Основні типи задач класифікації.

У подальшому будуть розглянуті деякі постановки задач класифікації в залежності від об'єму апріорної інформації про вектор ознак $\vec{x} = \vec{x}(\omega) \in X$.

I тип. У найкращому випадку для вектора ознак $\vec{x}$ може бути доступна повна інформації про його розподіл.

II тип. Гірше, коли розподіл вектора ознак $\vec{x}$ заданий з точністю до деякого вектора невідомих параметрів. У такій ситуації цю невизначеність, як правило, компенсують наявністю об'єктів з відомими класифікаціями, тобто вважають, що нам доступна *навчальна вибірка (training sample)*. Тобто маємо задачу *класифікації з вчителем*. При розв'язанні такого типу задач класифікації у нагоді стане *дискримінантний аналіз*.

III тип. У найгіршому випадку будь-яка інформація про розподіл вектора ознак $\vec{x}$ може бути взагалі відсутня, а апріорна інформація про об'єкти, які належить класифікувати, вичерпується тільки інформацією, що об'єкти у кожному з класів близькі (схожі) у деякому розумінні. Якщо припускається, що навчальна вибірка відсутня, тоді кажуть, що мають справу з задачею *класифікації без вчителя*. Цей спектр задач розв'язують методами *кластерного аналізу*.

Задача класифікації у випадку m класів.

Почнемо розгляд задач класифікації з ситуації, коли для вектора ознак $\vec{x}$ доступна досить багата інформація, а саме відома функція щільності для кожного класу Ω_i, $i = \overline{1,m}$. Припустимо, що класи apriori визначені. Потрібно віднести кожен об'єкт ω до одного з цих класів.

Постановка задачі та основні припущення

Нехай маємо справу з об'єктами ω з m заданих класів $\{\Omega_i\}_{i=1}^{m}$, тобто

$$\Omega = \bigcup_{i=1}^{m} \Omega_i, \ \Omega_i \bigcap \Omega_j = \varnothing, \forall i \neq j.$$

Припустимо, що доступна така апріорна інформація:

I. для $n-$вимірного вектора ознак $\vec{x}=\vec{x}(\omega)\in X$ для кожного класу Ω_i відома функція щільності $p(\vec{x}/i), i=\overline{1,m}$; (3)

II. для кожного класу Ω_i відома апріорна ймовірність $p_i, i=\overline{1,m}$; $\sum_{i=1}^{m} p_i = 1$; (4)

III. відома невід'ємна *функція втрат* $C(j/i)$, яка дорівнює величині втрат (штрафу) при віднесенні об'єкту ω до $j-$ого класу, коли він насправді з $i-$ого класу $\left(i,j=\overline{1,m}\right)$, причому $C(i/i)=0, i=\overline{1,m}$. (5)

Необхідно побудувати правило прийняття рішення про віднесення об'єкту ω до одного з класів $\{\Omega_i\}_{i=1}^{m}$ на базі вектора ознак $\vec{x}=\vec{x}(\omega)$ так, щоб були найменшими середні втрати:

$$Q=\sum_{i=1}^{m}\left\{p_i\left[\sum_{j=1}^{m} C(j/i)P(j/i)\right]\right\}, \tag{6}$$

де $P(j/i)$- ймовірність віднесення об'єкту ω до $j-$ого класу, коли він насправді з $i-$ого класу $\left(i,j=\overline{1,m}\right)$.

Зауваження. Середні втрати при помилковій класифікації об'єктів з $i-$ого класу визначаються виразом

$$\sum_{j=1}^{m} C(j/i)P(j/i).$$

Зрозуміло, що

$$\sum_{j=1}^{m} P(j/i)=1,\ i=\overline{1,m}.$$

Причому необхідні ймовірності $P(j/i)$ можна підрахувати таким чином

$$P(j/i) = \int_{X_j} p(\vec{x}/i)d\vec{x}, \quad i,j = \overline{1,m}, \qquad (7)$$

де множини X_j згідно (1) мають вигляд

$$X_j = \left\{x:\ d_j(\vec{x}) > d_i(\vec{x}), \forall i \neq j\right\},\ j = \overline{1,m},$$

а $\left\{d_i(\vec{x})\right\}_{i=1}^{m}$ - відповідний набір дискримінантних функцій.

Фактично потрібно буде знайти такий набір функцій $\left\{d_i(\vec{x})\right\}_{i=1}^{m}$, який забезпечує розбиття простору ознак X на підмножини

$$\left\{X_i\right\}_{i=1}^{m} \left(X = \bigcup_{i=1}^{m} X_i,\ X_i \bigcap X_j = \varnothing, \forall i \neq j\right),$$

які не перетинаються, причому так, щоб середні втрати Q були найменшими.

Побудова дискримінантних функцій

Переходимо до знаходження набору дискримінантних функцій $\left\{d_i(x)\right\}_{i=1}^{m}$ таких, що забезпечують потрібне розбиття простору ознак X:

$$X = \bigcup_{i=1}^{m} X_i,\ X_i \bigcap X_j = \varnothing, \forall i \neq j; \qquad (8)$$

де $X_j = \left\{x:\ d_j(x) > d_i(x), \forall i \neq j\right\},\ j = \overline{1,m}$.

Проаналізуємо критерій якості (6). Для цього перепишемо його у іншому вигляді, прийнявши до уваги (7):

$$Q=\sum_{i=1}^{m}\left\{p_i\left[\sum_{j=1}^{m}C(j/i)P(j/i)\right]\right\}=$$

$$=\sum_{i=1}^{m}\left\{p_i\left[\sum_{j=1}^{m}C(j/i)\int_{X_j}p(\vec{x}/i)d\vec{x}\right]\right\}=$$

$$=\sum_{i=1}^{m}\left\{\sum_{j=1}^{m}\left[p_iC(j/i)\int_{X_j}p(\vec{x}/i)d\vec{x}\right]\right\}=$$

$$=\sum_{j=1}^{m}\left\{\sum_{i=1}^{m}\left[p_iC(j/i)\int_{X_j}p(\vec{x}/i)d\vec{x}\right]\right\}=$$

$$=\sum_{j=1}^{m}\left\{\sum_{i=1}^{m}\int_{X_j}\left[p_iC(j/i)p(\vec{x}/i)\right]d\vec{x}\right\}=$$

$$= \sum_{j=1}^{m} \left\{ \int_{X_j} \left(\sum_{i=1}^{m} \left[p_i C(j/i) p(\vec{x}/i) \right] \right) d\vec{x} \right\} =$$

$$= \sum_{j=1}^{m} \left\{ \int_{X_j} g_j(\vec{x}) d\vec{x} \right\}, \tag{9}$$

де в останньому перетворенні використано позначення

$$g_j(\vec{x}) = \sum_{i=1}^{m} \left[p_i C(j/i) p(\vec{x}/i) \right].$$

Припустимо, що для довільної підмножини індексів

$$1 \le j_1 < j_2 < \ldots < j_q \le m$$

міра Лебега множини

$$\left\{ \vec{x} : g_{j_1}(\vec{x}) = g_{j_2}(\vec{x}) = \ldots g_{j_q}(\vec{x}) \right\}$$

дорівнює нулеві.

Тоді враховуючи умову (8), неважко бачити, що вираз (9) досягає свого мінімуму, якщо підмножини $\{X_i\}_{i=1}^{m}$ вибрати у такому вигляді:

$$X_j = \left\{ x : \ g_j(\vec{x}) < g_i(\vec{x}), \forall i \ne j \right\}, \ j = \overline{1,m},$$

причому найменше значення функціонала буде дорівнювати

$$Q = \int_{X} \left[\min_{j=\overline{1,m}} g_j(\vec{x}) \right] d\vec{x}.$$

Це дозволяє взяти в якості потрібного набору дискримінантних функцій $\{d_i(\vec{x})\}_{i=1}^{m}$ функції виду:

$$d_j(\vec{x}) = -g_j(\vec{x}) = -\sum_{i=1}^{m} \left[p_i C(j/i) p(\vec{x}/i) \right], \ j = \overline{1,m}.$$

Висновок. Для задачі класифікації у випадку m класів $\{\Omega_i\}_{i=1}^{m}$ у припущеннях (3)-(5) шуканий набір дискримінантних функцій має вигляд

$$d_j(\vec{x}) = -\sum_{i=1}^{m} \left[p_i C(j/i) p(\vec{x}/i) \right], \ j = \overline{1,m}. \tag{10}$$

Випадок використання простої функції втрат

Формула (10) надає вигляд дискримінантних функцій для задачі класифікації у припущеннях (3)-(5). Проаналізуємо розв'язок вищезгаданої задачі у випадку, коли в якості *прикладу функції втрат* $C(j/i)$ вибрана така функція:

$$C(j/i)=c\left(1-\delta_{ji}\right), c>0, \tag{11}$$

де δ_{ji} - символ Кронекера, тобто

$$\delta_{ji}=\begin{bmatrix} 1, \ \textit{якщо} \ j=i, \\ 0, \ \textit{якщо} \ j\neq i. \end{bmatrix}$$

Почнемо з аналізу середніх втрат (6):

$$Q=\sum_{i=1}^{m}\left\{p_i\left[\sum_{j=1}^{m}C(j/i)P(j/i)\right]\right\}=$$

$$=\sum_{i=1}^{m}\left\{p_i\left[\sum_{j=1}^{m}c\left(1-\delta_{ji}\right)P(j/i)\right]\right\}=$$

$$=c\sum_{i=1}^{m}\left\{p_i\left[\sum_{\substack{j=1\\ j\neq i}}^{m}P(j/i)\right]\right\}=$$

$$=c\sum_{i=1}^{m}\left\{p_i\left[1-P(i/i)\right]\right\}=$$

$$=c\left\{1-\sum_{i=1}^{m}p_iP(i/i)\right\}.$$

Останній вираз приводить до висновку, що мінімізація функції середніх втрат (6) у випадку, коли функція втрат $C(j/i)$ має вигляд (11), звелася до мінімізація *ймовірності помилкової класифікації*

$$\sum_{i=1}^{m}\left\{p_i\left[\sum_{\substack{j=1\\j\neq i}}^{m}P(j/i)\right]\right\}=1-\sum_{i=1}^{m}p_iP(i/i). \tag{12}$$

У цій ситуації замість дискримінантних функцій (10) можна використати набір дискримінантних функцій більш простого вигляду.

Для цього спочатку проаналізуємо набір дискримінантних функцій (10):

$$\begin{aligned}d_j(\vec{x})&=-\sum_{i=1}^{m}\left[p_iC(j/i)p(\vec{x}/i)\right]=\\&=-\sum_{i=1}^{m}\left[p_ic\left(1-\delta_{ji}\right)p(\vec{x}/i)\right]=\\&=-c\sum_{\substack{i=1\\i\neq j}}^{m}\left[p_ip(\vec{x}/i)\right]=\\&=-c\left[p(\vec{x})-p_jp(\vec{x}/j)\right]=\\&=c\left[p_jp(\vec{x}/j)-p(\vec{x})\right],\ j=\overline{1,m}.\end{aligned} \tag{13}$$

Тут скористалися, тим що

$p(\vec{x})=\sum_{i=1}^{m}\left[p_ip(\vec{x}/i)\right]$, де $p(\vec{x})$ - щільність вектора ознак.

Тоді у випадку, який розглядається, області X_j будуть задаватися таким чином:

$$X_j = \left\{x:\ d_j(\vec{x}) > d_i(\vec{x}), \forall i \neq j\right\} =$$
$$= \left\{x:\ c\left[p_j p(\vec{x}/j) - p(\vec{x})\right] > c\left[p_i p(\vec{x}/i) - p(\vec{x})\right], \forall i \neq j\right\} =$$
$$= \left\{x:\ p_j p(\vec{x}/j) > p_i p(\vec{x}/i), \forall i \neq j\right\},\ j = \overline{1,m}. \quad (14)$$

Представлення (14) для областей $\left\{X_i\right\}_{i=1}^{m}$ дає змогу стверджувати, що замість дискримінантних функцій виду (13) можна використовувати дискримінантні функції вигляду

$$d_j^{(1)}(\vec{x}) = p_j p(\vec{x}/j),\ j = \overline{1,m}, \quad (15)$$

а області X_j будуть задаватися таким чином

$$X_j = \left\{x:\ d_j^{(1)}(\vec{x}) > d_i^{(1)}(\vec{x}), \forall i \neq j\right\},\ j = \overline{1,m}. \quad (16)$$

Висновок 1. Для задачі класифікації у випадку m класів $\left\{\Omega_i\right\}_{i=1}^{m}$ згідно критерію мінімуму ймовірності помилкової класифікації (12) вигляд потрібного набору дискримінантних функцій $\left\{d_i^{(1)}(\vec{x})\right\}_{i=1}^{m}$ спрощується і має представлення (15), а шукані області $\left\{X_i\right\}_{i=1}^{m}$ набувають вигляду (16).

Результат (14) дозволяє зробити також ще один корисний висновок.

Дійсно, згідно формули Байєса, апостеріорна ймовірність $p_{j/\vec{x}}$ для j-го класу Ω_j, після отримання вектора ознак $\vec{x}$, задається таким виразом

$$p_{j/\vec{x}} = \frac{p_j p(\vec{x}/j)}{p(\vec{x})},\ j = \overline{1,m}.$$

Тоді з (14) отримуємо

$$X_j = \left\{x:\ p_j p\left(\vec{x} / j\right) > p_i p\left(\vec{x} / i\right), \forall i \neq j\right\} =$$

$$= \left\{x:\ \frac{p_j p\left(\vec{x} / j\right)}{p\left(\vec{x}\right)} > \frac{p_i p\left(\vec{x} / i\right)}{p\left(\vec{x}\right)}, \forall i \neq j\right\} =$$

$$= \left\{x:\ p_{j/\vec{x}} > p_{i/\vec{x}}, \forall i \neq j\right\}.$$

У результаті можемо стверджувати таке.

Висновок 2. Розв'язуюче правило для задачі класифікації у випадку m класів $\left\{\Omega_i\right\}_{i=1}^{m}$ згідно критерію мінімуму ймовірності помилкової класифікації (12) можна розглядати як задачу класифікації для m класів згідно критерію *максимуму апостеріорної ймовірності.*

Випадок нормально розподілених векторів ознак

Додатково до припущень з попереднього розділу будемо вважати, що n- вимірні вектори ознак $\vec{x}$ є нормально розподіленими, а саме

$$\vec{x} = \vec{x}\left(\omega\right) \sim \mathcal{N}\left(\vec{m}_i, V_i\right), V_i > 0,\ \textit{якщо}\ \ \omega \in \Omega_i, i = \overline{1,m},$$

де параметри $\left\{\vec{m}_i, V_i\right\}_{i=1}^{m}$- задані.

Тобто

$$p\left(\vec{x} / i\right) = \frac{1}{\left(2\pi\right)^{\frac{n}{2}}\left(\det\left(V_i\right)\right)^{\frac{1}{2}}} e^{-\frac{1}{2}\left\|\vec{x} - \vec{m}_i\right\|^2_{V_i^{-1}}}, i = \overline{1,m},$$

де $\left\|\vec{x}\right\|^2_Q = \vec{x}^T Q\vec{x}, Q > 0.$

Тоді згідно (16) області X_j набувають вигляду

$$X_j = \left\{x:\ d_j^{(1)}(\vec{x}) > d_i^{(1)}(\vec{x}), \forall i \neq j\right\} =$$

$$= \left\{x:\ p_j p(\vec{x}/j) > p_i p(\vec{x}/i), \forall i \neq j\right\} =$$

$$= \left\{x:\ \frac{p_j}{(2\pi)^{\frac{n}{2}}\left(\det\left(V_j\right)\right)^{\frac{1}{2}}} e^{-\frac{1}{2}\left\|\vec{x}-\vec{m}_j\right\|^2_{V_j^{-1}}} > \right.$$

$$\left. > \frac{p_i}{(2\pi)^{\frac{n}{2}}\left(\det\left(V_i\right)\right)^{\frac{1}{2}}} e^{-\frac{1}{2}\left\|\vec{x}-\vec{m}_i\right\|^2_{V_i^{-1}}}, \forall i \neq j\right\} =$$

$$= \left\{x:\ \ln\left(p_j\right) - \frac{1}{2}\ln\left(\det\left(V_j\right)\right) - \frac{1}{2}\left\|\vec{x}-\vec{m}_j\right\|^2_{V_j^{-1}} > \right.$$

$$\left. > \left\{x: \ln\left(p_i\right) - \frac{1}{2}\ln\left(\det\left(V_i\right)\right) - \frac{1}{2}\left\|\vec{x}-\vec{m}_i\right\|^2_{V_i^{-1}}, \forall i \neq j\right\} =$$

$$= \left\{x:\ 2\ln\left(p_j\right) - \ln\left(\det\left(V_j\right)\right) - \left\|\vec{x}-\vec{m}_j\right\|^2_{V_j^{-1}} > \right.$$

$$\left. > \left\{x: 2\ln\left(p_i\right) - \ln\left(\det\left(V_i\right)\right) - \left\|\vec{x}-\vec{m}_i\right\|^2_{V_i^{-1}}, \forall i \neq j\right\},\ j = \overline{1,m}.$$

Останнє дозволяє замість дискримінантних функцій (15) використовувати дискримінантні функції вигляду

$$d_j^{(2)}(\vec{x}) = -\left\|\vec{x}-\vec{m}_j\right\|^2_{V_j^{-1}} + 2\ln\left(p_j\right) - \ln\left(\det\left(V_j\right)\right),\ j = \overline{1,m}. \qquad (17)$$

Висновок 1. У випадку нормально розподілених векторів ознак $\vec{x}$ можна скористатися набором дискримінантних функцій (17), які вже є квадратичними функціями.

В свою чергу області $\left\{X_j\right\}_{j=1}^{m}$ можна представити таким чином

$$X_j = \left\{x:\ d_j^{(2)}(\vec{x}) > d_i^{(2)}(\vec{x}), \forall i \neq j\right\},\ j = \overline{1,m}.$$

Окремо звернемо увагу на випадок, коли коваріаційні матриці нормальних розподілів векторів ознак для усіх класів однакові, тобто:

$$\vec{x} = \vec{x}(\omega) \sim \mathcal{N}\left(\vec{m}_i, V\right), V > 0,\ \textit{якщо}\ \ \omega \in \Omega_i, i = \overline{1,m}.$$

Це дозволяє дещо спростити вигляд потрібних дискримінантних функцій. Дійсно

$$\begin{aligned}
X_j &= \left\{x:\ d_j^{(2)}(\vec{x}) > d_i^{(2)}(\vec{x}), \forall i \neq j\right\} = \\
&= \Big\{x:\ 2\ln\left(p_j\right) - \ln\left(\det(V)\right) - \left\|\vec{x} - \vec{m}_j\right\|_{V^{-1}}^2 > \\
&\qquad > 2\ln\left(p_i\right) - \ln\left(\det(V)\right) - \left\|\vec{x} - \vec{m}_i\right\|_{V^{-1}}^2, \forall i \neq j\Big\} = \\
&= \Big\{x:\ 2\ln\left(p_j\right) - \left\|\vec{x}\right\|_{V^{-1}}^2 + 2\vec{x}^T V^{-1}\vec{m}_j - \left\|\vec{m}_j\right\|_{V^{-1}}^2 > \\
&\qquad > 2\ln\left(p_i\right) - \left\|\vec{x}\right\|_{V^{-1}}^2 + 2\vec{x}^T V^{-1}\vec{m}_i - \left\|\vec{m}_i\right\|_{V^{-1}}^2, \forall i \neq j\Big\} = \\
&= \Big\{x:\ \ln\left(p_j\right) + \vec{x}^T V^{-1}\vec{m}_j - \frac{1}{2}\left\|\vec{m}_j\right\|_{V^{-1}}^2 > \\
&\qquad > \ln\left(p_i\right) + \vec{x}^T V^{-1}\vec{m}_i - \frac{1}{2}\left\|\vec{m}_i\right\|_{V^{-1}}^2, \forall i \neq j\Big\},\ j = \overline{1,m}.
\end{aligned}$$

Таким чином, у цій ситуації замість дискримінантних функцій можна використовувати дискримінантні функції спрощеного виду

$$d_j^{(3)}(\vec{x}) = \vec{x}^T V^{-1}\vec{m}_j + \ln(p_j) - \frac{1}{2}\|\vec{m}_j\|_{V^{-1}}^2, \quad j = \overline{1,m}. \tag{18}$$

Висновок 2. Дискримінантні функції (18) у випадку, коли коваріаційні матриці вектора ознак для усіх класів однакові, вже є лінійними функціями від вектора ознак. Останнє суттєво спрощує їх геометричну інтерпретацію.

А це у свою чергу дозволяє потрібні області $\{X_j\}_{j=1}^m$ представити у такому вигляді

$$X_j = \left\{x:\ d_j^{(3)}(\vec{x}) > d_i^{(3)}(\vec{x}), \forall i \neq j\right\}, \quad j = \overline{1,m}.$$

Випадок двох класів

Поглянемо як можна спростити останні результати у випадку двох класів $\{\Omega_i\}_{i=1}^2$. Нехай припущення про нормальність вектора ознак $\vec{x}$ залишається в силі

$$\vec{x} = \vec{x}(\omega) \sim \mathcal{N}(\vec{m}_i, V_i), V_i > 0, \ \text{якщо} \ \omega \in \Omega_i, i = \overline{1,2}, \tag{19}$$

де параметри $\{\vec{m}_i, V_i\}_{i=1}^2$ вважаються відомими.

Для цієї ситуації, коли $m = 2$, дискримінантні функції мають вигляд (17), але можна обмежитися тільки однією дискримінантною функцією $d^{(2)}(\vec{x})$, яка згідно (2) набуває вигляду

$$\begin{aligned} d^{(2)}(\vec{x}) &= d_1^{(2)}(\vec{x}) - d_2^{(2)}(\vec{x}) = \\ &= -\|\vec{x}-\vec{m}_1\|^2_{V_1^{-1}} + 2\ln(p_1) - \ln(\det(V_1)) + \\ &\quad + \|\vec{x}-\vec{m}_2\|^2_{V_2^{-1}} - 2\ln(p_2) + \ln(\det(V_2)) = \\ &= \|\vec{x}-\vec{m}_2\|^2_{V_2^{-1}} - \|\vec{x}-\vec{m}_1\|^2_{V_1^{-1}} + 2\ln\left(\frac{p_1}{p_2}\sqrt{\frac{\det(V_2)}{\det(V_1)}}\right). \end{aligned}$$

Тобто ця єдина дискримінантна функція $d^{(2)}(\vec{x})$ залишається квадратичною функцією від вектора ознак.

Тоді відповідний алгоритм класифікації об'єкту ω буде мати таке представлення:

$$\left[\begin{array}{l} \text{якщо } d^{(2)}(\vec{x}(\omega)) > 0, \quad \text{то } \omega \in \Omega_1, \\ \text{якщо } d^{(2)}(\vec{x}(\omega)) < 0, \quad \text{то } \omega \in \Omega_2. \end{array}\right.$$

А області $\{X_j\}_{j=1}^{2}$ в свою чергу набувають вигляду

$$X_1 = \{x : d^{(2)}(\vec{x}) > 0\},$$
$$X_2 = \{x : d^{(2)}(\vec{x}) < 0\}.$$

Якщо додатково до припущення вважати, що коваріаційні матриці вектора ознак для цих класів однакові, тобто:

$$\vec{x} = \vec{x}(\omega) \sim \mathcal{N}(\vec{m}_i, V), V > 0, \text{ якщо } \omega \in \Omega_i, i = \overline{1,2},$$

то відповідна єдина дискримінантна функція $d^{(3)}(\vec{x})$ набуде більш простого вигляду

$$
\begin{aligned}
d^{(3)}(\vec{x}) &= d_1^{(3)}(\vec{x}) - d_2^{(3)}(\vec{x}) = \\
&= \vec{x}^T V^{-1}\vec{m}_1 + \ln(p_1) - \frac{1}{2}\|\vec{m}_1\|_{V^{-1}}^2 - \\
&\quad - \vec{x}^T V^{-1}\vec{m}_2 - \ln(p_2) + \frac{1}{2}\|\vec{m}_2\|_{V^{-1}}^2 = \\
&= \vec{x}^T V^{-1}(\vec{m}_1 - \vec{m}_2) + \ln\left(\frac{p_1}{p_2}\right) + \frac{1}{2}\left(\|\vec{m}_2\|_{V^{-1}}^2 - \|\vec{m}_1\|_{V^{-1}}^2\right) = \\
&= \vec{x}^T V^{-1}(\vec{m}_1 - \vec{m}_2) + \ln\left(\frac{p_1}{p_2}\right) + \frac{1}{2}(\vec{m}_2 - \vec{m}_1)^T V^{-1}(\vec{m}_2 + \vec{m}_1)
\end{aligned}
$$

В останньому переході була використана тотожність, яка легко перевіряється (на с/р):

$$
\|\vec{x}_1\|_Q^2 - \|\vec{x}_2\|_Q^2 = (\vec{x}_1 - \vec{x}_2)^T Q(\vec{x}_1 + \vec{x}_2),
$$

де $Q > 0$, $\vec{x}_1, \vec{x}_2$ - вектори відповідної розмірності.

А відповідний алгоритм класифікації об'єкту ω та вигляд областей $\left\{X_j\right\}_{j=1}^2$ записуються подібним чином, але вже через дискримінантну функцію $d^{(3)}(\vec{x})$.

[redacted]

Підрахування ймовірності помилкової класифікації у випадку двох класів

Якість розв'язання задачі класифікації можна оцінити по значенню ймовірності помилкової класифікації.

Продемонструємо її підрахування у ситуації, яка була розглянута останньою у попередньому розділі. Тобто у випадку двох класів $\{\Omega_i\}_{i=1}^2$, коли для вектора ознак $\vec{x}$ справедливо

$$\vec{x} = \vec{x}(\omega) \sim \mathcal{N}(\vec{m}_i, V), V > 0, \text{ якщо } \omega \in \Omega_i, i = \overline{1,2}.$$

Ймовірність помилкової класифікації у цій ситуації згідно (12) визначається таким чином

$$p_1 P(2/1) + p_2 P(1/2).$$

Залишилося тільки обчислити ймовірності $P(2/1), P(1/2)$, які згідно (7) задаються формулою

$$P(j/i) = \int_{X_j} p(\vec{x}/i)\,d\vec{x}, \quad i, j = \overline{1,2},$$

де

$$p(\vec{x}/i) = \frac{1}{(2\pi)^{\frac{n}{2}}(\det(V))^{\frac{1}{2}}} e^{-\frac{1}{2}\|\vec{x}-\vec{m}_i\|^2_{V^{-1}}}, i = \overline{1,2};$$

$$X_1 = \{x : d^{(3)}(\vec{x}) > 0\},$$

$$X_2 = \{x : d^{(3)}(\vec{x}) < 0\},$$

а $d^{(3)}(\vec{x}) = \vec{x}^T V^{-1}(\vec{m}_1 - \vec{m}_2) + \ln\left(\frac{p_1}{p_2}\right) + \frac{1}{2}(\vec{m}_2 - \vec{m}_1)^T V^{-1}(\vec{m}_2 + \vec{m}_1)$. (на с/р)

Згідно останнього співвідношення дискримінантна функція $d^{(3)}(\vec{x})$ є лінійне перетворення нормально розподіленого вектора ознак $\vec{x}$ і тому теж буде мати нормальний розподіл, а саме: … (на с/р), goto с.49.

Задача класифікації з вчителем

Розглянемо тепер постановку задачі класифікації об'єктів ω у випадку m класів

$$\{\Omega_i\}_{i=1}^{m} \quad \left(\Omega = \bigcup_{i=1}^{m} \Omega_i,\ \Omega_i \bigcap \Omega_j = \varnothing, \forall i \neq j\right),$$

але з більш меншим об'ємом апріорної інформації.

А саме вважаємо, що:

Ґ. для $n-$вимірного вектора ознак $\vec{x} = \vec{x}(\omega) \in X$ для кожного класу Ω_i функція щільності $p(\vec{x}, \alpha_i / i)$ відома з точністю до деякого вектора невідомих параметрів α_i, $i = \overline{1,m}$;

III. задана невід'ємна функція втрат $C(j/i)$ при віднесенні об'єкту ω до $j-$ого класу, коли він насправді з $i-$ого класу $\left(i, j = \overline{1,m}\right)$;

IV. доступна навчальна вибірка S, тобто набір векторів ознак з їх відомими однозначними класифікаціями; а саме

$$S = \bigcup_{i=1}^{m} S_i,\ S_i \bigcap S_j = \varnothing, \forall i \neq j,$$

де $S_i = \left\{\vec{x}^{ij} = \vec{x}^{ij}(\omega) : \omega \in \Omega_i, j = \overline{1,n_i}\right\}, i = \overline{1,m}, \quad N = \sum_{i=1}^{m} n_i$.

Тут S_i - підвибірка навчальної вибірки S, яка складається з векторів ознак $\vec{x}^{ij}(\omega)$, що відповідають об'єктам ω з класу Ω_i, $i = \overline{1,m}$.

А апріорні ймовірності p_i для кожного класу Ω_i залишилися невідомими, $i=\overline{1,m}$; $\sum_{i=1}^{m} p_i = 1$.

Таким чином наявна невизначеність у постановці задачі компенсована присутністю навчальної вибірки S. Тобто маємо задачу *класифікації з вчителем.*

Потрібно запропонувати таке правило прийняття рішення про віднесення об'єкту ω до одного з класів $\{\Omega_i\}_{i=1}^{m}$ на базі доступного вектора ознак $\vec{x}=\vec{x}(\omega)$, щоб середні втрати були найменшими.

Подібна задача класифікації, але при припущеннях (3)-(5) була розв'язана раніше. Скористаємося у подальшому цим результатом.

Для розв'язання поточної задачі пропонується використовувати *підстановочне розв'язуюче правило.* Суть його полягає у наступному.

Спочатку скористаємося навчальною вибіркою S для визначення:

- оцінки $\hat{p}_i = \frac{n_i}{N}$ для апріорної ймовірності p_i для кожного класу $\Omega_i, i=\overline{1,m}$;
- оцінки $\hat{\alpha}_i$ методом максимальної правдоподібності для вектора невідомих параметрів α_i, $i=\overline{1,m}$.

В результаті середні втрати, які потрібно мінімізувати, набувають вигляду

$$\tilde{Q} = \sum_{i=1}^{m}\left\{\hat{p}_i\left[\sum_{j=1}^{m} C(j/i)\tilde{P}(j/i)\right]\right\},$$

де

$$\tilde{P}(j/i) = \int_{X_j} p(\vec{x},\hat{\alpha}_i/i)d\vec{x}, \quad i,j=\overline{1,m}.$$

Але ця задача була розглянута раніше і її розв'язок згідно (10) має вигляд

$$\tilde{d}_j(\vec{x}) = -\sum_{i=1}^{m}\left[\hat{p}_i C(j/i)\, p(\vec{x}, \hat{\alpha}_i / i)\right],\; j = \overline{1,m}. \qquad (22)$$

У підсумку для розв'язання задачі класифікації об'єктів ω з m класів $\{\Omega_i\}_{i=1}^{m}$ в умовах наявності припущень I', III, IV можна скористатися набором дискримінантних функцій (22).

Приклад. Продемонструємо використання підстановочного розв'язуючого правила у випадку двох класів $\{\Omega_i\}_{i=1}^{2}$ з нормально розподіленими векторами ознак тобто, коли

$$\vec{x} = \vec{x}(\omega) \sim \mathcal{N}(\vec{m}_i, V_i), (V_i > 0),\; \textit{якщо}\; \omega \in \Omega_i, i = \overline{1,2}.$$

В якості невідомих параметрів у цій ситуації виступають пари $\{\vec{m}_i, V_i\}_{i=1}^{2}$.

Тоді, використовуючи навчальну вибірку S, маємо можливість визначити оцінки невідомих параметрів:

- $\hat{p}_i = \dfrac{n_i}{N}, i = \overline{1,2};$
- $\hat{\vec{m}}_i = \dfrac{1}{n_i}\sum_{j=1}^{n_i}\vec{x}^{ij}, \hat{V}_i = \dfrac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(\vec{x}^{ij} - \hat{\vec{m}}_i\right)\left(\vec{x}^{ij} - \hat{\vec{m}}_i\right)^T, i = \overline{1,2}.$

Залишилось тільки підставити ці оцінки у представлення (22) для функцій $\tilde{d}_j(\vec{x}),\; j = \overline{1,m}$.

Задача розпізнавання III типу

Познайомимося ще з одним колом задач розпізнавання, для яких доступний зовсім незначним є об'єм апріорної інформації про об'єкти, які підлягають розподілу по підмножинам. А саме відомо тільки те, що об'єкти у кожній такій підмножині є близькими (схожими) між собою згідно деякої міри, а об'єкти, які не є близькими (схожими), належать до різних таких підмножин. Тобто фактично потрібно розбити множину усіх об'єктів на деякі підмножини (*кластери*), які складаються з близьких (схожих) між собою об'єктів. Розв'язанням проблем такого плану і займається *кластерний аналіз (cluster analysis)*.

Отже, задача *кластерного аналізу* полягає в тому, що потрібно розбити множину усіх об'єктів на такі *кластери* (підмножини) так, щоб до кожної *кластеру* (підмножини) були віднесені тільки об'єкти близькі (схожі) між собою згідно вибраної міри близькості (схожості).

Більш детальна інформація про такі міри близькості та схожості буде наведена далі.

Зауважимо, що у загальному випадку кількість кластерів може бути заздалегідь невідомою.

Не можна також забувати і про те, що віднесення об'єкту до певного кластеру буде залежати від вибору одиниць виміру компонент вектора ознак.

Вхідна інформація у задачі кластерного аналізу

Доступна інформація про об'єкти, які підлягають кластерізації, може задаватися різним чином. Наведемо деякі можливі варіанти її представлення.

1) Для кожного об'єкту ω_i відомий його n-вимірний вектор ознак $\vec{x}_i$, $i = \overline{1,N}$. Тобто фактично на вході маємо матрицю, яка складається з векторів ознак
$$\left(\vec{x}_1 \quad \vec{x}_2 \quad \dots \quad \vec{x}_N\right).$$
2) Для кожної пари об'єктів ω_i та ω_j з векторами ознак $\vec{x}_i$ та $\vec{x}_j$ відповідно, відома відстань d_{ij} між ними. Інакше кажучи на вході маємо *матрицю відстаней (distance matrix)*
$$\begin{pmatrix} 0 & d_{12} & \dots & d_{1N} \\ d_{21} & 0 & \dots & d_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N1} & d_{N2} & \dots & 0 \end{pmatrix},$$

де $d_{ij} = d\left(\vec{x}_i, \vec{x}_j\right)$- відповідне значення *функції відстані (distance function)*, яка є дійсною функцією, що задовольняє відомим аксіомам:

1. $d\left(\vec{x}_i, \vec{x}_j\right) \geq 0, \forall \vec{x}_i, \vec{x}_j,$
2. $d\left(\vec{x}_i, \vec{x}_j\right) = 0 \Leftrightarrow \vec{x}_i = \vec{x}_j,$
3. $d\left(\vec{x}_i, \vec{x}_j\right) = d\left(\vec{x}_j, \vec{x}_i\right), \forall \vec{x}_i, \vec{x}_j,$
4. $d\left(\vec{x}_i, \vec{x}_j\right) \leq d\left(\vec{x}_i, \vec{x}_k\right) + d\left(\vec{x}_k, \vec{x}_j\right), \forall \vec{x}_i, \vec{x}_k, \vec{x}_j.$

Функцію відстані ще називають *метрикою (metric)*.

Зауваження 1. Очевидно, що функції відстані та схожості взаємно зв'язані. Дійсно, якщо оперують функцією відстані d_{ij}, то вона дозволяє побудувати деяку функцію схожості s_{ij}, наприклад таким чином:

$$s_{ij} = \frac{1}{1+d_{ij}}, \quad \forall i, j.$$

Зауваження 2. На перший погляд здається, що можна користуватися і зворотнім зв'язком

$$d_{ij} = \frac{1}{s_{ij}} - 1, \quad \forall i, j.$$

Тобто маючи у розпорядженні деяку функцію схожості s_{ij} запропонувати функцію відстані d_{ij}. Але з'ясувалося, що це у загальному випадку не так. А саме, не усі функції відстані побудовані таким чином задовольняють своїй останній аксіомі (нерівність трикутника), а тому не завжди можуть виступити у якості функції відстані.

Розглянемо деякі вектори ознак $\vec{x} = (x_1, x_2, \ldots, x_n)^T$ та $\vec{y} = (y_1, y_2, \ldots, y_n)^T$. Наведемо для них ряд прикладів функцій відстані (метрик).

1. *Манхеттенська відстань (Manhattan distance)*

$$d_1(\vec{x}, \vec{y}) = \sum_{i=1}^{n} |x_i - y_i|.$$

2. *Евклідова відстань (Euclidean distance)*

$$d_2(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

3. *Відстань Чебишова (Chebyshev distance)*(рівномірна метрика)

$$d_\infty(\vec{x}, \vec{y}) = \max_{i=\overline{1,n}} |x_i - y_i|.$$

4. *Відстань Мінковського (Minkowski distance)*

$$d_p(\vec{x},\vec{y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}, \ p \geq 1.$$

Зауважимо, що три попередніх приклада функцій відстані є частинними випадками функції відстані Мінковського $d_p(\vec{x},\vec{y})$ при $p=1$, $p=2$ та $p=\infty$ відповідно,

5. *Відстань Махаланобіса (Mahalanobis distance)*

$$d_{V^{-1}}^{(M)}(\vec{x},\vec{y}) = \sqrt{\|\vec{x}-\vec{y}\|_{V^{-1}}^2},$$

де V - матриця відповідної розмірності, $V > 0$.

Ієрархічний алгоритм

Одним із методів, який широко використовується на практиці при розв'язанні задачі кластерного аналізу, є ієрархічний алгоритм. Саме він знайшов своє широке застосування з огляду на його прозорість та простоту реалізації.

Нехай потрібно провести процедуру кластерізації для об'єктів ω_i, $i=\overline{1,N}$. Опишемо *ієрархічний алгоритм (hierarchical algorithm)* покроково.

1. Спочатку вважаємо, що кожен об'єкт ω_i уявляє собою окремий кластер Ω_i, $i=\overline{1,N}$. Тобто маємо $\Omega_i = \{\omega_i\}$, $i=\overline{1,N}$. А поточна кількість кластерів $\tilde{m}$ відповідно буде дорівнювати N.

2. Для довільної пари кластерів $\left\{\Omega_i, \Omega_j\right\}$ знаходимо міжкластерну відстань $D_{ij} = D\left(\Omega_i, \Omega_j\right), i < j$. (Приклади обчислення міжкластерної відстані будуть наведені у подальшому).
3. Знаходимо ту пару кластерів $\left\{\Omega_i, \Omega_j\right\}$ для якої міжкластерна відстань є найменшою і об'єднуємо ці кластери в один кластер.
4. Корегуємо поточну кількість кластерів $\tilde{m} := \tilde{m} - 1$.
5. Перевіряємо *умову зупинки ієрархічного алгоритму*. Якщо умова зупинки не виконується, то переходимо до кроку 2, інакше анулюємо останнє злиття, корегуємо значення $\tilde{m} := \tilde{m} + 1$ та зупиняємося. (Приклади умов зупинки ієрархічного алгоритму будуть наведені далі).

Зауваження. Якщо на третьому кроці алгоритму пар кластерів $\left\{\Omega_i, \Omega_j\right\}$ з найменшою міжкластерною відстанню виявиться декілька, то об'єднувати можна і декілька таких пар кластерів з відповідною корекцією скрізь поточної кількості кластерів $\tilde{m}$.

Приклади обчислення міжкластерної відстані

В залежності від метода, який використовується, конкретизується і формула для обчислення міжкластерної відстані. Наведемо деякі її приклади.

1. Метод найближчого сусіда

$$D_{ij} = D_{\min}\left(\Omega_i, \Omega_j\right) = \min_{\substack{\tilde{\omega}\in\Omega_i \\ \tilde{\tilde{\omega}}\in\Omega_j}} d\left(\vec{x}\left(\tilde{\omega}\right), \vec{x}\left(\tilde{\tilde{\omega}}\right)\right).$$

2. Метод найдальшого сусіда

$$D_{ij} = D_{\max}\left(\Omega_i, \Omega_j\right) = \max_{\substack{\tilde{\omega}\in\Omega_i \\ \tilde{\tilde{\omega}}\in\Omega_j}} d\left(\vec{x}\left(\tilde{\omega}\right), \vec{x}\left(\tilde{\tilde{\omega}}\right)\right).$$

3. Метод центроїдів

$$D_{ij} = D_c\left(\Omega_i, \Omega_j\right) = d\left(\bar{\bar{x}}_i, \bar{\bar{x}}_j\right)$$

де

$$\bar{\bar{x}}_k = \frac{1}{card\left(\Omega_k\right)} \sum_{\omega \in \Omega_k} \vec{x}(\omega), k = i, j.$$

4. Метод середнього зв'язку

$$D_{ij} = \bar{D}\left(\Omega_i, \Omega_j\right) = \frac{1}{card\left(\Omega_i\right) card\left(\Omega_j\right)} \sum_{\tilde{\omega} \in \Omega_i} \sum_{\tilde{\tilde{\omega}} \in \Omega_j} d\left(\vec{x}(\tilde{\omega}), \vec{x}\left(\tilde{\tilde{\omega}}\right)\right).$$

Варіанти умов зупинки ієрархічного алгоритму

В основі більшості умов зупинки ієрархічного алгоритму лежить міркування, що відстані між об'єктами всередині кластерів повинні бути меншими ніж між об'єктами з різних кластерів. Введемо деякі поняття.

Означення. Середня внутрішньо кластерна відстань кластеру Ω_i це числова характеристика $\bar{d}_{in}\left(\Omega_i\right)$, яка дорівнює середній відстані між різними об'єктами з кластеру Ω_i.

Означення. Середня зовнішньо кластерна відстань кластеру Ω_i це числова характеристика $\bar{d}_{out}\left(\Omega_i\right)$, яка дорівнює середній відстані від об'єктів з кластеру Ω_i до об'єктів з інших кластерів.

Наведемо деякі приклади широко вживаних умов зупинки ієрархічного алгоритму.

1. Порушується умова, що для довільного кластеру середня внутрішньо кластерна відстань менше середньої зовнішньо кластерної відстані, тобто умова зупинки має вигляд

$$\exists i \quad \bar{d}_{in}(\Omega_i) \geq \bar{d}_{out}(\Omega_i).$$

В цьому випадку кажуть, що мають справу з *кластерами типу згущення у середньому.*

2. Порушується умова, що для довільного кластеру середня внутрішньо кластерна відстань не менше ніж у $k\ (k>1)$ раз менша середньої зовнішньо кластерної відстані, тобто умова зупинки предстане у такому вигляді

$$\exists i \quad \bar{d}_{in}(\Omega_i) \geq \frac{\bar{d}_{out}(\Omega_i)}{k}.$$

Вважають, що у цій ситуації стикаються з *сильними кластерами.*

3. Нехай apriori задана загальна кількість кластерів m_*. Тоді алгоритм зупиняється, якщо останнє злиття кластерів привело до загальної кількості кластерів меншої ніж m_*. Тобто умова зупинки алгоритму набуває вигляду $\tilde{m} < m_*$.

[redacted]