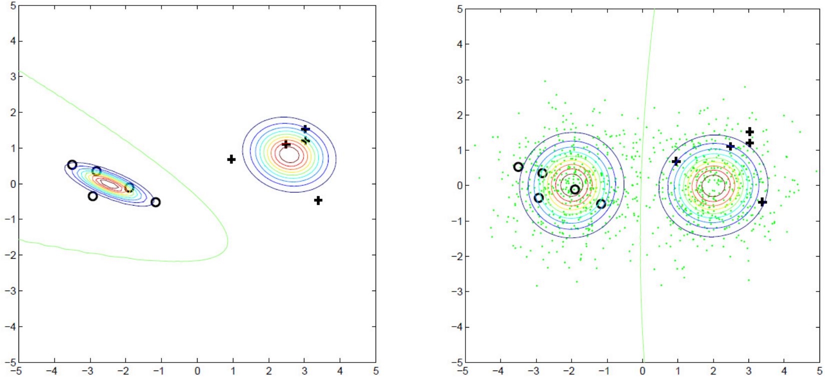
Дано:
множество объектов X, множество классов Y;
$X^\ell = \{x_1, \dots, x_\ell\}$ — размеченная выборка (labeled data);
$\{y_1, \dots, y_\ell\}$
$X^k = \{x_{\ell+1}, \dots, x_{\ell+k}\}$ — неразмеченная выборка (unlabeled data).

Два варианта постановки задачи:

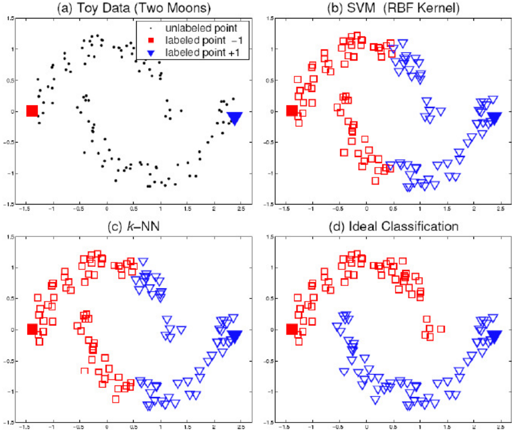
- *Частичное обучение* (semi-supervised learning): построить алгоритм классификации $a\colon X \to Y$.
- *Трансдуктивное обучение* (transductive learning): зная все $\{x_{\ell+1}, \dots, x_{\ell+k}\}$, получить метки $\{y_{\ell+1}, \dots, y_{\ell+k}\}$.

Типичные приложения:
классификация и каталогизация текстов, изображений, и т. п.

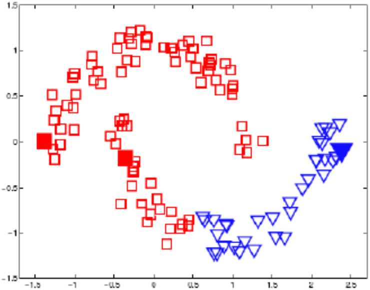
Пример 1. плотности классов, восстановленные:
по размеченным данным X^ℓ по полным данным $X^{\ell+k}$



Пример 2. Методы классификации не учитывают кластерную структуру неразмеченных данных



Пример 3. Методы кластеризации не учитывают приоритетность разметки.



Пусть $\mu\colon X^\ell \to a$ — произвольный метод обучения;
классификаторы имеют вид $a(x) = \arg\max\limits_{y\in Y} \Gamma_y(x)$;
Псевдоотступ — степень уверенности классификации $a_i = a(x_i)$:

$$M_i(a) = \Gamma_{a_i}(x_i) - \max_{y\in Y\setminus a_i} \Gamma_y(x_i).$$

Алгоритм self-training — обёртка (wrapper) над методом μ:

1: $Z := X^\ell$;
2: **пока** $|Z| < \ell + k$
3: $\quad a := \mu(Z)$;
4: $\quad \Delta := \{x_i \in X^k \setminus Z \mid M_i(a) \geqslant M_0\}$;
5: $\quad y_i := a(x_i)$ для всех $x_i \in \Delta$;
6: $\quad Z := Z \cup \Delta$;

M_0 можно определять, например, из условия $|\Delta| = 0.05\,k$

Метод co-training (Blum, Mitchell, 1998)

Пусть $\mu_1\colon X^\ell \to a_1$, $\mu_2\colon X^\ell \to a_2$ — два существенно различных метода обучения, использующих
— либо разные наборы признаков;
— либо разные парадигмы обучения (inductive bias);
— либо разные источники данных $X_1^{\ell_1}$, $X_2^{\ell_2}$.

$Z_1 := X_1^{\ell_1}$; $Z_2 := X_2^{\ell_2}$;
пока $|Z_1 \cup Z_2| < \ell + k$

$\quad a_1 := \mu_1(Z_1)$; $\Delta_1 := \{x_i \in X^k \backslash Z_1 \backslash Z_2 \mid M_i(a_1) \geqslant M_{01}\}$;
$\quad y_i := a(x_i)$ для всех $x_i \in \Delta_1$;
$\quad Z_2 := Z_2 \cup \Delta_1$;

$\quad a_2 := \mu_2(Z_2)$; $\Delta_2 := \{x_i \in X^k \backslash Z_1 \backslash Z_2 \mid M_i(a_2) \geqslant M_{02}\}$;
$\quad y_i := a(x_i)$ для всех $x_i \in \Delta_2$;
$\quad Z_1 := Z_1 \cup \Delta_2$;

Метод co-learning (deSa, 1993)

Пусть $\mu_t\colon X^\ell \to a_t$ — разные методы обучения, $t = 1, \ldots, T$.

Алгоритм co-learning — это self-training для композиции — простого голосования базовых алгоритмов $a_1, \ldots, a_T$:

$$a(x) = \arg\max_{y \in Y} \Gamma_y(x), \quad \Gamma_y(x_i) = \sum_{t=1}^{T} \bigl[a_t(x_i) = y\bigr].$$

тогда $M_i(a)$ — степень уверенности классификации $a(x_i)$.

1: $Z := X^\ell$;
2: **пока** $|Z| < \ell + k$
3: $\quad a := \mu(Z)$;
4: $\quad \Delta := \bigl\{x_i \in X^k \backslash Z \bigm| M_i(a) \geqslant M_0\bigr\}$;
5: $\quad y_i := a(x_i)$ для всех $x_i \in \Delta$;
6: $\quad Z := Z \cup \Delta$;

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.