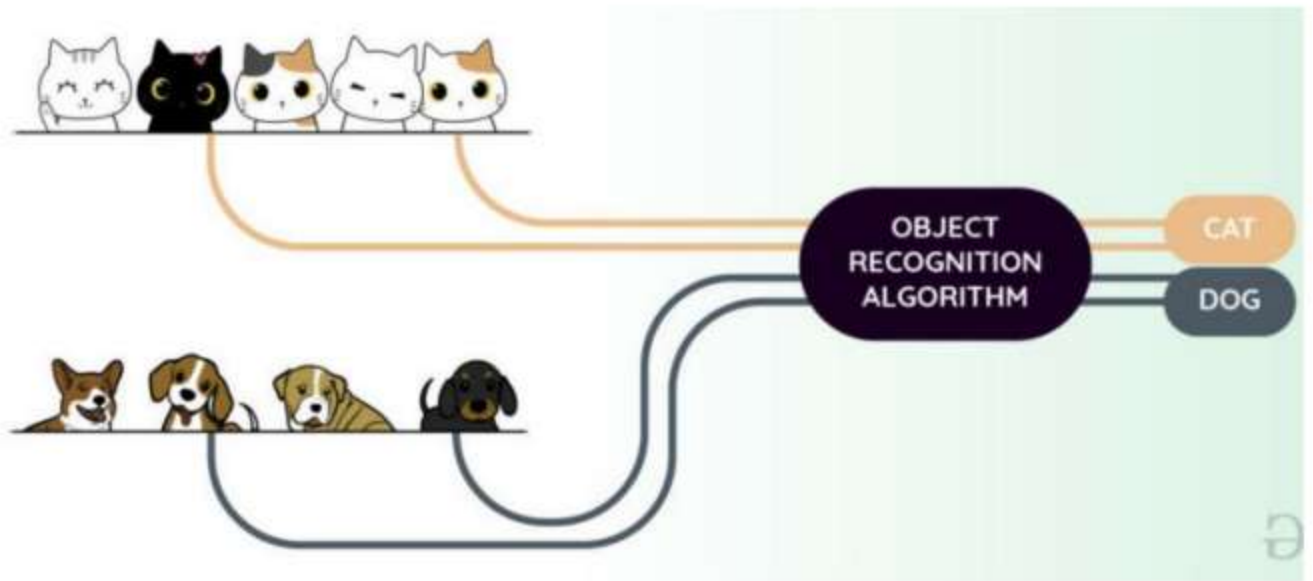


Розпізнавання образів



Сьогодні на лекції

1
PCO, комунікації,
види завдань

2
Статистичний
підхід

3
Метрики оцінки
якості
кластеризації

4
Задача
розпізнавання
образів

5
Методи вибору
ознак

6
Методи
кластеризації



PCO

$$\begin{aligned} & 3 \text{ лабораторних роботи} = 38 \text{ балів} \\ & \quad + \\ & 1 \text{ контрольна робота} = 22 \text{ бали} \\ & \quad + \\ & 4 \text{ домашні завдання} = 40 \text{ бали} \\ & \quad = \\ & 100 \text{ балів} \\ & \quad + \\ & \text{завдання на заохочувальні бали} \end{aligned}$$

PCO

Допуск до заліку: здані всі домашні та лабораторні роботи + 41 бал

Якщо допущені: менше 60 балів, то залік. Більше 60 ставлю ваші бали, або ви можете підвищити ваш бал - залік.



Орієнтовні строки здачі завдань

| завдання \ дата | 5.09- 18.09 | 19.09 - 2.10 | 3.10 - 16.10 | 17.10 - 30.10 | 31.10 - 13.11 | 14.11 - 27.11 | 28.11 - 11.12 | 12.12 - 25.12 | 26.12 - 31.12 |
|----------------------------|----------------|-----------------|-----------------|------------------|------------------|------------------|------------------|------------------|------------------|
| Домашнє завдання №1 | | | | | | | | | |
| Домашнє завдання №2 | | | | | | | | | |
| Домашнє завдання №3 | | | | | | | | | |
| Домашнє завдання №4 | | | | | | | | | |
| Лабораторна робота №1 | | | | | | | | | |
| Лабораторна робота №2 | | | | | | | | | |
| Лабораторна робота №3 | | | | | | | | | |
| Модульна контрольна робота | | | | | | | | | |

| | | |
|----------|----------------------|-----------------------|
| максимум | максимум мінус х бал | максимум мінус у бали |
|----------|----------------------|-----------------------|

в залежності
від
лабораторної
х від 1-2, у від
2-3

Комунікації

Гугл клас - <https://classroom.google.com/c/NTI2ODcyNDE1NTQ0?cjc=oamk2lr>

Пари за посиланням - <https://meet.google.com/cpm-jppz-obg>

Пошта - sharoval.nataliia@lil.kpi.ua

Телеграм - @nsharoval





Чи проходили ви курс інтелектуальний аналіз даних?

Так

Ні



Powered by  Poll Everywhere

Start the presentation to see live content. For screen share software, share the entire screen. Get help at pollev.com/app

Орієнтовні дати та теми лекцій

| Дата | Тема |
|-------|---|
| 6.09 | Вступ. Статистичний підхід. Кластерний аналіз. |
| 13.09 | Кластерний аналіз. |
| 20.09 | Ансамблі |
| 27.09 | Кольорові простори.Методи обробки зображень. Дескриптори та особливі точки. |
| 4.10 | Класичні методи сегментації. |
| 11.10 | Класичні методи розпізнавання облич |
| 18.10 | Виділення меж на зображенні |
| 25.10 | кр |
| 1.11 | Архітектури згорткових нейронних мереж |

Орієнтовні дати та теми лекцій

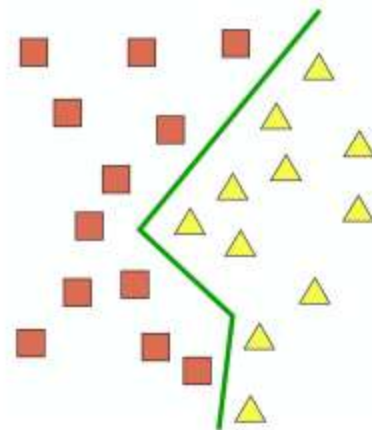
| Дата | Тема |
|-------|---|
| 7.11 | Архітектури згорткових нейронних мереж |
| 15.11 | Детекція об'єктів |
| 22.11 | Методи сегментації за допомогою ЗНМ |
| 29.11 | Генеративні згорткові мережі |
| 5.12 | Механізм уваги |
| 13.12 | Графові нейронні мережі |
| 20.12 | Розпізнавання мови. Розпізнавання рухів |
| 27.12 | Заключна лекція |
| 3.01 | залік |

Лекція 1.

Загальна характеристика розпізнавання образів. Статистичний підхід



Методи розпізнавання образів використовують:



Задача розпізнавання образів полягає в співвіднесенні вхідного образу x одному з класів ω_i .



Основні задачі теорії розпізнавання образів :

1. Математичний опис образів. Векторний простір - простір ознак.
2. Вибір найбільш інформативних ознак, що описують даний образ.
3. Опис класів розпізнаних образів.
4. Знаходження оптимальних вирішальних процедур (*методів класифікації*).
5. Оцінка достовірності класифікації образів .

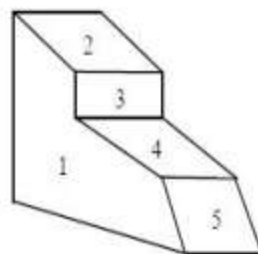
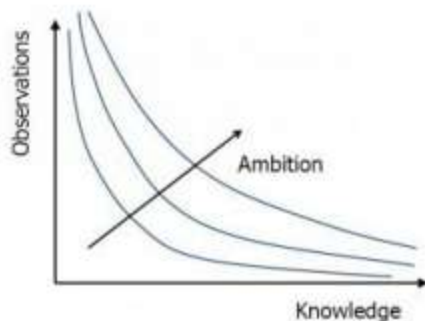


Рис. 1.3

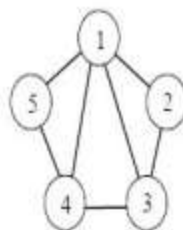


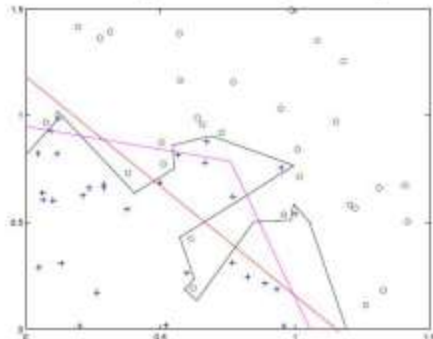
Рис. 1.4

Статистичний підхід в теорії РО

Поява того чи іншого образу є випадковою подією і ймовірність цієї події можна описати за допомогою закону розподілу ймовірностей цієї багатовимірної випадкової величини.

Знаючи елементи навчальної вибірки $X = \{x_1, \dots, x_N\}$, можна відновити ймовірнісні характеристики цього середовища.

Отже якщо вектори-образи є випадковими векторами, нам необхідно знайти таку вирішальну функцію, щоб помилка неправильної класифікації була мінімальною.



Ймовірнісні характеристики середовища

- функція щільності розподілу ймовірностей появи образу $f(\mathbf{x})$;
- умовні ймовірності приналежності деякого образу заданим класам $p(\omega_i | \mathbf{x}^0)$;
- ймовірності появи класів $p_i = p(\omega_i)$;
- функції умовних щільностей розподілу ймовірностей образів в середині класів $f_i(\mathbf{x}) = f(\mathbf{x} | \omega_i)$.

Непараметричне оцінювання

Ймовірнісні
характерист
ики

вирішальні функції



класифікатор



$$P_{\text{false}} \rightarrow \min$$

Залежно від кількості апріорної інформації про імовірнісних характеристиках середовища і про ціну неправильної класифікації, ймовірність такої класифікації може визначатися по-різному.

класифікатор - ?


σ, μ - ?

Постановка задачі баєсівської класифікації



Розглянемо спочатку для простоти випадок двох класів $\{w_1, w_2\}$. Задано деяке розбиття простору ознак R^n на дві області X_1 і X_2 : $X_1 \cap X_2 = \emptyset, X_1 \cup X_2 = R^n$

Причому, будемо вважати, що область X_i є областю переваги класу w_i $i=1,2$, тобто образ $x \in w_i$ якщо $x \in X_i$.
Тоді ймовірність неправильної класифікації можна обчислити за формулою

$$Q = \int_{X_1} p(w_2 | x) f(x) dx + \int_{X_2} p(w_1 | x) f(x) dx.$$


Наївний байєсівський класифікатор

Середня помилка буде мінімальною, якщо

$$X_1 := \{x \in R^n : p_1 f_1(x) > p_2 f_2(x)\},$$

$$X_2 := R^n \setminus X_1 = \{x \in R^n : p_2 f_2(x) > p_1 f_1(x)\}.$$

За формулою Байєса нерівність $p(\varpi_1 | x) > p(\varpi_2 | x)$
рівносильна нерівності $\frac{p_1 f_1(x)}{f(x)} > \frac{p_2 f_2(x)}{f(x)} \Leftrightarrow p_1 f_1(x) > p_2 f_2(x)$

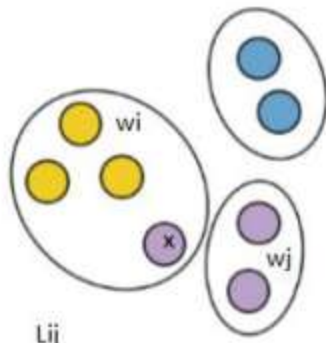
Звідки маємо: середня помилка
класифікації буде мінімальна якщо

$$X_1 := \{x \in R^n : p(\varpi_1 | x) > p(\varpi_2 | x)\},$$

$$X_2 := R^n \setminus X_1 = \{x \in R^n : p(\varpi_2 | x) > p(\varpi_1 | x)\}.$$

Або для довільного числа класів $x \in \varpi_i$,
якщо $p_i f_i(x) > p_j f_j(x)$ для всіх $j \neq i$.

Середні втрати



Оскільки образ x може належати будь-якому з W класов, що розглядаються, середня величина втрат, пов'язаних з віднесенням x до класу ω_j , дорівнює

$$r_j(x) = \sum_{k=1}^W L_{kj} p(\omega_k | x).$$

Відповідно до термінології теорії прийняття рішень, ця величина називається (умовним) середнім ризиком (або втратами).

Використовуючи формулу умовної ймовірності

$$p(A|B) = [p(A) p(B|A)] / p(B)$$

Функцію втрат можна записати як

$$r_j(\mathbf{x}) = \frac{1}{p(\mathbf{x})} \sum_{k=1}^W L_{kj} p(\mathbf{x} | \omega_k) P(\omega_k),$$

Оскільки множник $1/p(\mathbf{x})$ додатний та однаковий для всіх $r_j(\mathbf{x})$, $j = 1, 2, \dots, W$, його можна опустити, при цьому впорядкованість значень функцій $r_j(\mathbf{x})$ не зміниться. Тоді вираз для умовних середніх втрат (з точністю до постійного множника) зводиться до

$$r_j(\mathbf{x}) = \sum_{k=1}^W L_{kj} p(\mathbf{x} | \omega_k) P(\omega_k).$$

Класифікатор має можливість віднести невідомий образ, що надійшов до будь-якого з W класів. Якщо для кожного образу x обчислити функції $r_1(x), r_2(x), \dots, r_W(x)$ і приписати цей образ до того класу, для якого втрати мінімальні, то сумарне значення середніх втрат за всіма рішеннями буде мінімальним. Такий класифікатор, що мінімізує сумарну величину середніх втрат, називається Байєсівським класифікатором. Отже, байєсівський класифікатор відносить невідомий образ x до класу ω_i , якщо $r_i(x) < r_j(x)$ для $j = 1, 2, \dots, W; j \neq i$. Останню нерівність можна записати як:

$$\sum_{k=1}^W L_{ki} p(x|\omega_k) P(\omega_k) < \sum_{q=1}^W L_{qj} p(x|\omega_q) P(\omega_q) \text{ для всіх } j \neq i.$$

$L_{ij} = 1 - \delta_{ij}$, де $\delta_{ij} = 1$ при $i = j$ і $\delta_{ij} = 0$ при $i \neq j$,

Функцію δ_{ij} називають **симетричною або нуль-одичною функцією втрат**. Тоді середні втрати рівні:

$$r_j(x) = \sum_{k=1}^W (1 - \delta_{kj}) p(x | \omega_k) P(\omega_k) = p(x) - p(x | \omega_j) P(\omega_j).$$

Тоді байєсівський класифікатор приписує образ \mathbf{x} до класу ω_i , якщо

$$p(\mathbf{x}) - p(\mathbf{x} | \omega_i)P(\omega_i) < p(\mathbf{x}) - p(\mathbf{x} | \omega_j)P(\omega_j),$$

для всіх $j \neq i$

Бачимо, що байєсівський класифікатор у випадку нуль-одиничної функції втрат є не що інше, як обчислення дискримінантних функцій виду

$$p(\mathbf{x} | \omega_i)P(\omega_i) > p(\mathbf{x} | \omega_j)P(\omega_j) \quad j=1,2,\dots,W; j \neq i.$$

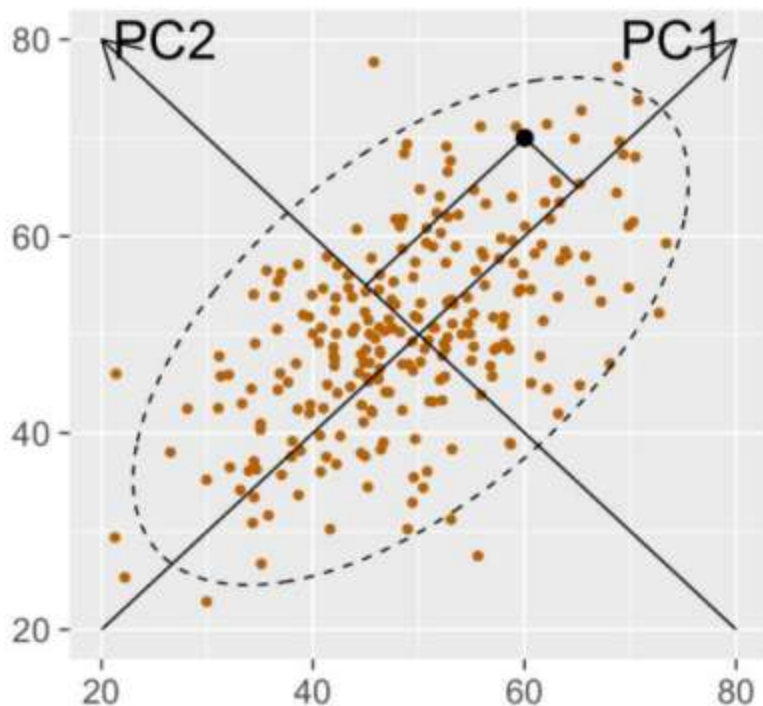
З віднесенням образу \mathbf{x} до того класу ω_i , для якого значення дискримінантної функції $d_i(\mathbf{x})$ виявиться найбільшим.

$$d_j(\mathbf{x}) = p(\mathbf{x} | \omega_j)P(\omega_j) \quad j=1,2,\dots,W$$

Методи вибору ознак



Задача зниження розмірності. Метод головних компонент



Метод главных компонент

$$R(\mathbf{x}) = \mathbf{x}\mathbf{x}^T.$$

$$R_i = \frac{1}{|\mathcal{O}_i|} \sum_{\mathbf{x} \in \mathcal{O}_i} R(\mathbf{x}),$$

$$R = \frac{1}{m} \sum_{i=1}^m R_i$$

$$\mathbf{x}' = S\mathbf{x}$$

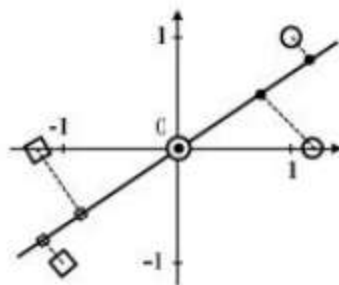
$$R' = \frac{1}{m} \sum_{i=1}^m R'_i = \frac{1}{m} \sum_{i=1}^m \frac{1}{|\mathcal{O}_i|} \sum_{\mathbf{x} \in \mathcal{O}_i} \mathbf{x}'\mathbf{x}'^T = \frac{1}{m} \sum_{i=1}^m \frac{1}{|\mathcal{O}_i|} \sum_{\mathbf{x} \in \mathcal{O}_i} S\mathbf{x}\mathbf{x}^T S^T = SRS^T.$$

Приклад

- Припустимо, що задані двовимірні образи – вектори $\mathbf{x}_4 = (-\sqrt{3}/2, 0)^T$, $\mathbf{x}_5 = (-1, -1)^T \in X_2$ і

$$\mathbf{x}_1 = (\sqrt{3}/2, 0)^T, \quad \mathbf{x}_2 = (0, 0)^T, \quad \mathbf{x}_3 = (1, 1)^T \in X_1$$

належать областям переваги X_1 і X_2 двох класів.



Знайдемо автокореляційні матриці образів в класах:

$$R_1 = \frac{1}{3}(\mathbf{x}_1\mathbf{x}_1^T + \mathbf{x}_2\mathbf{x}_2^T + \mathbf{x}_3\mathbf{x}_3^T) = \frac{1}{3}\left(\begin{pmatrix} 3/2 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\right) = \frac{1}{6}\begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix},$$

$$R_2 = \frac{1}{2}(\mathbf{x}_4\mathbf{x}_4^T + \mathbf{x}_5\mathbf{x}_5^T) = \frac{1}{2}\left(\begin{pmatrix} 3/2 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\right) = \frac{1}{4}\begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix}$$

і автокореляційну матрицю всієї вибірки

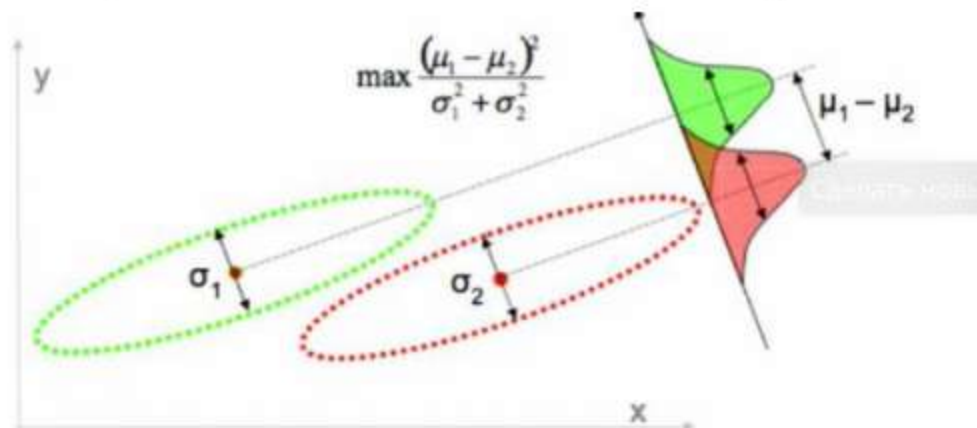
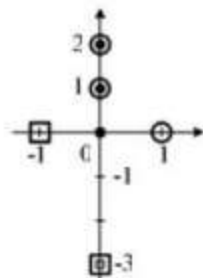
$$R = \frac{1}{2}(R_1 + R_2) = \frac{5}{24}\begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix}.$$

Знайдемо власні значення і власні вектори матриці R:

$$\lambda_1 = 5/4, \mathbf{e}_1 = \frac{1}{\sqrt{5}}(2,1)^T, \lambda_2 = 5/24, \mathbf{e}_2 = \frac{1}{\sqrt{5}}(-1,2)^T.$$

$$\mathbf{x}' = S\mathbf{x}, \text{ где } S = (\mathbf{e}_i^T): \mathbf{x}'_1 = \sqrt{30}/5, \mathbf{x}'_2 = 0, \mathbf{x}'_3 = 3\sqrt{20}/10, \\ \mathbf{x}'_4 = -\sqrt{30}/5, \mathbf{x}'_5 = -3\sqrt{20}/10$$

Лінійний дискримінант Фішера



Copyright © 2013 Victor Lavrenko

Будемо шукати проекції векторів на пряму з направляючим вектором w . Тоді $x' = w^T x$. Для знаходження w Фішер запропонував використовувати наступну функцію критерію

$$f(w) = \frac{|m'_1 - m'_2|^2}{s_1'^2 + s_2'^2}, \quad \text{де } m'_i = \frac{1}{|P_i|} \sum_{x \in P_i} x'$$

Вибіркові мат сподівання проекцій векторів i -го класа, $i = 1, 2$,

$$s_i'^2 = \sum_{x \in P_i} (x' - m'_i)^2$$

Розкид спроектованих вибірових значень в середині i -го класа, $i = 1, 2$.

Так як
$$\mathbf{m}'_i = \frac{1}{|\mathcal{O}_i|} \sum_{\mathbf{x} \in \mathcal{O}_i} \mathbf{x}' = \frac{1}{|\mathcal{O}_i|} \sum_{\mathbf{x} \in \mathcal{O}_i} \mathbf{w}^T \mathbf{x} = \mathbf{w}^T \mathbf{m}_i, \quad i = 1, 2.$$

то $|\mathbf{m}'_1 - \mathbf{m}'_2|^2 = |\mathbf{w}^T (\mathbf{m}_1 - \mathbf{m}_2)|^2 = \mathbf{w}^T (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{w} = \mathbf{w}^T S_m \mathbf{w}$, де $S_m = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T$
матриця розкиду між класами.

Аналогічно маємо
$$s_i'^2 = \sum_{\mathbf{x} \in \mathcal{O}_i} (\mathbf{x}' - \mathbf{m}'_i)^2 = \sum_{\mathbf{x} \in \mathcal{O}_i} (\mathbf{w}^T (\mathbf{x} - \mathbf{m}_i))^2 = \sum_{\mathbf{x} \in \mathcal{O}_i} \mathbf{w}^T (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T \mathbf{w} = \mathbf{w}^T S_i \mathbf{w},$$

$$\text{де } S_i = \sum_{\mathbf{x} \in \mathcal{O}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T, \quad i = 1, 2,$$

Матриці розкиду векторів в класах.

Нехай $S = S_1 + S_2$ – матриця розкиду
векторів всієї вибірки. Тоді $s_1'^2 + s_2'^2 = \mathbf{w}^T S \mathbf{w}$

і функція критерія :
$$f(\mathbf{w}) = \frac{\mathbf{w}^T S_m \mathbf{w}}{\mathbf{w}^T S \mathbf{w}}.$$

w – власний вектор узагальненої задачі
на власні значення $f(\mathbf{w}) \rightarrow \max \Leftrightarrow \mathbf{w} : S_m \mathbf{w} = \lambda S \mathbf{w}.$

якщо $S \neq 0$, то **w** – власний вектор
матриці $S^{-1}S_m$

Так як $S_m \mathbf{w} = k(\mathbf{m}_1 - \mathbf{m}_2)$, где $k = (\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{w}$, то

в якості вектора **w** можна взяти

$$\mathbf{w} = S^{-1}(\mathbf{m}_1 - \mathbf{m}_2).$$

Приклад

Нехай задані двомірні образи – вектори

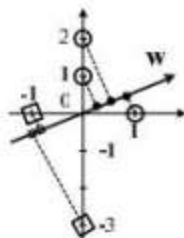
$$\mathbf{x}_1 = (0, 2)^T, \mathbf{x}_2 = (0, 1)^T, \mathbf{x}_3 = (1, 0)^T \in X_1 \text{ и } \mathbf{x}_4 = (-1, 0)^T, \mathbf{x}_5 = (0, -3)^T \in X_2$$

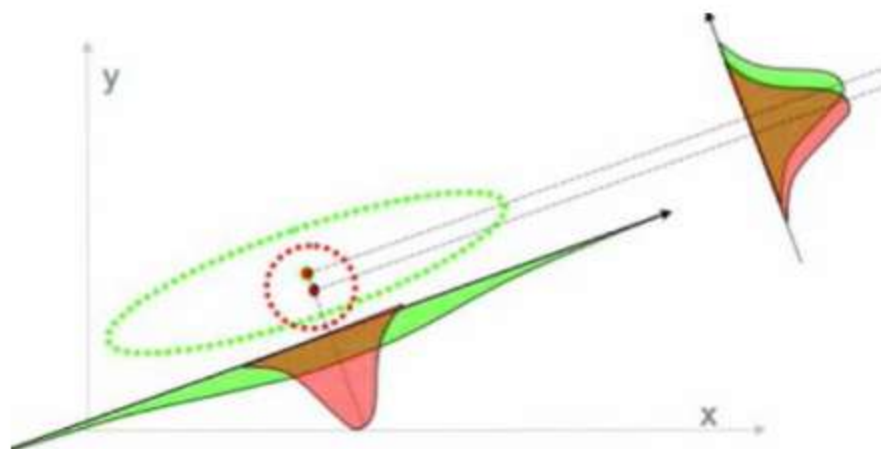
Тогда $\mathbf{m}_1 = (1/3, 1)^T$, $\mathbf{m}_2 = (-1/2, -3/2)^T$ и

$$S_1 = \frac{1}{3} \begin{pmatrix} 2 & -3 \\ -3 & 6 \end{pmatrix}, \quad S_2 = \frac{1}{2} \begin{pmatrix} 1 & -3 \\ -3 & 9 \end{pmatrix}, \quad S = S_1 + S_2 = \frac{1}{6} \begin{pmatrix} 7 & -15 \\ -15 & 39 \end{pmatrix},$$

$$\mathbf{w} = S^{-1}(\mathbf{m}_1 - \mathbf{m}_2) = \frac{1}{8} \begin{pmatrix} 39 & 15 \\ 15 & 7 \end{pmatrix} \cdot \frac{5}{6} \begin{pmatrix} 1 \\ 3 \end{pmatrix} = \frac{5}{4} \begin{pmatrix} 7 \\ 3 \end{pmatrix}.$$

$$\mathbf{x}'_1 = 15/2, \quad \mathbf{x}'_2 = 15/4, \quad \mathbf{x}'_3 = 35/4, \quad \mathbf{x}'_4 = -35/4, \quad \mathbf{x}'_5 = -45/4.$$





- Вибір ознак -

Обираємо **ознаки** з множини ознак
(включаючи будь-які нещодавно
розроблені) **не змінюючи** їх взагалі



- “Створення” ознак -

створення нових ознак з тих, які вже є

- Зменшення розмірності -

модифікує або перетворює елементи в
простір меншої розмірності



- Незалежне ранжування ознак за допомогою функції оцінки, обирається n найкращих



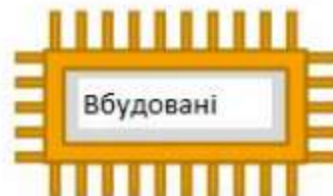
немає кореляції або надмірності



- Дослідження супернабору ознак, виміряти похибку узагальнення всіх підмножин
- Ціла комбінаторна задача оптимізації



Дилемма дослідження та використання



- Поєднання вибору ознак та навчання



немає кореляції або надмірності

Типи фільтрів

Одномірні фільтри оцінюють і класифікують одну ознаку за певними критеріями.

Вони розглядають кожну ознаку окремо та незалежно від простору ознак. Наприклад:

1. Ранжування ознаки за певними критеріями.
2. Вибір ознаки з найвищим рейтингом відповідно до критеріїв.

Багатовимірні фільтри, оцінюють весь простір ознак. Вони враховують зв'язки з іншими ознаками у наборі. Ці методи здатні обробляти дубльовані, надлишкові та корельовані ознаки.

Unsupervised

- Методи засновані на дисперсії
- Середня абсолютна різниця
- Критерій Лапласа

$$MAD_i = \frac{1}{n} \sum_{j=1}^n |X_{ij} - \bar{X}_i|$$

Supervised

- Методи засновані на кореляції (corr())
- Критерій Фішера
- Критерій хі-квадрат
- Корреляція Пірсона

Методи обгортки

X — початкова множина ознак, $|X|=n$

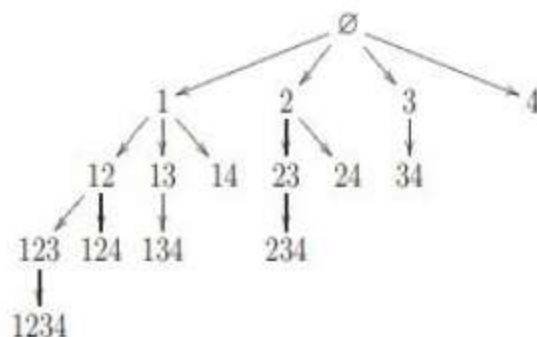
X' — підмножина початкової множини ознак, $X' \subseteq X$

$Q(X')$ — критерій якості навчання, наприклад:

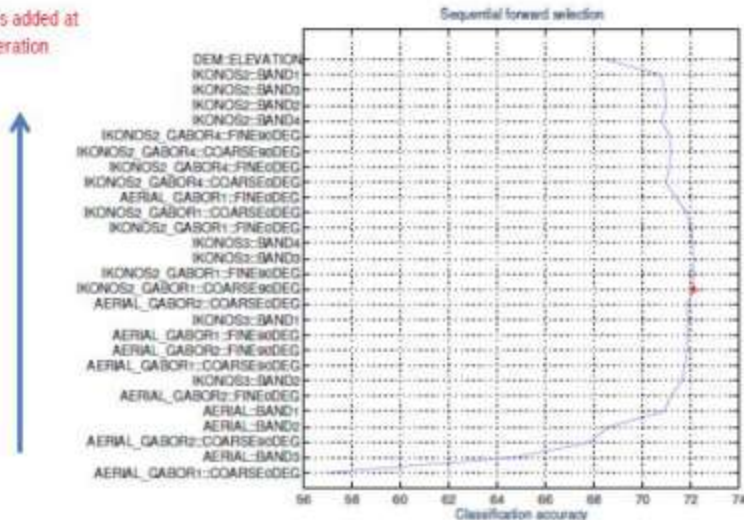
- Продуктивність моделі знижується.
- Продуктивність моделі зростає.
- Досягається заздалегідь визначена кількість ознак.

Наприклад, попередньо встановлені критерії можуть бути такими показниками, як ROC-AUC для класифікації або RMSE для регресії.

Послідовний прямий відбір (sequential forward selection)

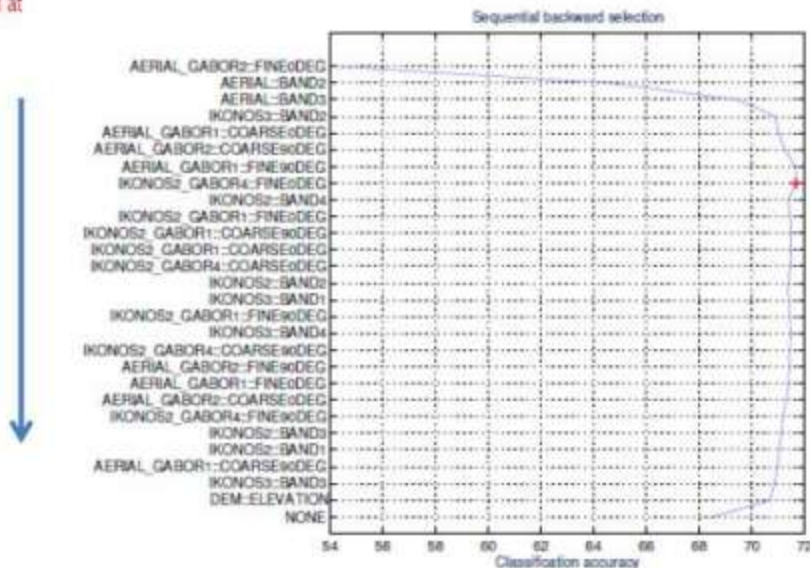


features added at
each iteration

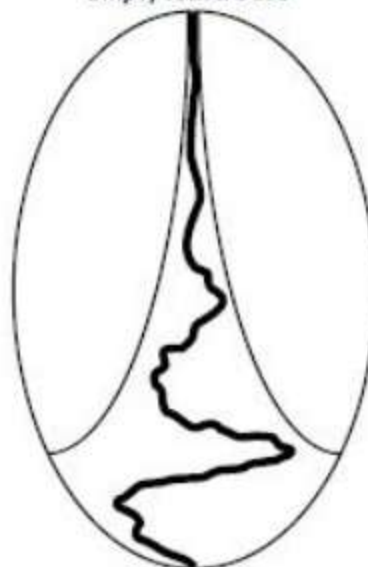


Послідовний зворотній відбір (sequential backward selection)

features removed at
each iteration



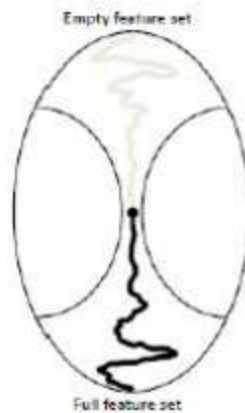
Empty feature set



Full feature set

Коли оптимальна
множина велика

- Двонаправлений пошук
- Повний перебір (exhaustive search)
- LRS або плюс-L, мінус-R
Якщо $L > R$, LRS починається з порожнього набору ознак:
 - Неодноразово додає **L** ознак
 - Неодноразово видаляє **R** ознак
- Послідовний плаваючий зворотній/прямий вибір (SFBS та SFFS)



Різниця між методами фільтра та обгортки



VS



Вбудовані методи

Вбудовані методи виконують вибір підмножини ознак в якості одного з етапів навчання, і тому специфічні для конкретної моделі. Переваги цих методів полягають в наступному:

- найкращим чином пристосовані до конкретної моделі;
- немає необхідності виділяти спеціальну тестову підмножину, на якій тестується поріг функції рангу для методів-фільтрів або виконується пошук найкращої підмножини для методів-обгортки;
- як наслідок з попереднього пункту, при використанні цього методу менше ризик перенавчання класифікатора

Регуляризація

існує три основних типи регуляризації для лінійних моделей:

- **лассо-регресія** або регуляризація L1
- **регресія хребта** або регуляризація L2
- **еластичні сітки** або регуляризація L1 / L2

L1 Regularization

$$\text{Cost} = \sum_{i=0}^N (y_i - \sum_{j=0}^M x_{ij} W_j)^2 + \lambda \sum_{j=0}^M |W_j|$$

L2 Regularization

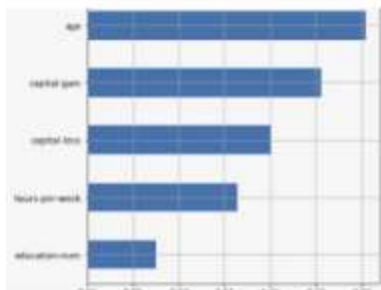
$$\text{Cost} = \underbrace{\sum_{i=0}^N (y_i - \sum_{j=0}^M x_{ij} W_j)^2}_{\text{Loss function}} + \lambda \underbrace{\sum_{j=0}^M W_j^2}_{\text{Regularization Term}}$$

Loss function

Regularization
Term

Комбінування методів

1. Використання методів фільтрів і обгортки
 2. Використання методів обгортки та вбудованих
- Наприклад:

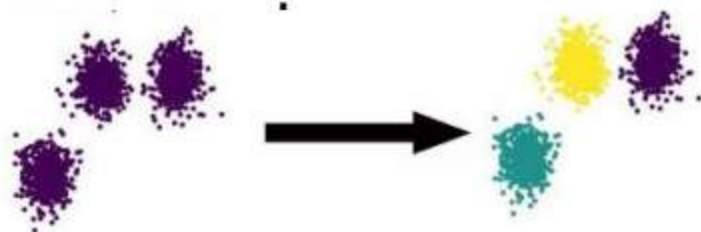


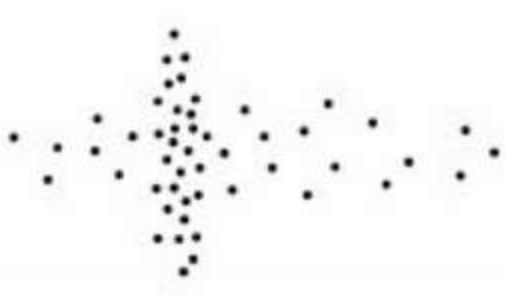
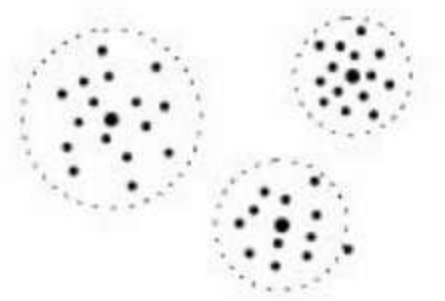
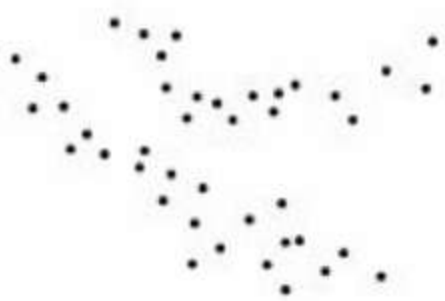
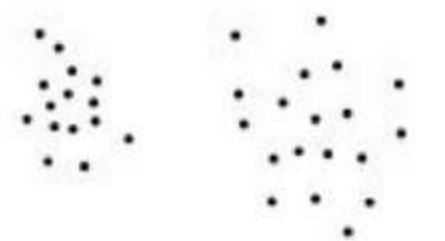
Задача кластеризації

Знайти таке розбиття навчальної вибірки $\Theta = \{x_1, \dots, x_N\}$ на непересічні підмножини (кластери) X_1, \dots, X_m : $X_1 \cup \dots \cup X_m = \Theta$, $X_i \cap X_j = \emptyset$ для будь-яких $i \neq j$, щоб всі точки одного кластера склалися зі «схожих» елементів, а точки різних кластерів істотно відрізнялися.

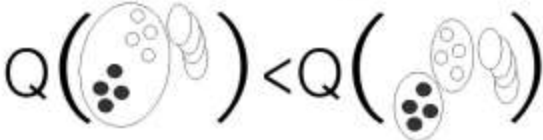
Основні параметри кластеризації:


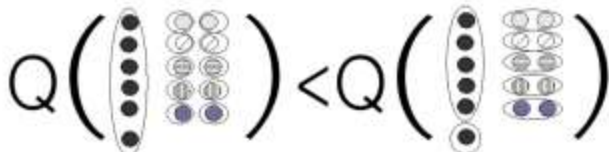
- критерій «схожості» елементів Q ;
- використовувана
- число кластерів.





Вимоги до метрики оцінки кластеризації

- Однорідність $Q(\text{Diagram 1}) < Q(\text{Diagram 2})$

- Повнота $Q(\text{Diagram 3}) < Q(\text{Diagram 4})$

- Rag Bag $Q(\text{Diagram 5}) < Q(\text{Diagram 6})$

- Розмір проти якості $Q(\text{Diagram 7}) < Q(\text{Diagram 8})$


Якість кластеризації

- Середня відстань в середині кластеру

$$Q^{(1)} = \sum_i \sum_{x,y \in X_i} d(x,y) \rightarrow \min;$$

- Середня міжкластерна відстань

$$Q^{(2)} = \sum_{i < j} \sum_{\substack{x \in X_i \\ y \in X_j}} d(x,y) \rightarrow \max;$$

- Сумарна вибіркова дисперсія розкиду елементів відносно центру кластерів

$$Q^{(3)} = \sum_i \frac{1}{|X_i|} \sum_{x \in X_i} d^2(x, c_i) \rightarrow \min, \text{ где } c_i = \frac{1}{|X_i|} \sum_{x \in X_i} x - \text{центр кластера } X_i.$$



Метрики в sklearn

| Зовнішні | Внутрішні |
|--|---|
| Adjusted Rand index Fowlkes-Mallows scores Mutual Information based scores Homogeneity Completeness V-measure | Silhouette Coefficient Davies-Bouldin Index Calinski-Harabasz Index * Dunn Index |
| | |