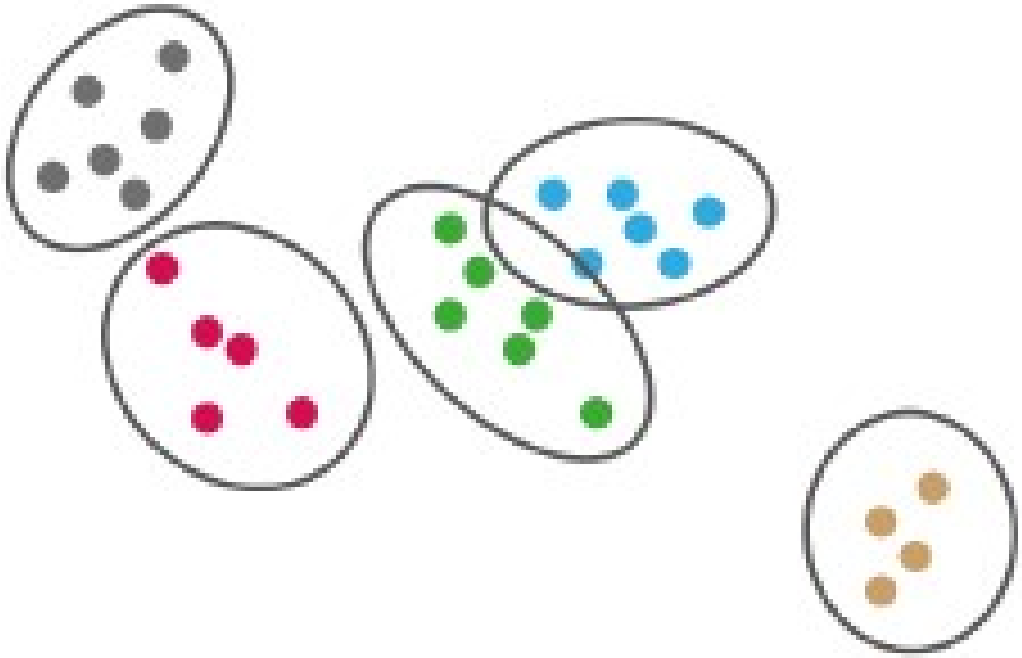
Лекція 21
Пошук аномалій. Візуалізація даних

§110 Задача виявлення аномалій

Зв'язок виявлення аномалій з іншими задачами машинного навчання

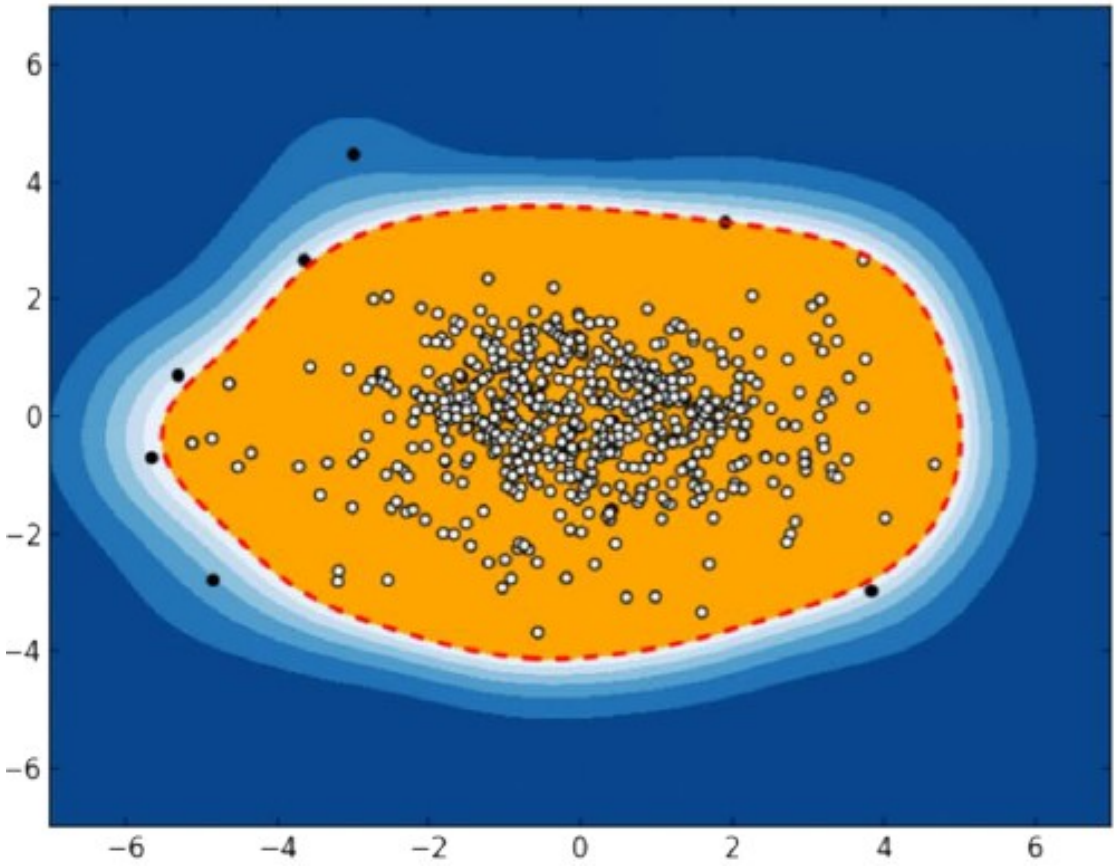
Згадаємо деякі задачі пошуку структури даних (навчання без учителя).
Кластеризація (потрібно знайти такі групи об'єктів, що об'єкти усередині кластера були схожі один на одного).



Задача зниження розмірності (вибірку потрібно спроектувати у простір меншої розмірності так, щоб зберегти якнайбільше інформації)

Задача пошуку аномалії (потрібно знайти у вибірці об'єкти, які не схожі на більшість об'єктів, які виокремлюються від інших, є аномальними)

Прикладів аномалій або немає взагалі, або їх дуже мало. Ця задача відноситься до навчання без учителя: дані не розмічені. **Необхідно навчитися розуміти, чи схожий новий об'єкт на інші відомі**.



Приклади пошуку аномалій

Перший приклад – клієнти банку.

Клієнта описуємо **характеристиками транзакцій**, **поведінкою в інтернет-банку і таке інше**.

Чи відрізняється його поведінка від середньої за усіма клієнтами?

Може це шахрай? І з ци клієнтом потрібно поводитись як з шахраєм?

Другий приклад — моніторинг складної комп'ютерної системи, що складається з великої кількості взаємозалежних машин.

Характеристики: завантаження процесорів, використання пам'яті на кожній машині, навантаження на мережу й т.д.

Чи відрізняється поточний стан системи від характеристик тих станів, про які відомо, що вони нормальні.

Якщо відрізняється, і система поводиться якось інакше, то це привід задуматися, чи не трапилася якась поломка, потрібно чи продіагностувати систему й щось полагодити в ній.

Третій приклад – програма (модель), яка за відгуком про банк визначає його тональність: позитивну або негативну.

Процедура наступна: клієнт заходить на сайт банку, у спеціальну форму вводить деякий відгук, далі цей відгук приходить на вхід моделі, що визначає, позитивний він або негативний. Якщо він негативний, то потрібно сповістити про це співробітникам банку, щоб вони вирішили виниклу проблему.

Але крім цього хотілося б розуміти, чи не зламалась така модель?

Коли приходить новий відгук, чи можна до нього застосовувати цю модель машинного навчання, чи можливо його класифікувати цією моделлю. Справа в тому, що **розподіл ознак цього об'єкта міг змінитися**. Наприклад, **банк міг поміняти назву продуктів**, і тому тепер слова, що зустрічаються у відгуці, зовсім інші, – модель до них не готова. Або **банк міг змінити обмеження на довжину відгуку**. **Через це модель може стати непридатної для рішення задачі.**

Якщо стало відомо, що **об'єкти стали іншими, аномальними, у порівнянні з тими, на яких навчалася модель, то це привід її навчити на нових розмічених даних**.

Методи виявлення аномалій

- Методи, які засновані на відновленні **густини розподілу імовірності**
- Методи, які засновані на **методах класифікації**

§111 Параметричне відновлення густини

Імовірнісний підхід до виявлення аномалій

- В імовірнісному підході для виявлення аномалій уважається, що **аномалія - це об'єкт, що був отриманий з розподілу, відмінного від того, за допомогою якого сгенерована навчальна вибірка**.
- **Виникає питання: як знайти розподіл, з якого була отримана вибірка**?
- **Якщо знайти цей розподіл, то можна оцінити ймовірність належності нового об'єкта цьому розподілу**.
- **Якщо ймовірність одержати новий об'єкт із цього розподілу дуже мала, то це, швидше за все, аномалія.**

Існує три основних підходи до відновлення імовірнісної густини:

- параметричний підхід,
- непараметричний підхід,
- відновлення сумішей.

Параметричний підхід

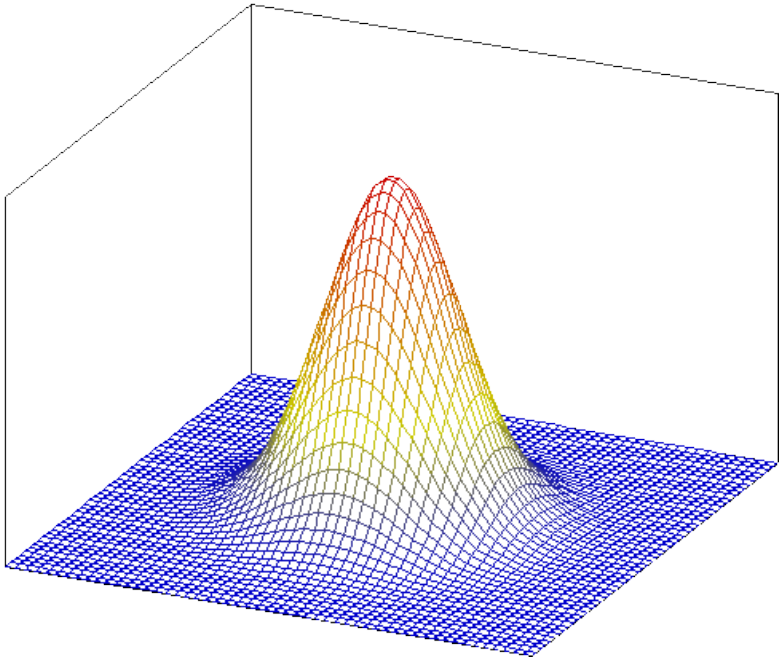
Існує деякий імовірнісний розподіл $p(x)$ на всіх об'єктах, які можна одержати. Приймаємо, що цей розподіл є параметричним (залежить від певних параметрів):

$$p(x) = \phi(x|\theta),$$

θ – задає параметри розподілу.

Приклад параметричного сімейства розподілів — це нормальний розподіл:

$$\phi(x|\theta) = \mathcal{N}(\mu, \Sigma)$$



Нормальний розподіл задається параметрами $\theta = (\mu, \Sigma)$; (центр та коваріаційна матриця). Параметр μ визначає, де знаходиться центр цього капелюха, а параметр Σ — те, наскільки об'єкти розкидані навколо центра.

Як знайти параметри розподілу?

Метод максимальної правдоподібності

Логічно **шукати параметр розподілу** так, щоб **ймовірність об'єктів навчальної вибірки** за цим розподілом **була максимальною** (максимізувати ймовірність).

Працювати із самою **правдоподібністю незручно, оскільки це — добуток значень густини у всіх точках навчальної вибірки**. Замість цього **можна взяти його логарифм** і намагатися максимізувати отриману суму:

$$\sum_{i=1}^{\ell} \log \phi(x_i|\theta) \to \max_{\theta}$$

Для **деяких розподілів** цю задачу можна вирішити **аналітично**, якщо обчислити частинні похідні й прирівняти їх до нуля.

Наприклад, для нормального розподілу такі рішення існують:

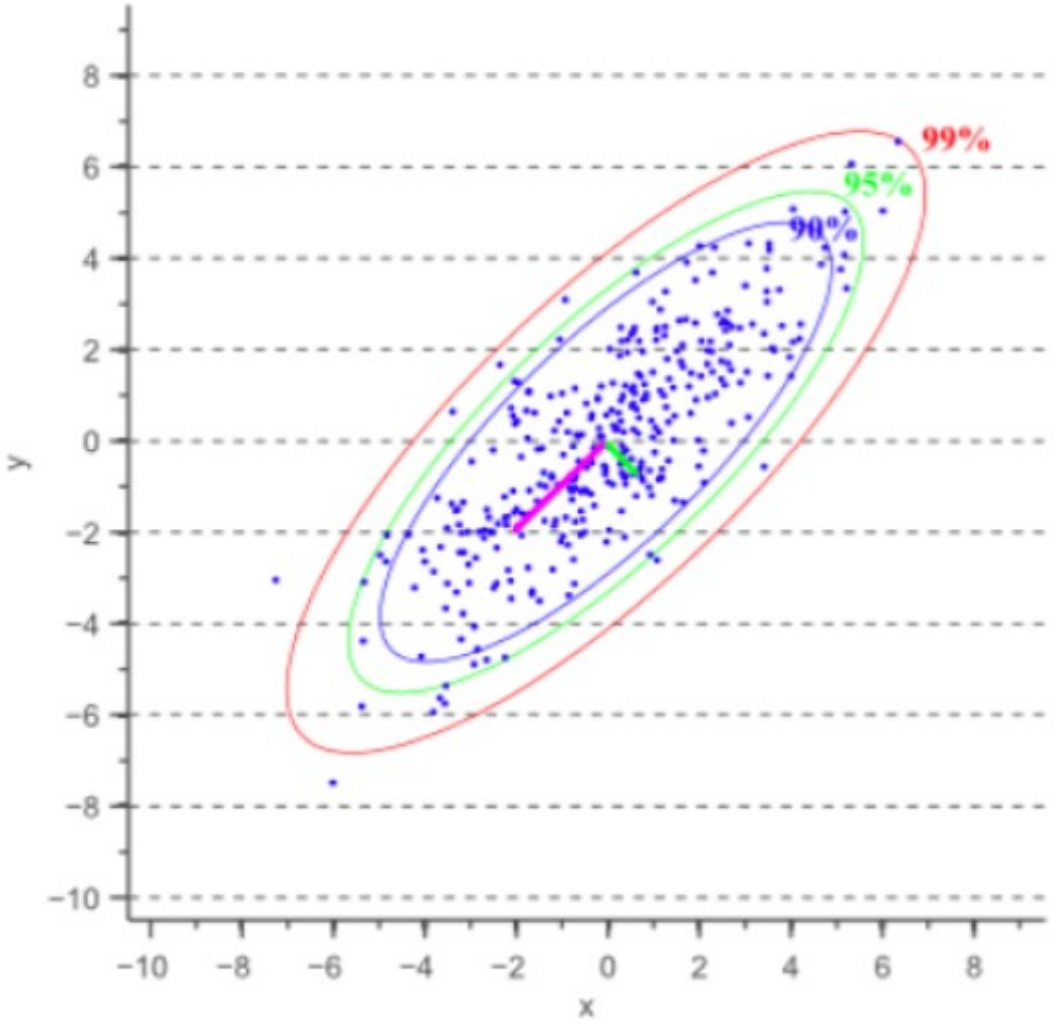
$$\mu = \frac{1}{\ell} \sum_{i=1}^{\ell} x_i$$

$$\Sigma = \frac{1}{\ell} \sum_{i=1}^{\ell} (x_i - \mu)(x_i - \mu)^T$$

Приклад: вибірка, що зображена на рисунку, створена **нормальним розподілом**.

Знаходимо параметри цього розподілу **методом максимальної правдоподібності**. За цим розподілом малюємо **лінії рівного рівня**. Синяї лінія рівня усередині знаходиться **90 %**, зелена - **95%**, червона - **99 %**.

Висновок: імовірність одержати об'єкт поза червоним еліпсом дуже мала. **Дві точки, які знаходяться поза червоним еліпсом – це аномалія.**



Алгоритм роботи:

- Знаходимо розподіл $p(x)$
- Для нового об'єкта x обчислюємо ймовірність, порівняємо її з деяким порогом t.
- Якщо ймовірність менше цього порога, об'єкт оголошується аномалією.

Як вибирати поріг t?

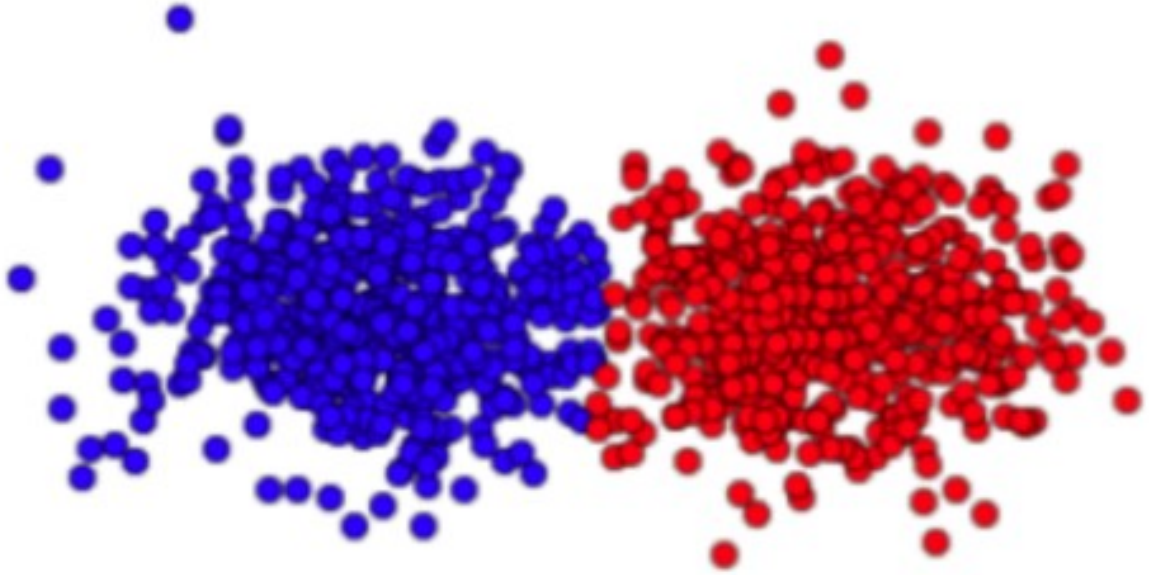
- Не існує однозначних рекомендацій.
  Можна вибирати його з апріорних міркувань (наприклад, поза лінією рівня 99%).
- Якщо є об'єкти, про які точно відомо, що це — аномалії, то можна підібрати поріг t так, щоб ці об'єкти були оголошені аномальними, а всі інші — згенерованими з розподілу $p(x)$.

Модель суміші розподілів EM-алгоритм

У деяких випадках параметричного підходу виявляється недостатньо.

Наприклад, на рисунку зображена вибірка, що згенерована із **двох нормальних розподілів** з однаковими матрицями коваріацій, але різними центрами, таким чином, отримуємо дві хмари точок.

Описати цю вибірку одним нормальним розподілом буде неможливо, зате для цього відмінно підходить модель суміші розподілів.



Сумішшю називається такий розподіл $p(x)$, що подається у вигляді зваженої суми інших розподілів:

$$p(x) = \sum_{j=1}^{K} w_j p_j(x), \qquad p_j(x) = \phi(x|\theta_j)$$

Розподіли $p_j(x)$ називаються **компонентами суміші**, і, як правило, вони є параметричними розподілами.

Властиво, кожний компонент p_j є членом параметричного сімейства $\phi(x)$ зі **своїм параметром** θ_j.

EM – алгоритм

$$p(x) = \sum_{j=1}^{K} \mathrm{w}_j \, p_j(x) \qquad p_j(x) = \phi(x|\theta_j)$$

- E- крок:

$$g_{ji} = p(j|x_i) = \frac{\mathrm{w}_j \, p_j(x_i)}{p(x_i)}$$

- M- крок:

$$\mathrm{w}_j = \frac{1}{N} \sum_{i=1}^{N} g_{ji}$$

$$\theta_j = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} g_{ji} \ln \varphi(\theta; x_i)$$

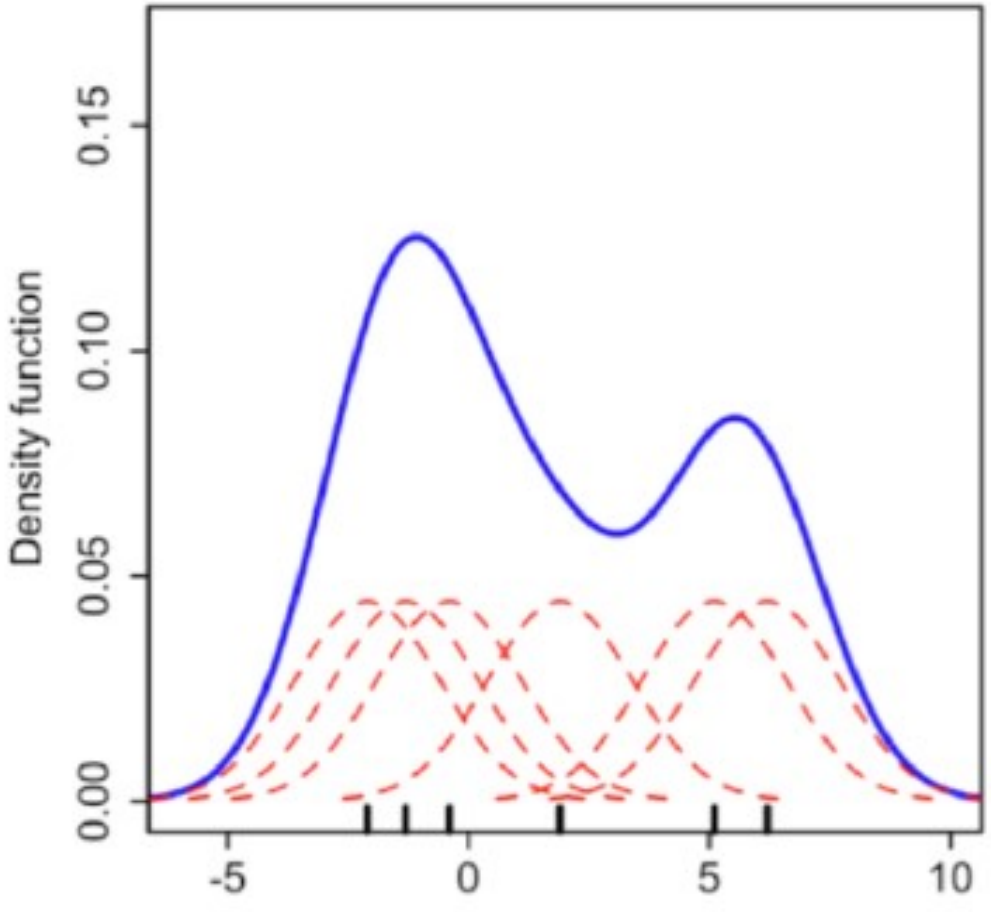
Завдяки цьому алгоритму можна визначити, яка саме суміш із *K* розподілів створює вибірку.

§113 Непараметричне відновлення густини

Формула Парзена-Розенблатта і його параметри

Непараметричний підхід до відновлення густини полягає в тому, що вид розподілу намагаються відновити, не вводячи ніяких сімейств розподілів, **використовуючи тільки самі дані**.

Нехай **є одновимірна вибірка** (**об'єкти** вибірки **позначені мітками** на осі абсцис). **У кожній точці** навчальної вибірки поміщають **центр гаусіани** (показані червоною лінією), таким, що усім точкам на осі привласнюються деякі ймовірності. Далі, **у кожній точці ці гаусіани підсумуються, і отримують підсумковий розподіл** (на рисунку показано синім).



Формально це можна зробити за допомогою формули Парзена-Розенблатта:

$$p_h(x) = \frac{1}{\ell h} \sum_{i=1}^{\ell} K\left(\frac{x - x_i}{h}\right)$$

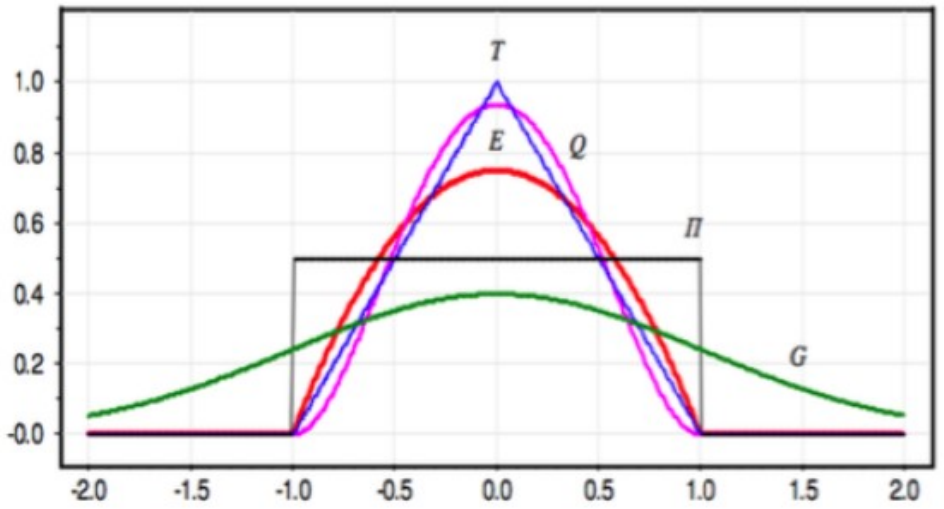
Тут *K*(*r*) — ядро (параметр методу, що характеризує ймовірність того, що точки x й x_i схожі одина на одну).

Ядро – це парна функція, також для нього повинна виконуватися умова

$$\int K(r)dr = 1$$

Існує багато різних ядер.

- $E(r) = \frac{3}{4}(1 - r^2)[|r| \leq 1]$ — оптимальне
- $Q(r) = \frac{15}{16}(1 - r^2)^2[|r| \leq 1]$ — квадратичне
- $T(r) = (1 - |r|)[|r| \leq 1]$ — трикутне
- $G(r) = (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}r^2\right)$ — гаусове
- $\Pi(r) = \frac{1}{2}[|r| \leq 1]$ — пряме



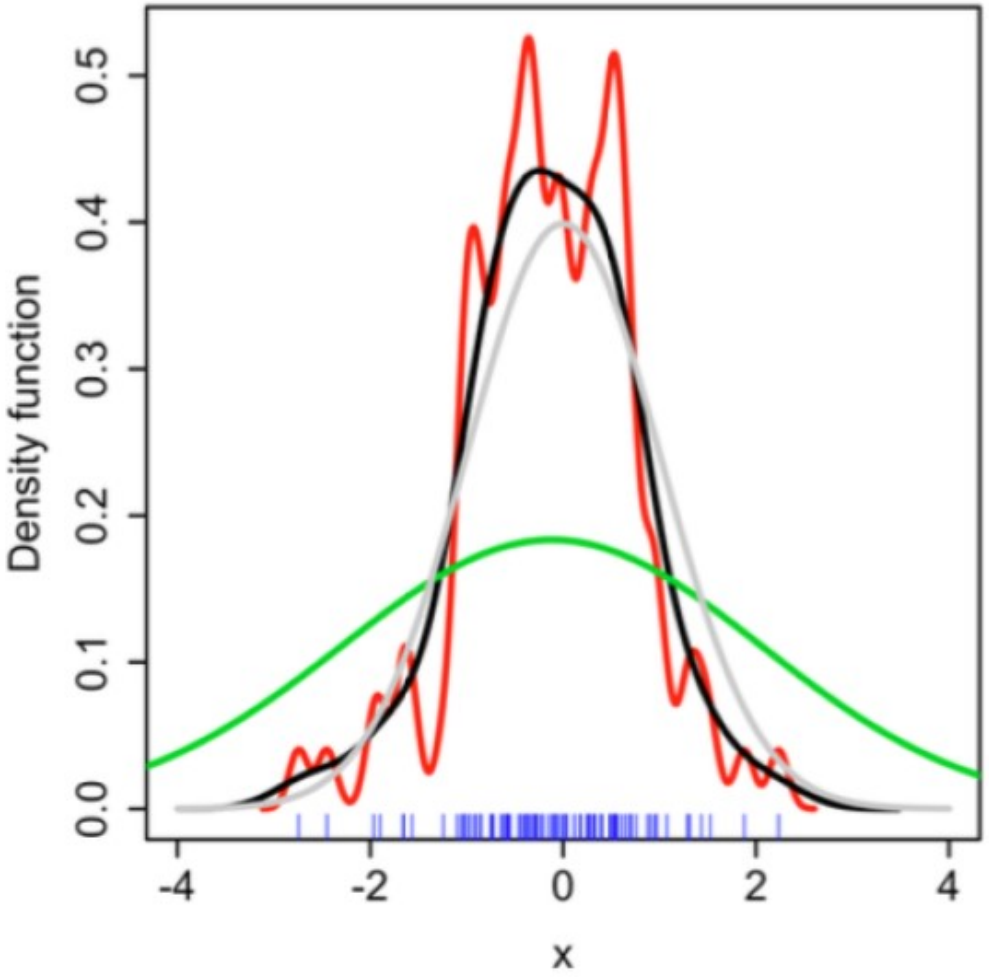
Ще **один параметр оцінки** Парзена-Розенблатта — це **ширина вікна** h.

$$p_h(x) = \frac{1}{\ell h} \sum_{i=1}^{\ell} K\left(\frac{x - x_i}{h}\right)$$

- h — ширина вікна

Вплив ширини вікна на відновлену густину імовірності

Елементи вибірки зображені синіми мітками на осі абсцис. **Червона крива** відповідає дуже маленькій ширині вікна, і результуюча густина дуже чутлива до всіх точок. **Чорна крива** відповідає більше високому значенню ширини вікна. **Зелена крива** відповідає оцінці густини з дуже великою шириною вікна.



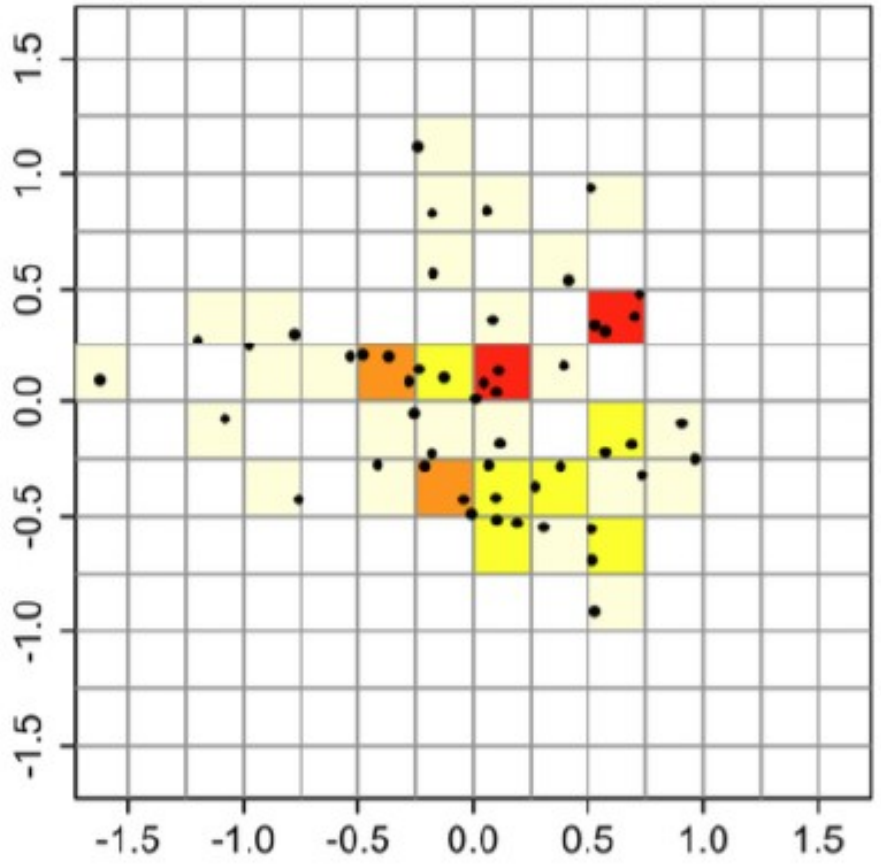
Багатомірний випадок, проблеми які виникають

Непараметричне оцінювання густини добре узагальнюється на багатомірний випадок, при цьому необхідно різницю між точками x й x_j замінити метрикою й увести нормувальну константу $V(h)$, щоб густина була нормована:

$$p_h(x) = \frac{1}{\ell V(h)} \sum_{i=1}^{\ell} K\left(\frac{\rho(x, x_i)}{h}\right)$$

- $V(h) = \int K\left(\frac{\rho(x, x_i)}{h}\right) dx$ —

Є одна **велика проблема: чим вище розмірність, тим більше потрібно об’єктів у вибірці**.



Якщо спроектувати цю вибірку на одну вісь, то **для одновимірного випадку є достатньо об'єктів, щоб відновити густину імовірності**.

У двовимірному просторі на один елемент площі маємо набагато менше об'єктів, ніж на відповідний йому відрізок в одномірному просторі.

Число об'єктів, необхідних для відновлення густини, росте експоненціально зі збільшенням розмірності простору, і на практиці цей підхід можна застосовувати в просторах не дуже високої розмірності.

Виявлення аномалій з використанням непараметричного підходу

Так само, як і з використанням параметричного відновлення густини

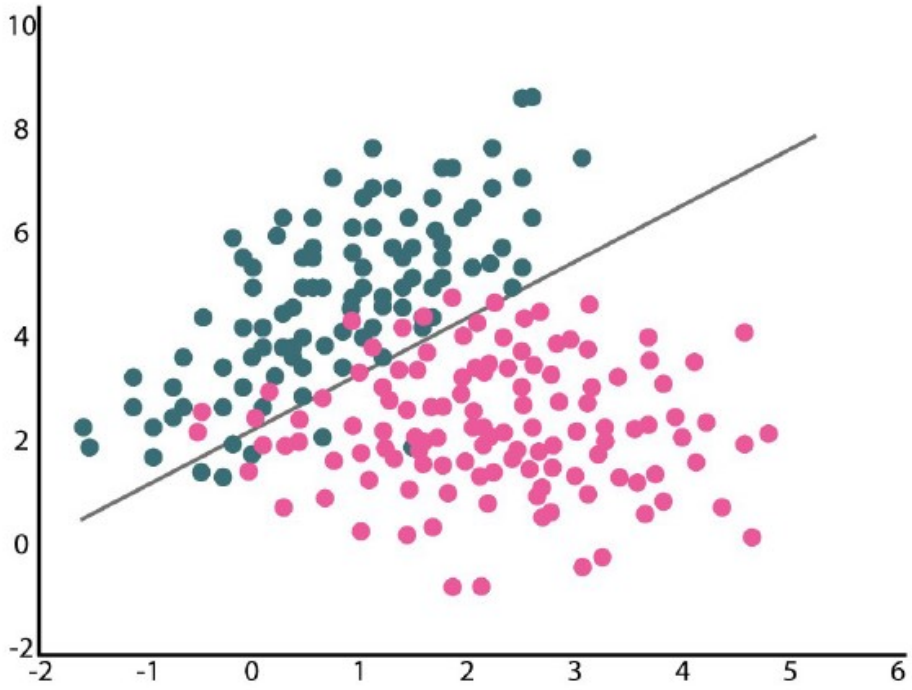
- Для кожного нового об'єкта обчислюється густина імовірності $p(x)$,
- Якщо $p(x) < t$, де t — заданий поріг, то об'єкт уважається аномалією.
- Вибір порога:
    - з апріорних (деяких теоретичних) міркувань
    - за відомими аномаліями

§114 Одноκласовий SVM (support vector machines, метод опорних векторів)

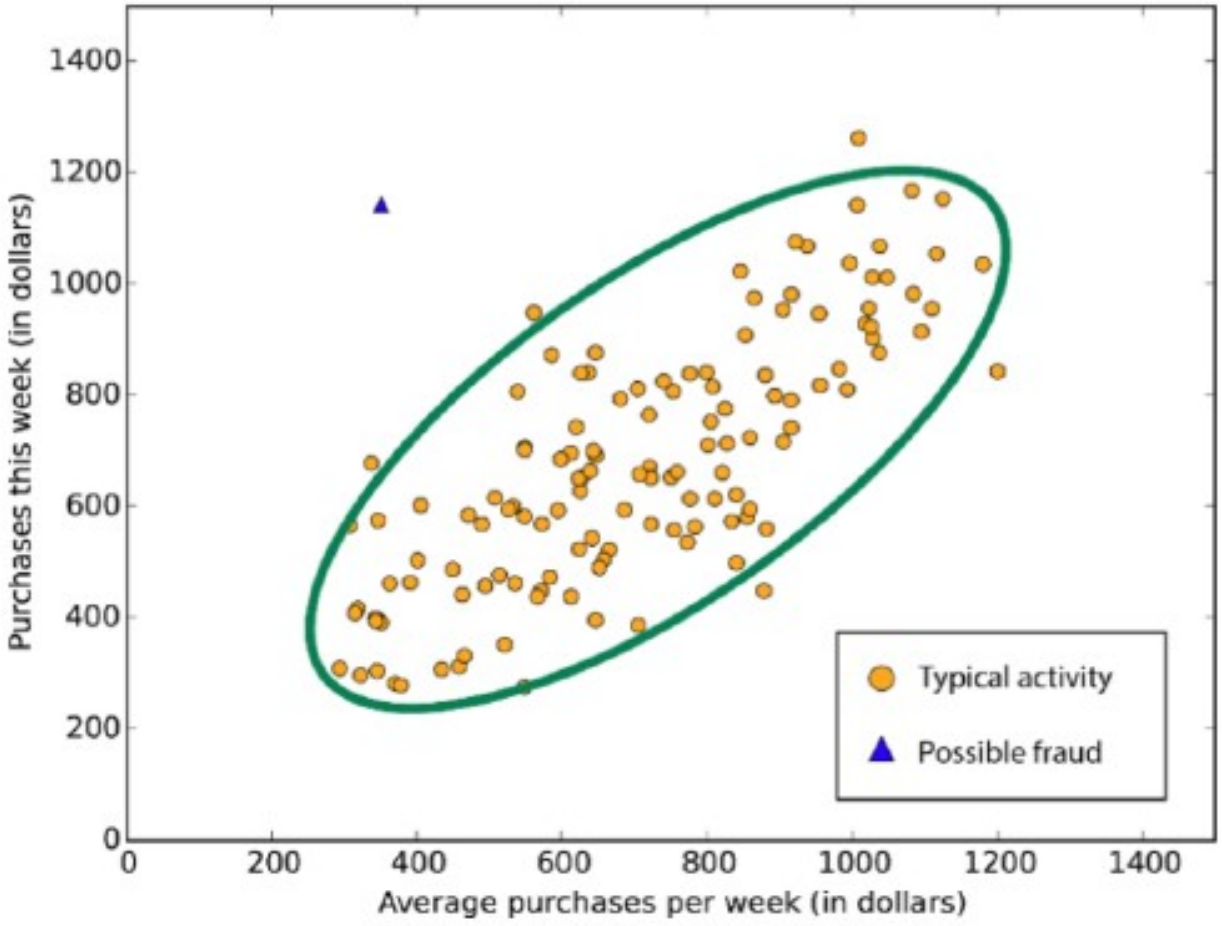
Зв'язок виявлення аномалій із задачею класифікації

Задачу виявлення аномалій можна зв'язати із задачею класифікацією.

Нагадаю задачу бінарної класифікації (див. рисунок) 2 класи необхідно розділити прямою (у просторах високої розмірності — гіперплощиною).



У задачі виявлення аномалій теж є вибірка, але **потрібно побудувати деяку криву, що відокремить вибірку від усього іншого** (див. рисунок). **Усе, що перебуває поза цією кривою, буде вважатись аномаліями**, тому що це щось, що не попадає в множину типових об'єктів з вибірки.



Як можна об’єднати ці дві задачі?

У **задачі виявлення аномалій теж необхідно побудувати деяку поділяючу поверхню**, але при цьому в наявності немає прикладів аномалій, є тільки приклади типових об'єктів.

- **Можна вважати, що в задачі присутні 2 класи**.
- **Перший клас - це нормальні об'єкти**, до нього відноситься вся навчальна вибірка.
- **Другий клас - це аномалії, і вважається, що аномальним є початок координат**.

Тепер можна вирішувати цю задачу, використовуючи методи класифікації.

Застосування методу опорних векторів для виявлення аномалій

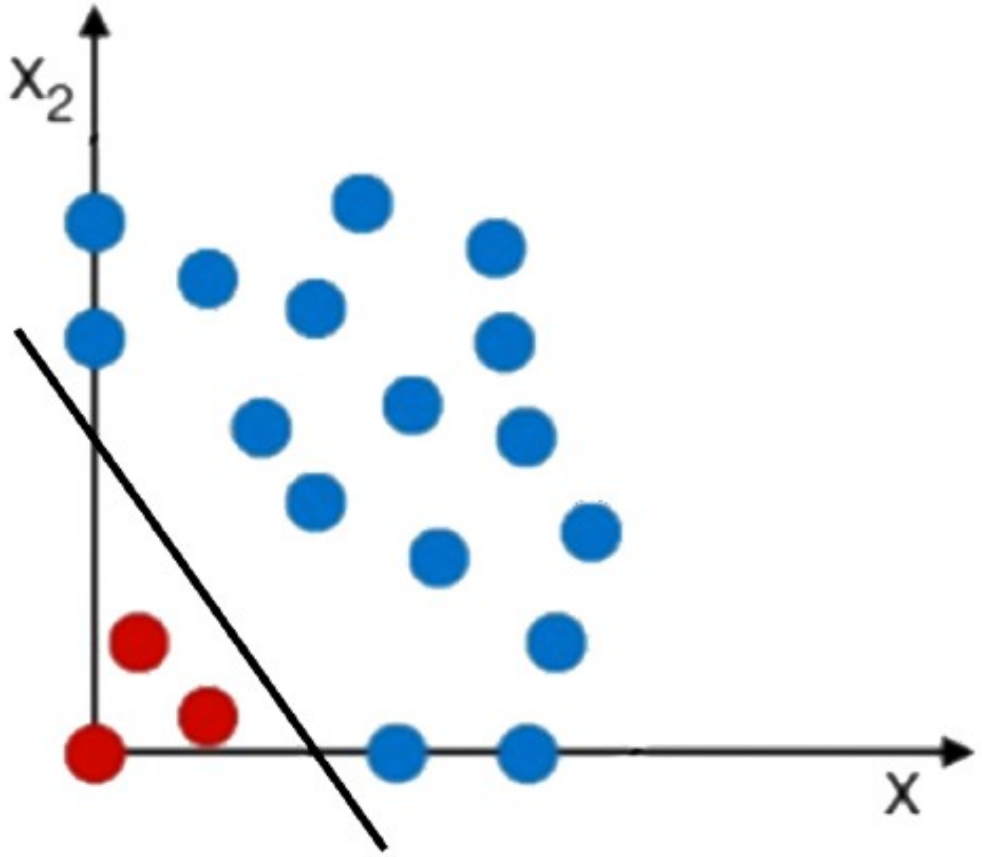
Для вирішення буде **використовуватися лінійний спосіб**, при цьому необхідно вибирати поділяючу гіперплощину так, щоб **вона давала максимальний розмір зазору**, тоді **ймовірність перенавчання буде мінімальною**. Це найкраще робити за допомогою **методу опорних векторів**, або SVM.

Задача відділення вибірки від початку координат формується так:

$$\begin{cases} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{v\ell}\sum_{i=1}^{\ell}\xi_i - \rho \to \min\limits_{\mathbf{w},\xi,\rho} \\ \langle \mathbf{w}, x_i \rangle \geq \rho - \xi_i, \quad \xi_i \geq 0 \end{cases}$$

Відмінність: важливий параметр задачі υ, задає верхню оцінку для частки аномальних об'єктів у вибірці.

Якщо вибрати υ так, що аномальними можуть бути оголошені 3 об'єкти, то поділ може відбутися як на рисунку



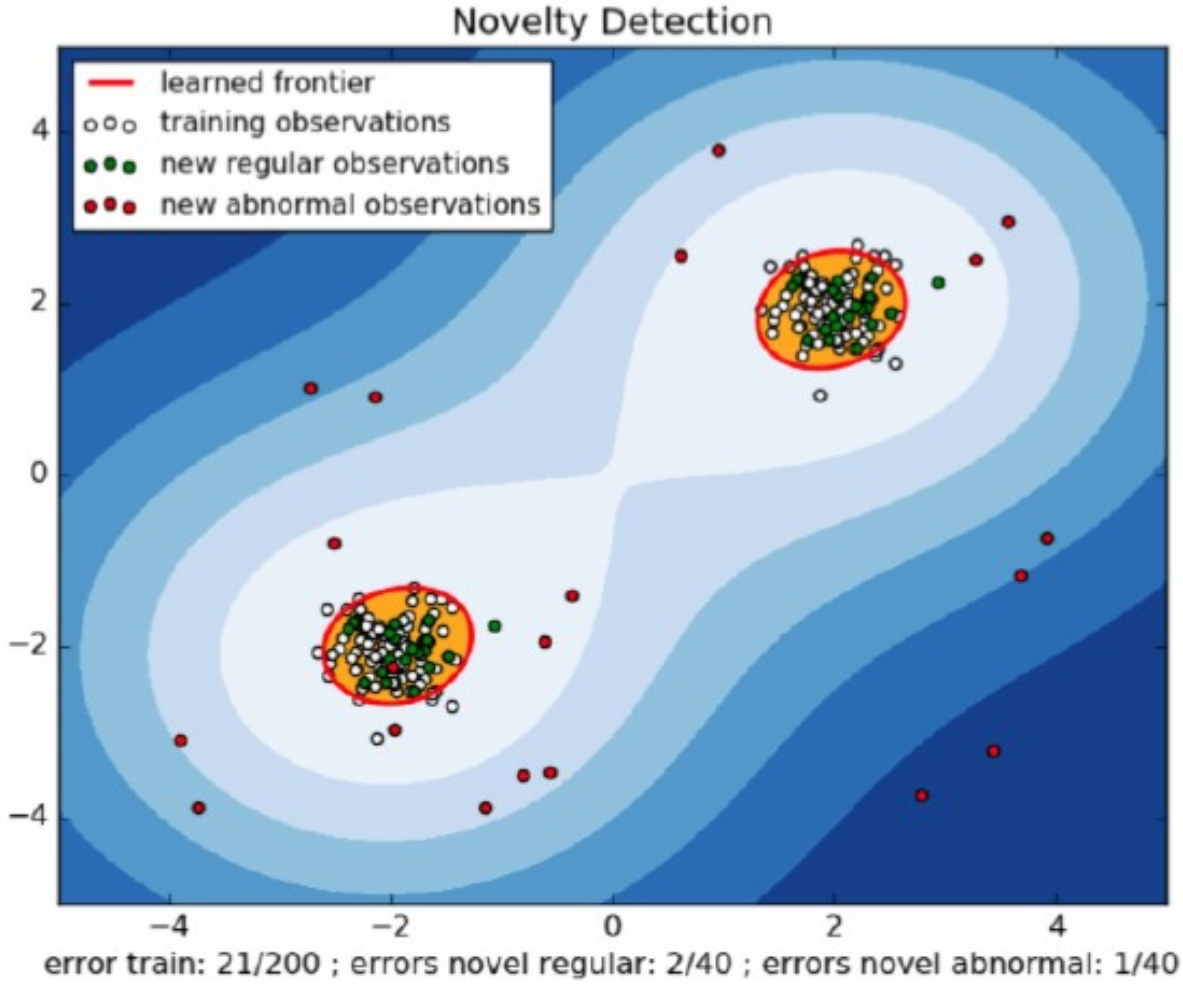
Класифікація з використанням нелінійного ядра

Припущення про те, що початок координат — це аномальний об'єкт, дуже дивне.

- Наприклад, якщо вибірка центрована навколо початку координат, те цей підхід у принципі не може дати правильний результат.
- Насправді, однокласовий SVM з лінійним ядром ніколи не використовується.
- **Скалярний добуток** у сформульованій задачі можна замінити **на ядро *K***.
- Тоді SVM дозволяє будувати нелінійні розділяючі поверхні
- Популярний вибір - це RBF-ядро, що обчислюється за формулою

$$K(x,z) = \exp\left(\frac{\|x-z\|^2}{\sigma^2}\right)$$

Після заміни ядра поділяюча площина буде будуватися в просторі більше високої розмірності. Приклад застосування ядрового однокласового SVM з RBF-ядром у вибірці показаний на рисунку.



§115 Задача візуалізації

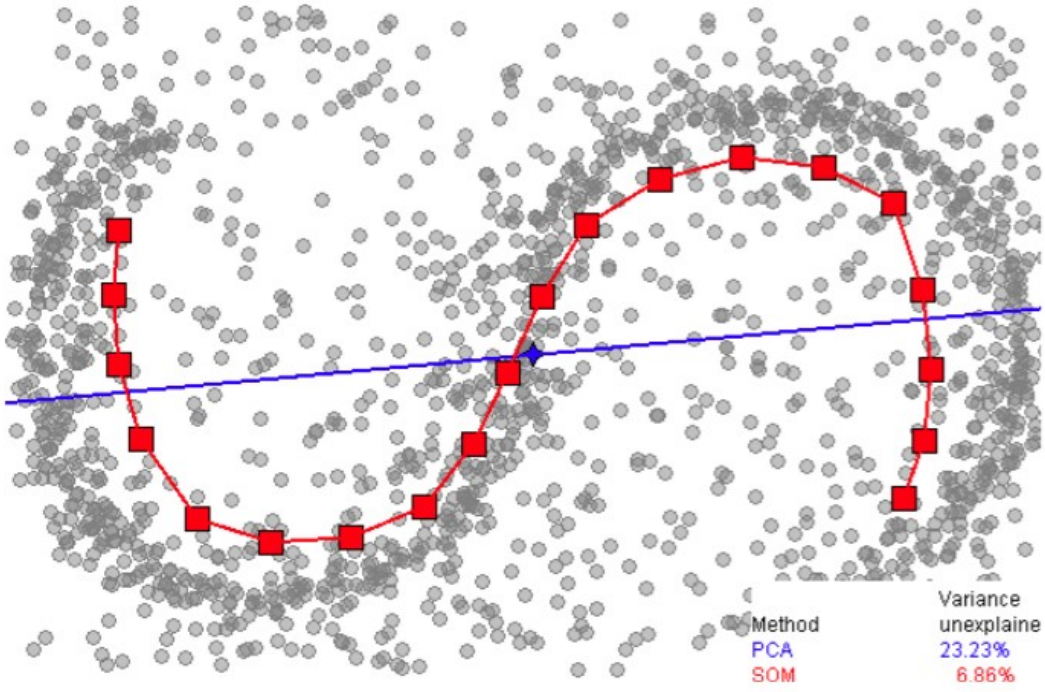
Зв'язок задачі візуалізації й методів зниження розмірності

Навіщо візуалізувати дані?

Приклад: є вибірка, потрібно понизити її розмірність до однієї ознаки.

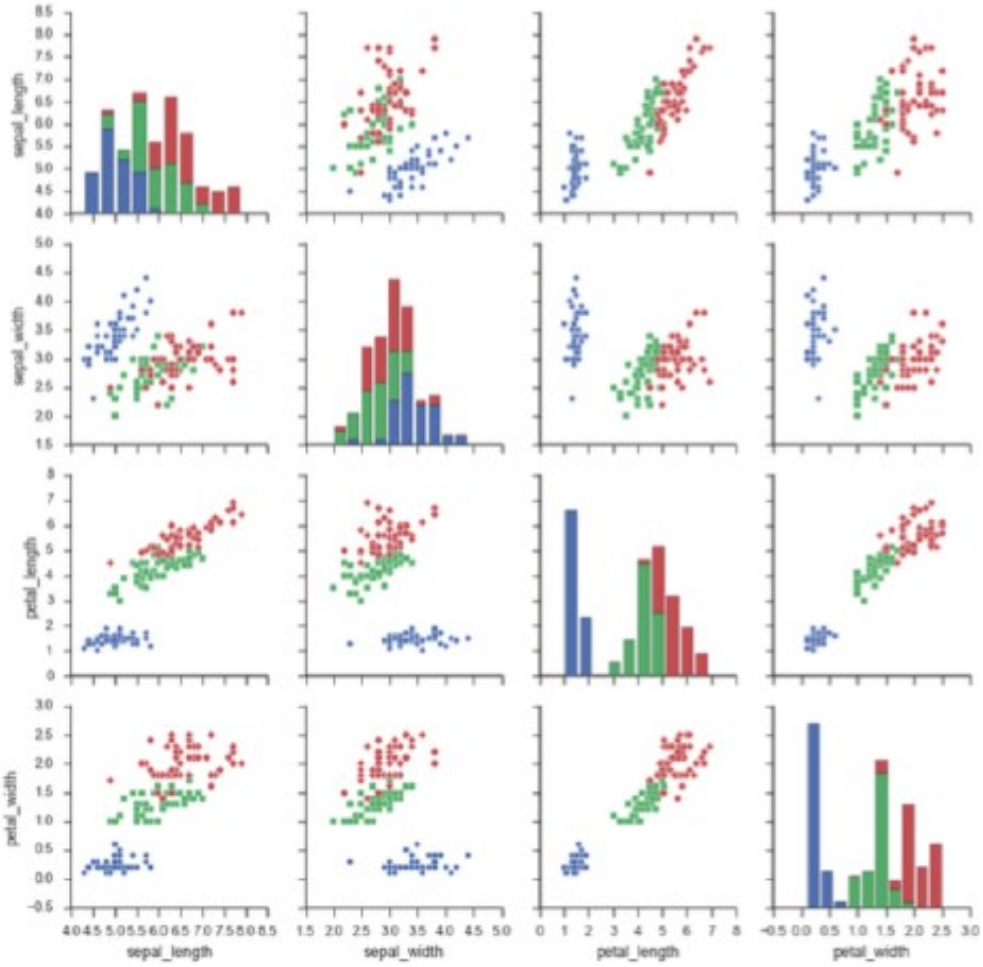
Якщо не знати структури даних, то **безуспішно намагатися проектувати об'єкти на пряму**, втрачаючи при цьому багато інформації.

Якщо подивитися на візуалізацію даних, то стає зрозуміло, що необхідно проектувати дані **на червону криву**, а для цього **потрібно використовувати нелінійні методи зниження розмірності**.



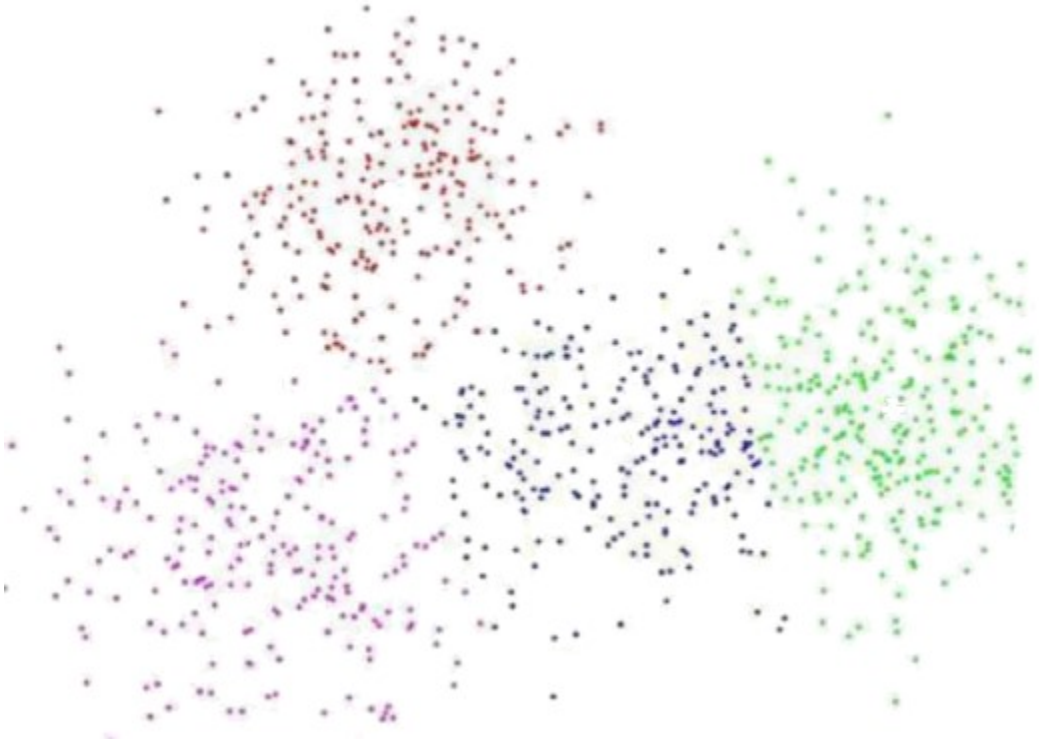
У машинному навчанні мають справу з вибірками, у яких **дуже багато ознак**. **Хочеться розуміти, як улаштовані дані**, які в них є **взаємозв'язки**, які ознаки **важливі**.

Для цього можна, як на рисунку, для **кожної пари ознак спроектувати вибірку на цю пару**, а для кожної ознаки окремо побудувати гістограму. Із цих графіків можна витягти багато інформації: помітити, **які ознаки краще розділяють класи, або навіть якийсь одна ознака їх добре розділяє**.



Недолік такого підходу — неможливість бачити всю вибірку в цілому. Хочеться відобразити дані **у двовимірний або тривимірний простір**, щоб була видна **їхня структура** (як на рисунку): **які класи добре роздільні, які — перемішані між собою**.

Задачі візуалізації даних - це окремий випадок нелінійного зниження розмірності, коли дані проектують на площину або в тривимірний простір. При цьому хочеться спроектувати дані так, щоб зберегти всю структуру й закономірності.



Пошук структури в даних за допомогою візуалізації

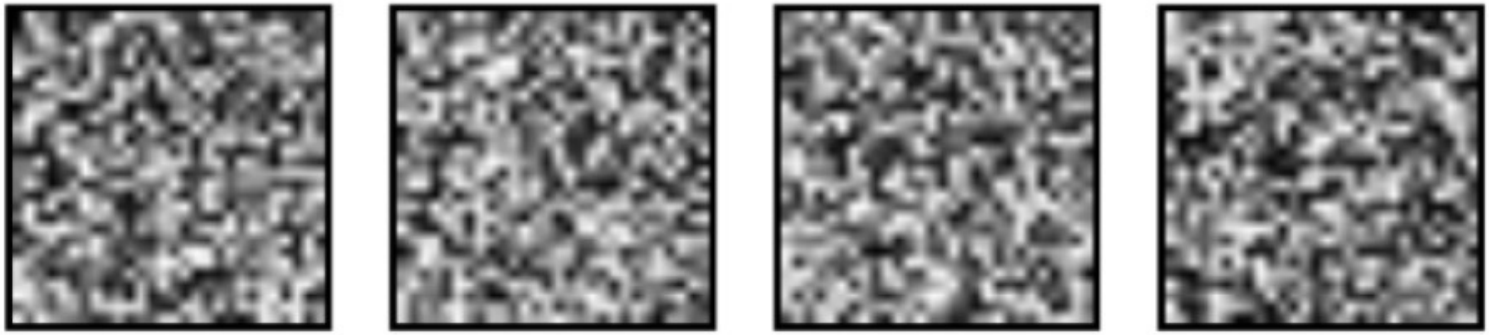
Приклад, за допомогою якого легко продемонструвати задачу візуалізації — це набір даних MNIST (рисунок). Це зображення цифр, написаних від руки, причому всі цифри дуже різні.

Потрібно **їх розділити на 10 класів**, тобто визначити за зображенням, яка цифра на ньому намальована. Кожна картинка в цьому наборі даних **має розмір 28x28 пікселів**, тобто всього в кожного об'єкта **784 ознаки**.

Як їх усі зобразити у двовимірну площину?

Можна використовувати набагато менше ознак, щоб характеризувати кожну рукописну цифру (**внутрішня розмірність цифр набагато нижча за 784**).

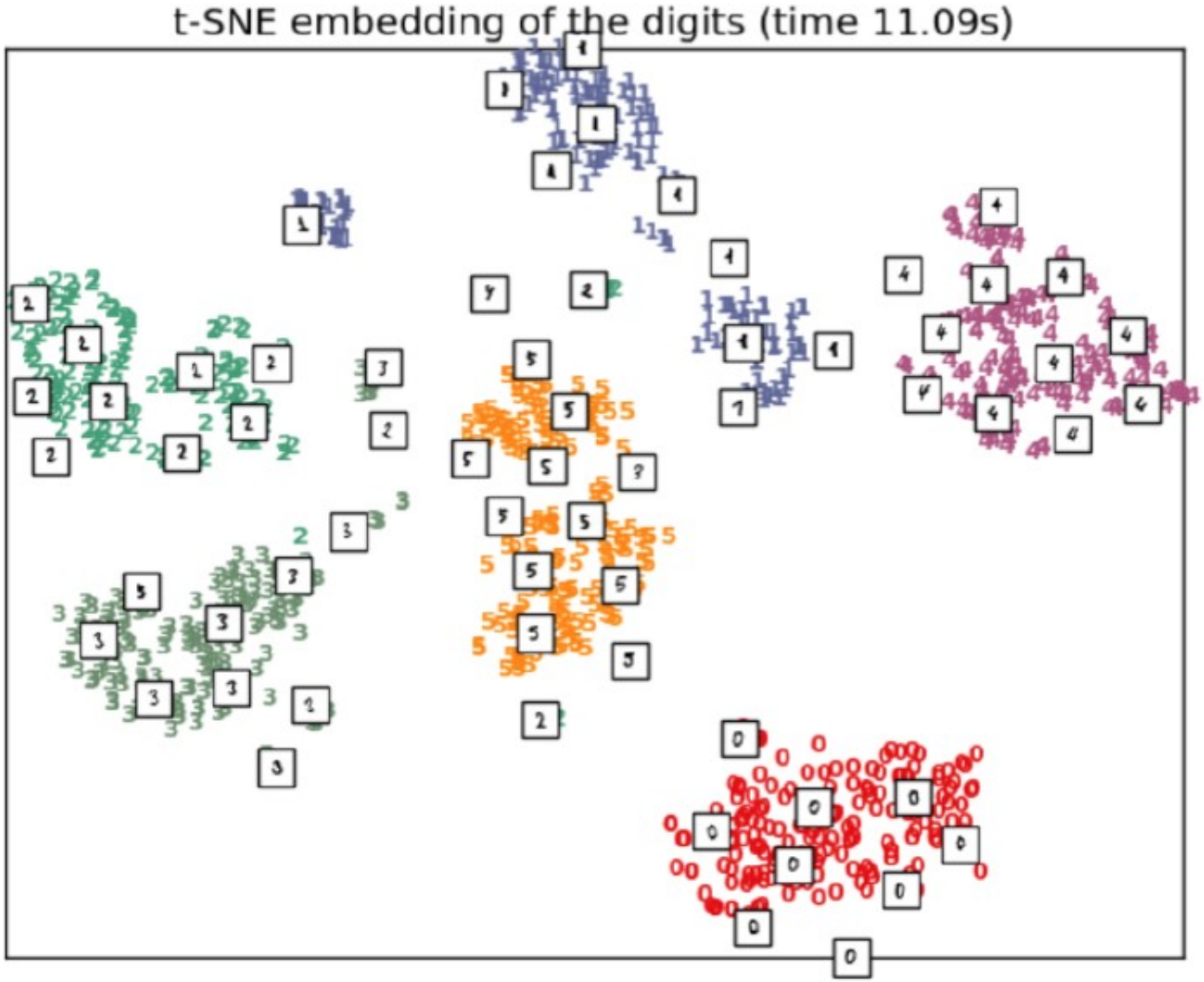
Це допомагає зрозуміти наступний приклад: якщо генерувати випадкові картинки розміром 28x28 пікселів, то результат буде схожий на картинки, показані на рисунку.



Порівняємо ці рисунки з рисуками з цифрами і бачимо, що на них **немає нічого загального із зображеннями цифр**.

Це дає зрозуміти, що **внутрішня розмірність набору цифр суттєво нижче**.

Якщо скористатися **методом візуалізації t-SNE** (докладніше про нього буде сказано далі), то отримаємо рисунок. Усього двох ознак досить, щоб зобразити ці цифри так, що всі класи будуть ідеально роздільні.

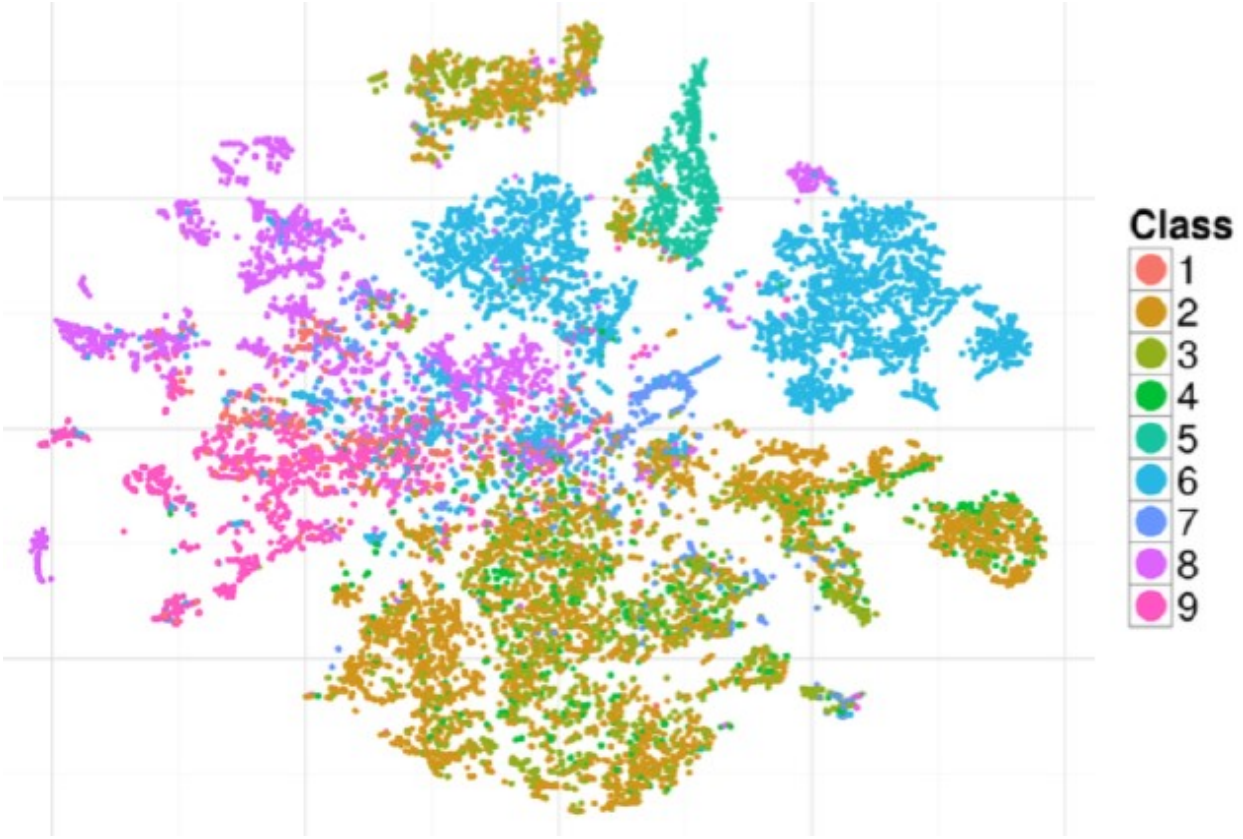


Ще один приклад.

Потрібно, використовуючи опис характеристики товару, визначити, до якій з **9 категорій** він відноситься. Усього ознак було **93**.

Якщо візуалізувати цей набір даних, то отримаємо зображення, показане на рисунку.

Якщо у вихідному ознаковому просторі використовувати для класифікації складні методи (випадковий ліс, градієнтний бустінг), то побачимо, **що в даних зберігаються ті ж закономірності, які показані на рисунку: добре роздільні на зображенні класи розділяються добре, перемішані класи — погано**.



§116 Багатовимірне шкалювання

Постановка задачі

До цього мова вже йшла про **лінійні методи зниження розмірності**, наприклад, описувався **метод випадкових проекцій**:

$$z_{ij} = \sum_{k=1}^{D} \mathrm{w}_{jk}\, x_{ik}$$

$$\mathrm{w}_{jk} \sim \mathcal{N}\left(0, \tfrac{1}{d}\right)$$

Також обговорювався **метод головних компонентів**, що вибирає ваги більш грамотно: він виражає їх через сингулярні вектори матриці « об'єкти-ознаки».

Ці методи не можуть визначити нелінійні залежності в даних.

Знайти їх може, наприклад, метод багатовимірного шкалювання (MDS). Використовує гіпотизу – під час візуалізації потрібно зберігати попарні відстані між об’єктами.

Для подальшого опису методу потрібно ввести формальні позначення:

- $x_1, \ldots, x_\ell$ — об'єкти у віхдному просторі
- $\tilde{x}_1, \ldots, \tilde{x}_\ell$ — об'єкти у маловимірному просторі
- $d_{ij} = \rho(x_i, x_j)$ — відстані у вихідному просторі
- $\tilde{d}_{ij} = \|\tilde{x}_i - \tilde{x}_j\|$ — відстані у маловимірному просторі

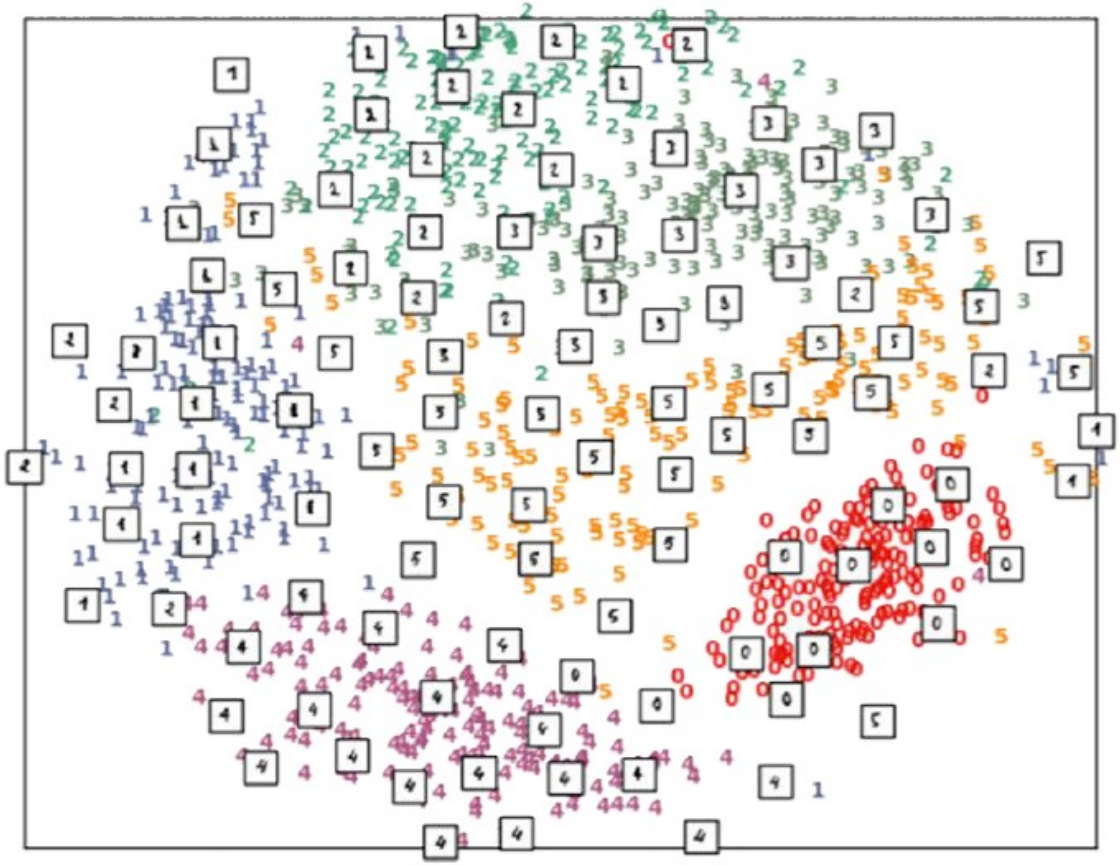
У задачі багатовимірного шкалювання потрібно, щоб попарні відстані між об'єктами змінювалися якнайменше:

$$\sum_{i<j}^{\ell} (\|\tilde{x}_i - \tilde{x}_j\| - d_{ij})^2 \to \min_{\tilde{x}_1,\ldots,\tilde{x}_\ell}$$

Звернемо увагу, що при використанні даного методу **не потрібно знати ознаковий опис об'єктів, досить лише вміти обчислювати відстані між ними**.

Оптимізація методу багатовимірного шкалювання

Ця задача **складніше** за ті, які зустрічалися до цього. Вирішити задачу можна за допомогою **методу оптимізації SMACOF**, але він досить складний, тому тут описаний не буде.



Результат багатовимірного шкалювання набору даних MNIST показаний на рисунку.

Дані непогано розділені, однак і об'єкти усередині одного класу, і самі класи розташовуються досить **близько один до одного**. Це результат того, що дані спроектовані з 728-мірного простору, а **чим більше розмірність простору, тим більше схожі відстані між різними парами об'єктів (прокляття розмірності)**.

§117 Метод t-SNE

Метод SNE

Рішення задачі **багатовимірного шкалювання**, яка описана раніше, **є не дуже якісними**, тому що **непросто зберегти попарні відстані між об'єктами при радикальному зменшенні розмірності простору**.

Метод SNE (Stochastic Neighbor Embedding) намагається вирішити цю проблему: **потрібно не збереження відстаней, а збереження пропорцій відстаней між об'єктами**.

- Stochastic Neighbor Embedding
- Зберігаємо пропорції відстаней
- Якщо $\rho(x_i, x_j) = \alpha\rho(x_i, x_k)$,

  то $\rho(\tilde{x}_i, \tilde{x}_j) = \alpha\rho(\tilde{x}_i, \tilde{x}_k)$

Вводимо наступні умовні ймовірності:

$$p(x_j|x_i) = \frac{\exp\left(||x_i - x_j||^2/2\sigma^2\right)}{\sum_{k\neq i} \exp\left(||x_i - x_k||^2/2\sigma^2\right)}$$

$$q(\tilde{x}_j|\tilde{x}_i) = \frac{\exp\left(||\tilde{x}_i - \tilde{x}_j||^2/2\sigma^2\right)}{\sum_{k\neq i} \exp\left(||\tilde{x}_i - \tilde{x}_k||^2/2\sigma^2\right)}$$

Цей розподіл ураховує тільки співвідношення між відстанями. **Щоб зрівняти ці два розподіли**, необхідно використовувати функціонал, що знаходить розходження між двома імовірнісними розподілами. Для цього добре підходить **дивергенція Кульбака-Лейблера**:

$$\sum_{i=1}^{\ell} \sum_{j \neq i} p(x_j|x_i) \log \frac{p(x_j|x_i)}{q(\tilde{x}_j|\tilde{x}_i)} \to \min_{\tilde{x}_1,\ldots,\tilde{x}_\ell}$$

Для знаходження опису об'єкта в маловимірному ознаковому просторі **необхідно мінімізувати цей функціонал** (наприклад, методом стохастичного градієнтного спуску).

Метод t-SNE

У методі SNE є проблема: об'єкти в багатовимірному просторі легше розмістити поруч, чим у маловимірному, а **евклідова метрика занадто сильно штрафує за незбереження пропорцій**.

Метод t-SNE намагається вирішити цю проблему, використовуючи інший спосіб обчислення розподілів у новому просторі:

$$q(\tilde{x}_j|\tilde{x}_i) = \frac{\left(1 + \|\tilde{x}_i - \tilde{x}_j\|^2\right)^{-1}}{\sum_{k \neq i} \left(1 + \|\tilde{x}_i - \tilde{x}_k\|^2\right)^{-1}}$$

При використанні цього методу поділ виявляється набагато більше чітким.