніж багатозначні. Документи з класів *UK*, *China*, *poultry* і *coffee* можуть бути релевантними декільком темам одновременно, наприклад якщо прем'єр-міністр Великобританії (клас *UK*) відвідав Китай (клас *China*), щоб провести переговори про торгівлю *кавою* (клас *coffee*) і *домашнім птахом* (клас *poultry*).

Втім, ми часто будемо приймати припущення про однозначну класифікацію, навіть якщо класи насправді не є взаємовиключними. Для визначення мови документа припущення про однозначну класифікацію є цілком обґрунтованим, оскільки більшість документів написана тільки однією мовою. У таких випадках накладення умови однозначності може підвищити ефективність класифікації, оскільки при цьому виключаються помилки, що виникають через те, що документ приписується декільком класам чи жодному класу взагалі.

Часто J гіперплощин не розділяють простір $\mathbb{R}^{|V|}$ на J диз'юнктних областей. Отже, для рішення задачі однозначної класифікації за допомогою бінарних лінійних класифікаторів ми повинні застосовувати їхню комбінацію. Простіше всього ранжувати класи, а потім вибрати клас з найбільшим рангом. З геометричної точки зору ранжування зводиться до обчислення відстаней від J поділяючих гіперплощин. Імовірність неправильно класифікувати документ, що лежить ближче до межі класу, вище, тому чим далі документ від межі, тим вище імовірність, що він класифікований правильно. Як альтернативу можна безпосередньо обчислити довірчу оцінку для рангу класу, наприклад імовірність приналежності класу. Цей алгоритм однозначної класифікації за допомогою лінійних класифікаторів можна описати в такий спосіб.

1. Будуємо класифікатор для кожного класу, причому навчальна множина складається з набору документів, що належать класу (позитивні мітки), і його доповнення (негативні мітки).
2. До заданого тестового документа застосовуємо кожен класифікатор окремо.
3. Відносимо тестовий документ до класу, що має
   - максимальний ранг чи
   - максимальний довірчий рівень, чи
   - максимальну імовірність.

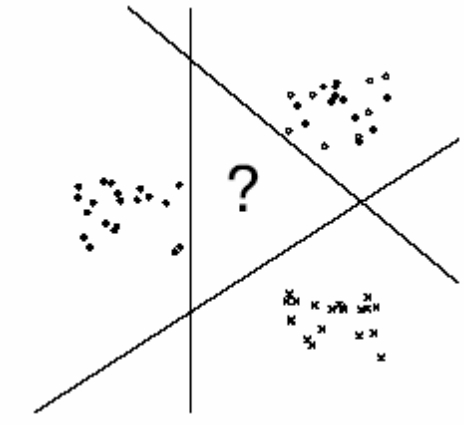

Рис. 6.5. J гіперплощин не розділяють простір $\mathbb{R}^{|V|}$ на J диз'юнктних областей

Важливим інструментом аналізу ефективності класифікатора при $J > 2$ класах є *матриця неточностей*. Для кожної пари класів $<c_1, c_2>$ вона показує, скільки документів із класу c_1 помилково віднесені до класу c_2. У табл. 6.2 показані результати розв'язування задачі, у якій класифікатор повинний був відрізнити три класи, присвячених фінансам

(*money-fx, trade* і *interest*), від трьох класів сільськогосподарської тематики (*wheat, corn* і *grain*), але зробив багато помилок усередині цих двох груп. Матриця неточностей допомагає виявити можливості для підвищення правильності роботи системи. Наприклад, щоб усунути другу по величині помилку в табл. 6.2, можна спробувати ввести ознаки, що відрізняють документи класу *wheat* від документів класу *grain*.

Таблиця 6.2. Матриця неточностей для колекції Reuters-21578. Наприклад, чотирнадцять документів із класу *grain* были помилково віднесені до класу *wheat* (Picca et al., 2006)

Щирий клас \ Приписаний клас	*money-fx*	*trade*	*Interest*	*Wheat*	*corn*	*grain*
money-fx	95	0	10	0	0	0
trade	1	1	90	0	1	0
interest	13	0	0	0	0	0
wheat	0	0	1	34	3	7
corn	1	0	2	13	26	5
grain	0	0	2	14	5	10

Вправа 6.2. Створіть навчальну множину, що складається з 300 документів, по 100 для кожної мови (наприклад, англійської, французької й іспанської). Точно так само створіть тестову множину, додавши до неї 100 документів на четвертій мові. Проведіть навчання однозначного і багатозначного класифікаторів на цій навчальній множині й оцініть їх за тестовою множиною. Чи існують які-небудь цікаві розходження між поводженням цих класифікаторів при розв'язуванні поставленої задачі?

Компроміс між зсувом і дисперсією

Нелінійні класифікатори потужніші ніж лінійні. Для деяких задач існують нелінійні класифікатори з нульовою помилкою класифікації, але немає аналогічних лінійних класифікаторів. Чи значить це, що для досягнення оптимальної ефективності статистичної класифікації текстів завжди варто використовувати нелінійні класифікатори?

Для відповіді на це питання в даному розділі описується компроміс між зсувом і дисперсією — одна з найбільш важливих концепцій у теорії машинного навчання. Цей компроміс допомагає пояснити, чому не існує оптимального методу навчання. Отже, вибір придатного методу навчання є невід'ємною частиною розв'язання задачі класифікації текстів.

Протягом цього розділу ми використовуємо лінійні і нелінійні класифікатори як прототипи "менш потужного" і "більш потужного" методів навчання відповідно. Такий підхід є спрощенням з кількох причин. По-перше, багато нелінійних моделей в окремих випадках зводяться до лінійних. Наприклад, нелінійний метод навчання, такий як kNN, у деяких випадках породжує лінійний класифікатор. По-друге, існують нелінійні моделі, що простіше лінійних. Наприклад, квадратний багаточлен із двома параметрами простіше, ніж 10 000-вимірний лінійний класифікатор. По-третє, складність навчання насправді не є властивістю класифікатора, оскільки існує багато факторів навчання (наприклад, вибір ознак, регуляризація й обмеження), що підвищують або знижують точність методу навчання незалежно від типу класифікатора, тобто остаточний результат навчання залежить не тільки від того, чи є класифікатор лінійним або нелінійним. У цьому розділі лінійні і нелінійні класифікатори служать як приклад менш потужного і більш потужного методів навчання в задачах класифікації текстів.

Спочатку слід уточнити мету класифікації текстів. Як говорилося раніше, ми сказали, що хочемо мінімізувати помилку класифікації на тестовій множини. При цьому було зроблено неявне припущення, що навчальні і тестові документи генеруються відповідно за тим самим розподілом. Позначимо цей розподіл як $P(\langle d, c\rangle)$, де d — це документ, а c — мітка його класу.

У цьому розділі замість кількості правильно класифікованих документів (чи, що еквівалентно, рівня помилок на тестових документах) як показник якості класифікації ми будемо використовувати оцінку, що враховує неперебірну невизначеність класифікації. У багатьох задачах класифікації текстів те саме представлення документа можна одержати з документів, що належать різним класам. Це відбувається тому, що документи з різних класів можуть відображатися в те саме представлення документа. Наприклад, документи, що складаються з пропозицій *China sues France* (Китай подав судовий позов до Франції) і *France sues China* (Франція подала судовий позов до Китаю), у моделі "мішка слів" відображаються в те саме представлення d' = {China, France, sues}. Однак класу c' = *юридичні дії, початі Францією*, що може бути визначений, наприклад, за запитом якого-небудь фахівця з міжнародної торгівлі, релевантний тільки другий документ.

Для спрощення обчислень при оцінці класифікатора ми не будемо підраховувати кількість помилок на тестовій множини, а звернемо увагу на те, наскільки добре класифікатор оцінює умовну імовірність $P(c|d)$ для документа з класу. У наведеному вище прикладі $P(c'|d') = 0{,}5$.

Мета класифікації текстів тепер полягає в пошуку класифікатора γ, що після усереднення по всіх документах d гарантував би оцінку $\gamma(d)$, як можна більш близьку до імовірності $P(c|d)$. Цей показник ми будемо вимірювати за допомогою середньоквадратичної помилки.

$$MSE(\gamma) = E_d\left[\gamma(d) - P(c|d)\right]^2 \tag{6.4}$$

Тут E_d — математичне сподівання $P(d)$. Середньоквадратична помилка приписує часткову довіру рішенням класифікатора γ, що близькі до правильних, навіть якщо вони не цілком збігаються з ними.

Класифікатор γ називається *оптимальним* щодо розподілу $P(<d, c>)$, якщо він мінімізує середньоквадратичну помилку MSE(γ).

Мінімізація середньоквадратичної помилки — бажана властивість *класифікатора*. Крім того, нам необхідний критерій для *методу навчання*. Нагадаємо, що методом навчання Г називається функція, аргументом якої є розмічена навчальна множина $\mathbb{D}$, а значенням — класифікатор γ.

Нашою метою є такий метод навчання Г, що у середньому по всім навчальним множинам створює класифікатор γ з мінімальною середньоквадратичною помилкою (MSE). Цю вимогу можна формалізувати за допомогою *помилки навчання*.

$$\text{Помилка навчання } (\Gamma) = E_{\mathbb{D}}[\text{MSE}(\Gamma(\mathbb{D}))] \tag{6.5}$$

Тут $E_{\mathbb{D}}$ — математичне сподівання на розмічених навчальних множинах. Для простоти припустимо, що навчальні множини мають фіксований розмір, тоді розподіл $P(<d, c>)$ визначає розподіл $P(\mathbb{D})$ по навчальних множинах.

Вибір методу навчання в задачі статистичної класифікації текстів можна здійснювати на основі помилки навчання. Метод навчання Г є *оптимальним* для розподілу $P(\mathbb{D})$, якщо він мінімізує помилку навчання.

$$\begin{aligned} E[x-\alpha]^2 &= Ex^2 - 2Ex\alpha + \alpha^2 = (Ex)^2 - 2Ex\alpha + \alpha^2 + Ex^2 - 2(Ex)^2 + (Ex)^2 = \\ &= [Ex-\alpha]^2 - E2x(Ex) + E(Ex)^2 = [Ex-\alpha]^2 + E[x-Ex]^2 \end{aligned} \tag{6.6}$$

$$E_{\mathbb{D}}E_d[\Gamma_{\mathbb{D}}(d) - P(c|d)]^2 = E_dE_{\mathbb{D}}[\Gamma_{\mathbb{D}}(d) - P(c|d)]^2 = E_d[[E_{\mathbb{D}}\Gamma_{\mathbb{D}}(d) - P(c|d)]^2 + E_{\mathbb{D}}\,[\Gamma_{\mathbb{D}}(d) - E_{\mathbb{D}}\Gamma_{\mathbb{D}}(d)]^2] \tag{6.7}$$

При виводі формули (6.7) ми ввели позначення $\alpha = P(c|d)$ і $x = \Gamma_{\mathbb{D}}(d)$ у формулі (6.7)

Для спрощення запису замінимо символи $\Gamma(\mathbb{D})$ позначенням $\Gamma_{\mathbb{D}}$. Тоді формулу (6.5) можна переписати в такий спосіб.

Помилка навчання $(\Gamma) = E_{\mathbb{D}}[\text{MSE}(\Gamma(\mathbb{D}))] =$

$$= E_{\mathbb{D}}E_d[\Gamma_{\mathbb{D}}(d) - P(c|d)]^2 = \tag{6.8}$$

$$= E_d[\text{зсув}(\Gamma, d) + \text{дисперсія}(\Gamma, d)], \tag{6.9}$$

$$\text{зсув}(\Gamma, d) = [P(c|d) - E_{\mathbb{D}}\Gamma_{\mathbb{D}}(d)]^2, \tag{6.10}$$

$$\text{дисперсія}(\Gamma, d) = E_{\mathbb{D}}[\Gamma_{\mathbb{D}}(d) - E_{\mathbb{D}}\Gamma_{\mathbb{D}}(d)]^2 \tag{6.11}$$

Тут еквівалентність формул (6.8) і (6.9) випливає з формули (6.7). Зверніть увагу на те, що документ d і множина $\mathbb{D}$ не залежать друг від друга. У принципі, для випадкового документа d і випадкової навчальної множини $\mathbb{D}$ множина $\mathbb{D}$ може не містити позначений екземпляр документа d.

Зсув (bias) — це квадрат різниці між істинною умовною імовірністю $P(c|d)$ того, що документ d належить класу c, і $\Gamma_{\mathbb{D}}(d)$ — прогнозною оцінкою, отриманою за допомогою навченого класифікатора й усередненою по всіх навчальних множинах. Якщо метод навчання породжує погані класифікатори, то зсув великий. Якщо ж 1) класифікатори гарні чи 2) різні навчальні множини породжують помилки на різних документах чи 3) різні навчальні множини породжують позитивні і негативні помилки на тих самих документах, але їх середнє дорівнює нулю, то зсув малий. Якщо виконується одна з перерахованих вище умов, то $E_{\mathbb{D}}\Gamma_{\mathbb{D}}(d)$, тобто математичне сподівання по всіх навчальних множинах, близьке до імовірності $P(c|d)$.

Лінійні методи, такі як метод Роккіо і наївний байєсівський метод, на нелінійних задачах мають великий зсув, оскільки вони можуть моделювати тільки один різновид меж між класами — гіперплощини. Якщо модель $P<d, c>$ складну нелінійну межу між класами, то зсув у рівності (6.9) буде великим, оскільки велика кількість точок буде класифікована неправильно. Наприклад, круглий анклав у попередньому прикладі не враховується лінійною моделлю і може породити велику кількість помилок.

Зсув можна інтерпретувати як результат знань про предметну область (чи їх недостачу), убудованих у класифікатор. Якщо відомо, що істинна межа між класами є лінійною, то, імовірніше за все, саме метод навчання, що породжує лінійні класифікатори, буде точнішим, ніж нелінійний метод. Однак якщо істинні межі між класами є нелінійними і ми помилково вибрали лінійний класифікатор, то середня точність класифікації буде низкою.

Нелінійні методи, такі як kNN, мають невеликий зсув. Межі між класами в методі kNN дуже мінливі і залежать від розподілу документів у навчальній множині. У результаті кожен документ має шанс бути класифікованим правильно на деякій навчальній множині. Отже, середня оцінка $E_{\mathbb{D}}\Gamma_{\mathbb{D}}(d)$ близька до імовірності $P(c|d)$, а зсув менше, ніж у лінійного методу навчання.

Дисперсія (variance) — це варіація оцінок лінійних класифікаторів, квадрат різниць між оцінкою $\Gamma_{\mathbb{D}}(d)$ і її середнім значенням $E_{\mathbb{D}}\Gamma_{\mathbb{D}}(d)$, усереднений по всіх навчальних множинах. Якщо різні навчальні множини $\mathbb{D}$ породжують зовсім різні класифікатори $\Gamma_{\mathbb{D}}$, то дисперсія велика. Якщо ж навчальна множина мало впливає на класифікацію $\Gamma_{\mathbb{D}}$, незалежно від того, правильна вона чи неправильна, то дисперсія мала. Дисперсія оцінює, наскільки суперечливими є рішення, незалежно від того, правильні вони чи неправильні.

Лінійні методи навчання мають невелику дисперсію, оскільки більшість випадково обраних навчальних множин породжує близькі поділяючі гіперплощини. Поділяючі лінії, породжені лінійними методами навчання, мало відхиляються від істинних меж між класами при зміні навчальних множин, але на класифікацію переважної більшості документів (за винятком документів, близьких до меж між класами) це не впливає. Анклав у формі кола буде постійно класифікуватися неправильно.

Нелінійні методи, такі як kNN, мають високу дисперсію. Метод kNN може моделювати дуже складні межі між двома класами. З цієї причини він дуже чуттєвий до шумових документів. У результаті дисперсія в рівності (6.9) для методу kNN виявляється великою. Тестові документи іноді класифікуються невірно (якщо вони знаходяться недалеко від шумових документів), а іноді вірно (якщо в їх околах немає шумових документів). Це призводить до великої варіації результатів при переході від однієї навчальної множини до іншої.

Методи навчання з великою дисперсією піддаються *перенавчанню* по навчальних вибірках. Мета класифікації — настроїти класифікатор на навчальних вибірках, щоб врахувати основні властивості базового розподілу $P(<d, c>)$. Однак якщо відбувається перенавчання, то результат містить у собі шумову інформацію. Перенавчання збільшує середньоквадратичну помилку і часто породжує проблеми при використанні методів навчання з великою дисперсією.

Дисперсію можна також інтерпретувати як *міру складності моделі*, чи *ємність запам'ятовування* методу навчання, тобто наскільки добре він запам'ятовує характеристики навчальної множини, а потім застосовує їх до нових даних. Ця ємність відповідає кількості незалежних параметрів, доступних для підгонки до навчальної множини. Кожна множина сусідів S_k у методі kNN породжує окреме незалежне рішення про класифікацію документа. У цьому випадку параметром є оцінка $\hat{P}(c|S_k)$. Таким чином, ємність методу kNN обмежена только обсягом навчальної множини. Він може запам'ятовувати досить великі навчальні множини. На противагу цьому кількість параметрів у методі Роккіо фіксована — J параметрів по кожній розмірності, один для кожного центроїда — і не залежить від розміру навчальної множини. Класифікатор Роккіо (тобто його визначальні центроїди) не може “запам'ятати” тонкі деталі розподілу документів у навчальній множині.

Відповідно до рівності (6.5) наша мета — вибрати метод навчання, що мінімізує помилку навчання. Основну ідею, що відображена у формулі (6.9), можна коротко сформулювати в такий спосіб: помилка навчання являє собою суму зсуву і дисперсії, тобто складається з двох компонентів, що неможливо мінімізувати одночасно. При порівнянні двох методів навчання, Γ_1 і Γ_2, у більшості випадків виявляється, що в одного з них більше зсув і менше дисперсія, а в іншого — менше зсув і більше дисперсія. Таким чином, вибір методу навчання не зводиться до пошуку методу, що стійко породжує якісні класифікатори по навчальних множинах (невелика дисперсія) чи налаштовує їх на розв’язування задач зі складними межами між класами (невеликий зсув). Замість цього необхідно зважити відносні переваги зсуву і дисперсії для нашої прикладної задачі і на підставі цієї інформації прийняти рішення. Цей компроміс називається *компромісом між зсувом і дисперсією*.

На рис. 6.3 проілюстрована досить штучна, але корисна для розуміння ситуація Допустимо, що деякий текст китайською мовою містить англійські слова, набрані буквами латинського алфавіту, наприклад CPU, ONLINE і GPS. Розглянемо задачу, у якій необхідно відрізнити веб-сторінки, створені винятково китайською мовою, від сторінок, що містять суміш китайських і англійських слів. Пошукова система запропонує користувачам, що не володіють

англійською (але розуміють зміст таких запозичень, як CPU), можливість відфільтрувати документи на суміші мов. Для розв'язування цієї задачі пропонується використовувати дві ознаки: кількість букв латинського алфавіту і кількість китайських символів на веб-сторінці. Як указувалося раніше, відповідно до розподілу $P(<d, c>)$, більшість згенерованих змішаних (відповідно, китайських) документів лежить вище (відповідно, нижче) пунктирної лінії, але існують шумові документи.

- **Класифікатор за однією ознакою.** Йому відповідає горизонтальна лінія, що складається з точок. Цей класифікатор використовує тільки одну ознаку — кількість букв латинського алфавіту. Якщо метод навчання мінімізує кількість неправильних відповідей у навчальній множині, то положення горизонтальної поділяючої межі слабко залежить від розходжень між навчальними вибірками (тобто від шумових документів). Метод навчання, що породжує такий тип класифікатора, має невелику дисперсію, але його зсув великий, тому що він послідовно буде неправильно класифікувати квадратики в лівому нижньому куті і чорні кружечки (документи, що містять більш 50 латинських букв).

- **Лінійний класифікатор.** Йому відповідає пунктирна лінія з довгими рисками. Цей метод навчання має невеликий зсув; неправильно будуть класифіковані тільки шумові документи і, можливо, кілька документів, що прилягають до межі між двома класами. Дисперсія цього методу вище, ніж дисперсія класифікатора по одній ознаці, але залишається невеликою. Пунктирна лінія з довгими рисками ненабагато відхиляється від справжньої межі між двома класами, як практично всі лінійні поділяючі межі, побудовані за навчальними множинами. Таким чином, лише декілька документів (що прилягають до межі) будуть класифіковані неправильно.

- **Класифікатор з ідеальною підгонкою до навчальної множини**. Йому відповідає суцільна лінія. У даному випадку навчальний метод побудував поділяючу межу, що ідеально розділяє класи в навчальній множині. Цей метод має найменший зсув, оскільки жодний документ не класифікований неправильно — класифікатор іноді правильно класифікує навіть шумові документи в тестовій множини. Однак дисперсія цього методу велика. Оскільки шумові документи можуть довільно переходити через поділяючу межу, тестові документи, близькі до шумового в навчальній множині, будуть класифіковані неправильно, що нехарактерно для лінійного методу навчання.

Багато добре відомих методів класифікації текстів є лінійними. Деякі з цих методів, зокрема лінійний метод опорних векторів, регуляризована логістична регресія і регуляризована лінійна регресія, належать до категорії найбільш ефективних методів. Зрозуміти причини їхнього успіху дозволяє компроміс між зсувом і дисперсією. Типові класи в задачах класифікації текстів мають складну структуру, і малоймовірно, що їхні межі є лінійними. Однак у просторах високої розмірності, що характерні для класифікації текстів, ці інтуїтивні представлення виявляються невірними. В міру збільшення розмірності імовірність лінійної роздільності швидко зростає. Таким чином, лінійні моделі в просторах великої розмірності є досить потужними, незважаючи на свою лінійність. Незважаючи на те що більш потужні нелінійні методи навчання здатні моделювати поділяючі межі, складність яких набагато перевищує складність гіперплощини, вони більш чуттєві до шумів, що містяться у навчальних даних. Іноді, коли обсяг навчальних даних великий, нелінійні методи навчання працюють краще лінійних, але далеко не у всіх випадках.

Вправа 6.3. Який із трьох векторів, зображених на рис. 6.6 ($\vec{a}$, $\vec{b}$ чи $\vec{c}$), 1) більше за інших схожий на вектор $\vec{x}$ відповідно до міри подібності на основі скалярного добутку, 2) більше за інших схожий на вектор $\vec{x}$ відповідно до косінусної міри подібності і 3) ближче за інших до вектора $\vec{x}$ відповідно до евклідової відстані.

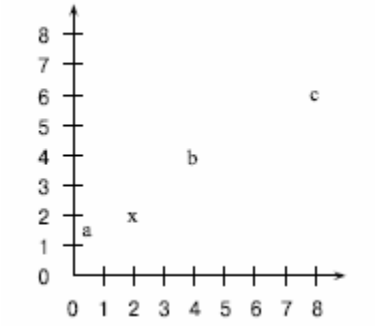

Рис. 6.6. Приклад розходжень між евклідовою відстанню, схожістю по скалярному добутку і косінусною мірою подібності. Вектори $\vec{a} = (0{,}5 \quad 1{,}5)^T$, $\vec{x} = (2 \quad 2)^T$, $\vec{b} = (4 \quad 4)^T$ і $\vec{a} = (8 \quad 6)^T$.

Вправа 6.4. Покажіть, що поділяючі межі в класифікації Роккіо, як і в методі kNN, відповідають діаграмі Вороного.

Вправа 6.5. Доведіть, що область площини, що складає з усіх точок, що мають тих самих *k* сусідів, є опуклим багатокутником.

Вправа 6.6. Розробіть алгоритм, що виконує ефективний пошук по методу 1NN по одній розмірності (ефективність оцінюється щодо кількості документів *N*). Яка часова складність цього алгоритму?

Вправа 6.6. Розробіть алгоритм, що виконує ефективний пошук за методом 1NN по двом розмірностям за поліноміальний час відносно до *N*.

Вправа 6.8. Чи можна розробити точний ефективний алгоритм за методом 1NN при дуже великому *M*, базуючись на ідеях, що були використані при виконанні попередньої вправи?

Вправа 6.9. Покажіть, що рівність (6.2) визначає гіперплощину з параметрами $\vec{w}(c_1)-\vec{\mu}(c_2)$ і $b=0,5*\left(|\vec{\mu}(c_1)|^2-|\vec{\mu}(c_2)|^2\right)$.

Вправа 6.10. Можна легко побудувати нероздільні множини даних у просторі великої розмірності, просто зануривши одну множину в іншу, як показано на рис. 6.6. Розглянете цей варіант у тривимірному просторі, а потім змініть положення точок на невелику величину у випадковому напрямку. Чи можна очікувати, що отримана конфігурація виявиться лінійно роздільною? Наскільки ймовірною є нероздільна множина, що складається з $m<<M$ точок у M-вимірному просторі?

Рис. 6.6. Проста нероздільна множина точок

Вправа6.11. Уявіть собі два класи і покажіть, що частка нероздільних вершин гіперкуба при збільшенні *M* зменшується, наприклад при $M = 1$ частка нероздільних вершин дорівнює нулю, при $M = 2$ – 2/16. Один з варіантів нероздільних вершин при $M = 2$ показаний на рис. 6.6. Інший варіант можна побудувати, застосувавши дзеркальне відображення. Розв’яжіть цю задачу аналітично або за допомогою моделювання.

Література

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