Постановка задачи частичного обучения

Дано:

множество объектов X, множество классов Y;

$X^\ell = \{x_1, \dots, x_\ell\}$ — размеченная выборка (labeled data);
$\{y_1, \dots, y_\ell\}$

$X^k = \{x_{\ell+1}, \dots, x_{\ell+k}\}$ — неразмеченная выборка (unlabeled data).

Два варианта постановки задачи:

- *Частичное обучение* (semi-supervised learning):
  построить алгоритм классификации $a\colon X \to Y$.
- *Трансдуктивное обучение* (transductive learning):
  зная все $\{x_{\ell+1}, \dots, x_{\ell+k}\}$, получить метки $\{y_{\ell+1}, \dots, y_{\ell+k}\}$.

Типичные приложения:

классификация и каталогизация текстов, изображений, и т. п.

Кластеризация как задача дискретной оптимизации

Пусть $\rho(x, x')$ — *функция расстояния* между объектами.
Веса на парах объектов (близости): $w_{ij} = \exp(-\beta\rho(x_i, x_j))$,
где β — параметр.

Задача кластеризации:

$$\sum_{i=1}^{\ell+k} \sum_{j=i+1}^{\ell+k} w_{ij} \bigl[a_i \neq a_j\bigr] \to \min_{\{a_i \in Y\}}.$$

Задача частичного обучения:

$$\sum_{i=1}^{\ell+k} \sum_{j=i+1}^{\ell+k} w_{ij} \bigl[a_i \neq a_j\bigr] + \lambda \sum_{i=1}^{\ell} \bigl[a_i \neq y_i\bigr] \to \min_{\{a_i \in Y\}}.$$

где λ — ещё один параметр.

Задано число кластеров K.

Графовый алгоритм КНП (кратчайший незамкнутый путь)

1: Найти пару вершин $(x_i, x_j) \in X^{\ell+k}$ с наименьшим $\rho(x_i, y_j)$ и соединить их ребром;
2: **пока** в выборке остаются изолированные точки
3: найти изолированную точку, ближайшую к некоторой неизолированной;
4: соединить эти две точки ребром;
5: удалить $K - 1$ самых длинных рёбер;

Задача частичного обучения: земенить только шаг 5...

Задано число кластеров K.

Графовый алгоритм КНП (кратчайший незамкнутый путь)

1: Найти пару вершин $(x_i, x_j) \in X^{\ell+k}$ с наименьшим $\rho(x_i, y_j)$ и соединить их ребром;
2: **пока** в выборке остаются изолированные точки
3: найти изолированную точку, ближайшую к некоторой неизолированной;
4: соединить эти две точки ребром;
5: ~~удалить $K - 1$ самых длинных рёбер;~~
6: **пока** есть путь между двумя вершинами разных классов
7: удалить самое длинное ребро на этом пути.

Алгоритм иерархической кластеризации (Ланс, Уильямс, 1967)

1: $C_1 := \bigl\{\{x_1\}, \dots, \{x_{\ell+k}\}\bigr\}$ — все кластеры 1-элементные;
   $R_{\{x_i\}\{x_j\}} := \rho(x_i, x_j)$ — расстояния между ними;
2: **для всех** $t = 2, \dots, \ell + k$ $\quad$ (t — номер итерации):
3: $\quad$ найти в C_{t-1} пару кластеров (U, V) с минимальным R_{UV};
4: $\quad$ слить их в один кластер:
   $\quad W := U \cup V$;
   $\quad C_t := C_{t-1} \cup \{W\} \setminus \{U, V\}$;
5: $\quad$ **для всех** $S \in C_t$
6: $\quad\quad$ вычислить R_{WS} по формуле Ланса-Уильямса:
   $\quad\quad R_{WS} := \alpha_U R_{US} + \alpha_V R_{VS} + \beta R_{UV} + \gamma \bigl| R_{US} - R_{VS} \bigr|$;

Алгоритм иерархической кластеризации (Ланс, Уильямс, 1967)

1: $C_1 := \{\{x_1\}, \dots, \{x_{\ell+k}\}\}$ — все кластеры 1-элементные;
$R_{\{x_i\}\{x_j\}} := \rho(x_i, x_j)$ — расстояния между ними;
2: **для всех** $t = 2, \dots, \ell + k$ (t — номер итерации):
3: найти в C_{t-1} пару кластеров (U, V) с минимальным R_{UV},
при условии, что в $U \cup V$ нет объектов с разными метками;
4: слить их в один кластер:
$W := U \cup V$;
$C_t := C_{t-1} \cup \{W\} \setminus \{U, V\}$;
5: **для всех** $S \in C_t$
6: вычислить R_{WS} по формуле Ланса-Уильямса:
$R_{WS} := \alpha_U R_{US} + \alpha_V R_{VS} + \beta R_{UV} + \gamma |R_{US} - R_{VS}|$;

1: начальное приближение центров μ_y, $y \in Y$;
2: **повторять**
3: **E-шаг:**
отнести каждый x_i к ближайшему центру:
$y_i := \arg\min_{y \in Y} \rho(x_i, \mu_y), \ \ i = 1, \dots, \ell + k;$
4: **M-шаг:**
вычислить новые положения центров:
$$\mu_y := \frac{\sum_{i=1}^{\ell+k} [y_i = y]\, x_i}{\sum_{i=1}^{\ell+k} [y_i = y]}, \ \text{для всех } y \in Y;$$
5: **пока** y_i не перестанут изменяться;

1: начальное приближение центров μ_y, $y \in Y$;

2: **повторять**

3: **E-шаг:**

отнести каждый $x_i \in X^k$ к ближайшему центру:

$$y_i := \arg\min_{y \in Y} \rho(x_i, \mu_y), \quad i = \ell+1, \dots, \ell+k;$$

4: **M-шаг:**

вычислить новые положения центров:

$$\mu_y := \frac{\sum_{i=1}^{\ell+k} [y_i = y]\, x_i}{\sum_{i=1}^{\ell+k} [y_i = y]}, \quad \text{для всех } y \in Y;$$

5: **пока** y_i не перестанут изменяться;

Резюме

- Задача SSL занимает промежуточное положение между классификацией и кластеризацией, но не сводится к ним.
- Простые методы-обёртки требуют многократного обучения, что вычислительно неэффективно.
- *Методы кластеризации* легко адаптируются к SSL путём введения ограничений (constrained clustering), но, как правило, вычислительно трудоёмки.
- *Методы классификации* адаптируются сложнее, но приводят к более эффективному частичному обучению.