Кластеризация и частичное обучение

К. В. Воронцов
vokov@forecsys.ru

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Машинное обучение (курс лекций, К.В.Воронцов)»

11 мая 2017

Содержание

Постановка задачи кластеризации

Дано:

X — пространство объектов;

$X^\ell = \{x_1, \ldots, x_\ell\}$ — обучающая выборка;

$\rho\colon X \times X \to [0, \infty)$ — функция расстояния между объектами.

Найти:

Y — множество кластеров,

$a\colon X \to Y$ — алгоритм кластеризации,

такие, что:

— каждый кластер состоит из близких объектов;

— объекты разных кластеров существенно различны.

Это задача *обучения без учителя* (unsupervised learning).

Некорректность задачи кластеризации

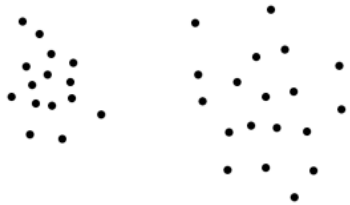
Решение задачи кластеризации принципиально неоднозначно:

- точной постановки задачи кластеризации нет;
- существует много критериев качества кластеризации;
- существует много эвристических методов кластеризации;
- число кластеров $|Y|$, как правило, неизвестно заранее;
- результат кластеризации сильно зависит от метрики ρ, выбор которой также является эвристикой.

Цели кластеризации

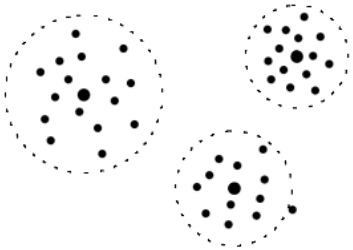
- **Упростить дальнейшую обработку данных**,
  разбить множество X^ℓ на группы схожих объектов
  чтобы работать с каждой группой в отдельности
  (задачи классификации, регрессии, прогнозирования).
- **Сократить объём хранимых данных**,
  оставив по одному представителю от каждого кластера
  (задачи сжатия данных).
- **Выделить нетипичные объекты**,
  которые не подходят ни к одному из кластеров
  (задачи одноклассовой классификации).
- **Построить иерархию множества объектов**
  (задачи таксономии).

Типы кластерных структур



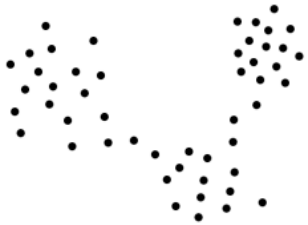
внутрикластерные расстояния, как правило, меньше межкластерных

ленточные кластеры



кластеры с центром

Типы кластерных структур



кластеры могут соединяться перемычками

кластеры могут накладываться на разрежен-
ный фон из редко расположенных объектов

кластеры могут перекрываться

Типы кластерных структур

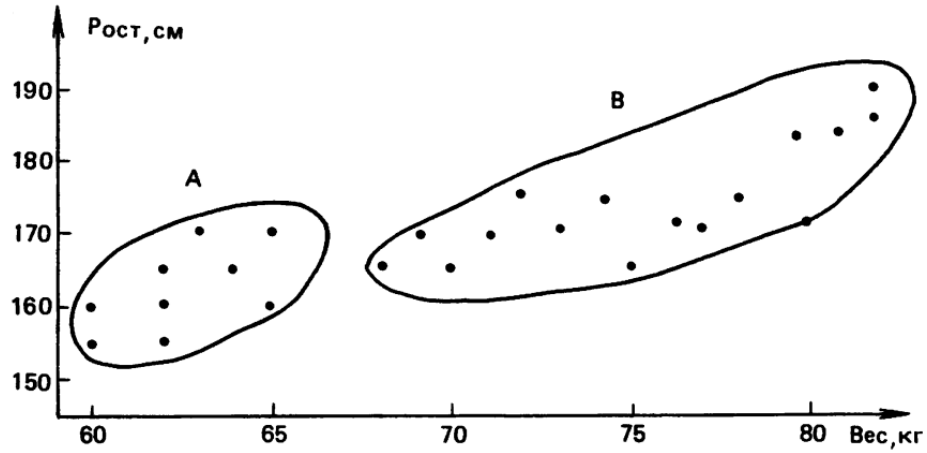
кластеры могут образовываться не по сходству, а по иным типам регулярностей

кластеры могут вообще отсутствовать

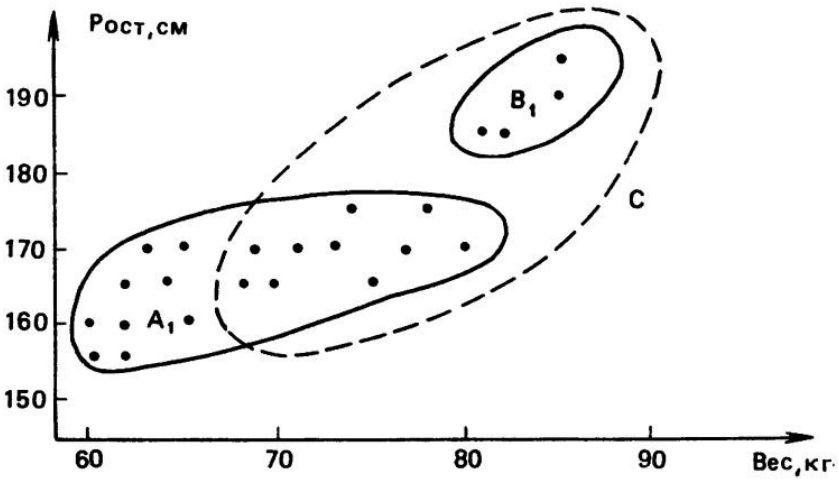
- Каждый метод кластеризации имеет свои ограничения и выделяет кластеры лишь некоторых типов.
- Понятие «тип кластерной структуры» зависит от метода и также не имеет формального определения.

Проблема чувствительности к выбору метрики

Результат зависит от нормировки признаков:



A — студентки,
B — студенты



после перенормировки
(сжали ось «вес» вдвое)

Постановка задачи частичного обучения

Дано:

множество объектов X, множество классов Y;

$X^k = \{x_1, \ldots, x_k\}$ — размеченные объекты (labeled data);
$\{y_1, \ldots, y_k\}$

$U = \{x_{k+1}, \ldots, x_\ell\}$ — неразмеченные объекты (unlabeled data).

Два варианта постановки задачи:

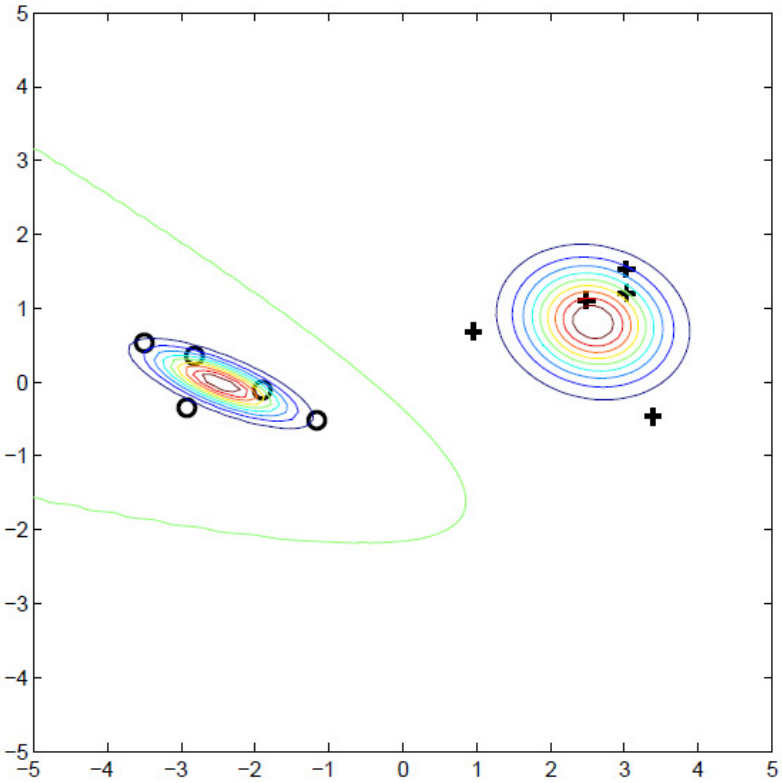
- *Частичное обучение* (semi-supervised learning):
  построить алгоритм классификации $a\colon X \to Y$.
- *Трансдуктивное обучение* (transductive learning):
  зная все $\{x_{k+1}, \ldots, x_\ell\}$, получить метки $\{a_{k+1}, \ldots, a_\ell\}$.

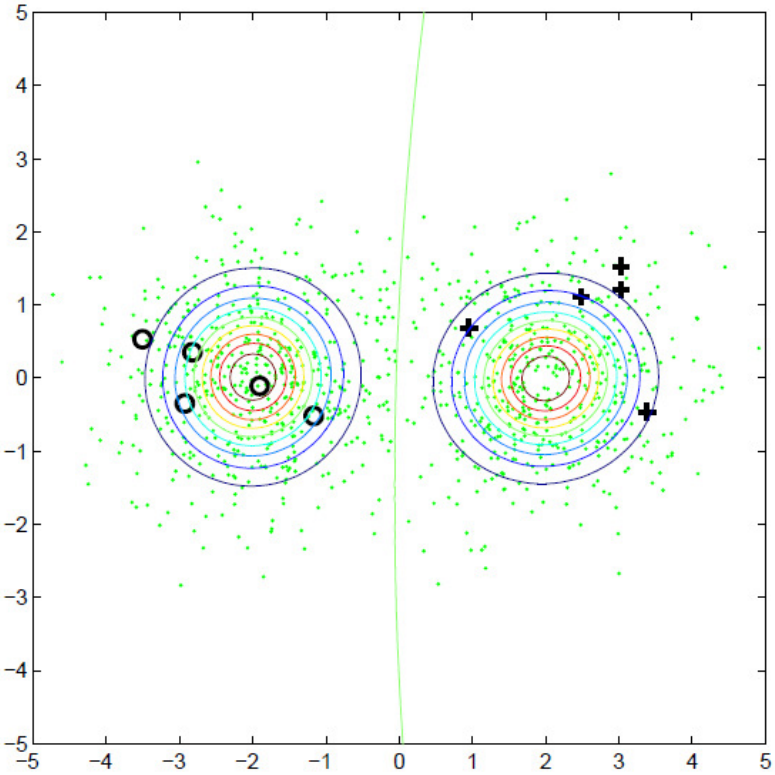
Типичные приложения:

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

SSL не сводится к классификации

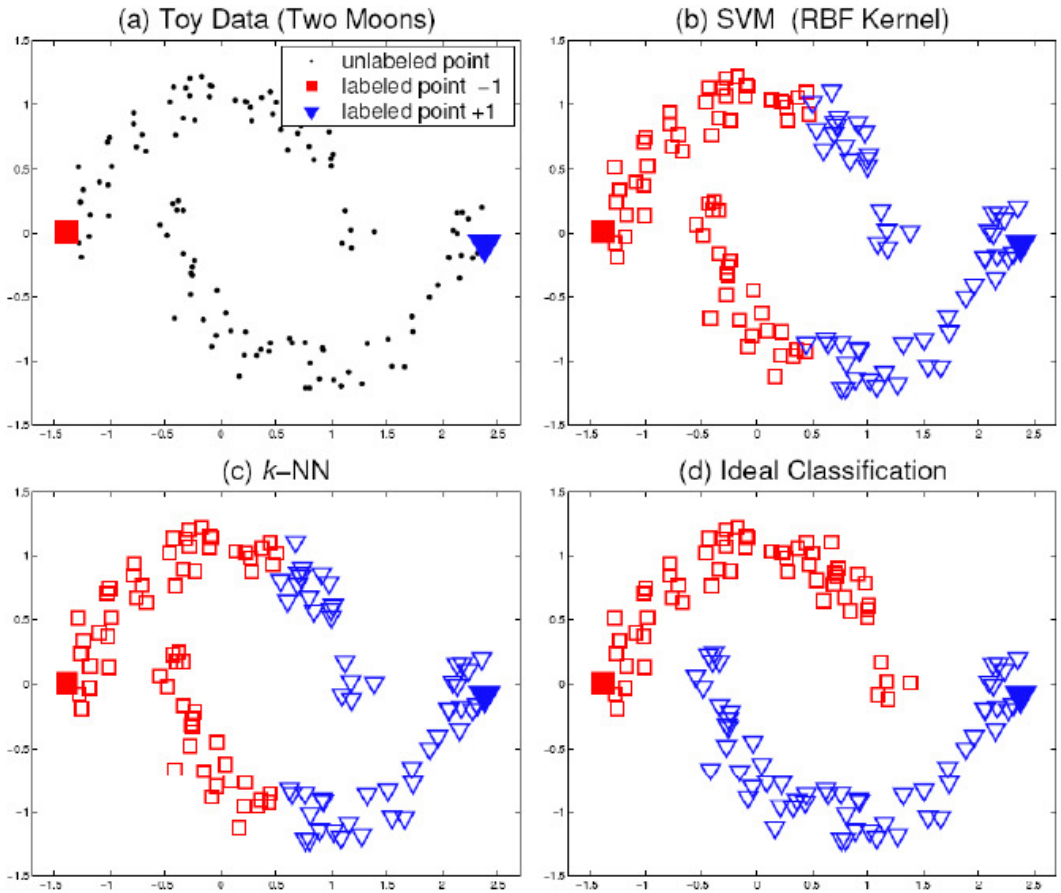
Пример 1. плотности классов, восстановленные:
по размеченным данным X^k по полным данным X^ℓ





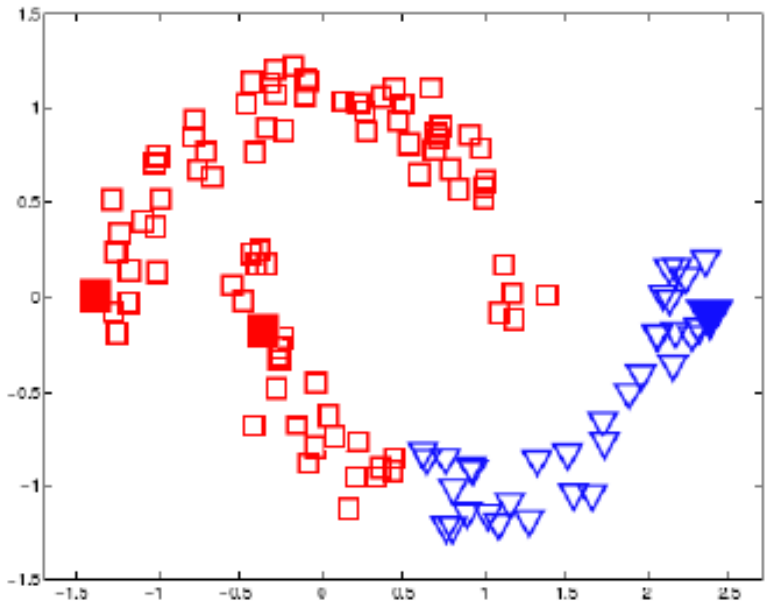
SSL не сводится к классификации

Пример 2. Методы классификации не учитывают кластерную структуру неразмеченных данных



Однако и к кластеризации SSL также не сводится

Пример 3. Методы кластеризации не учитывают приоритетность разметки над кластерной структурой.



Качество кластеризации в метрическом пространстве

- Среднее внутрикластерное расстояние:

$$F_0 = \frac{\sum\limits_{i<j} [a_i = a_j]\, \rho(x_i, x_j)}{\sum\limits_{i<j} [a_i = a_j]} \to \min .$$

- Среднее межкластерное расстояние:

$$F_1 = \frac{\sum\limits_{i<j} [a_i \neq a_j]\, \rho(x_i, x_j)}{\sum\limits_{i<j} [a_i \neq a_j]} \to \max .$$

- Отношение пары функционалов: $F_0 / F_1 \to \min$.

Качество кластеризации в линейном векторном пространстве

Объекты x_i задаются векторами признаков $\big(f_1(x_i), \ldots, f_n(x_i)\big)$.

- Сумма средних внутрикластерных расстояний:

$$\Phi_0 = \sum_{a \in Y} \frac{1}{|X_a|} \sum_{i:\ a_i = a} \rho(x_i, \mu_a) \to \min,$$

$X_a = \{x_i \in X^\ell \mid a_i = a\}$ — кластер a,
μ_a — центр масс кластера a.

- Сумма межкластерных расстояний:

$$\Phi_1 = \sum_{a,b \in Y} \rho(\mu_a, \mu_b) \to \max.$$

- Отношение пары функционалов: $\Phi_0 / \Phi_1 \to \min$.

Кластеризация как задача дискретной оптимизации

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

Задача кластеризации: найти метки кластеров a_i

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

Задача частичного обучения:

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

где λ — ещё один параметр.

Метод K-средних (K-means) для кластеризации

Минимизация суммы квадратов внутрикластерных расстояний:

$$\sum_{i=1}^{\ell} \|x_i - \mu_{a_i}\|^2 \to \min_{\{a_i\},\,\{\mu_a\}}, \quad \|x_i - \mu_a\|^2 = \sum_{j=1}^{n} \big(f_j(x_i) - \mu_{aj}\big)^2$$

Алгоритм Ллойда, упрощённый аналог EM-алгоритма.

Вход: X^ℓ, $K = |Y|$. **Выход:** центры μ_a, $a \in Y$

1: $\mu_a :=$ начальное приближение центров, для всех $a \in Y$;
2: **повторять**
3: **E-шаг:** отнести каждый x_i к ближайшему центру:
$a_i := \arg\min_{a \in Y} \|x_i - \mu_a\|, \quad i = 1, \ldots, \ell$;
4: **M-шаг:** вычислить новые положения центров:
$\mu_{aj} := \dfrac{\sum_{i=1}^{\ell} [a_i = a] x_i}{\sum_{i=1}^{\ell} [a_i = a]}, \quad a \in Y$;
5: **пока** a_i не перестанут изменяться;

Метод K-средних для частичного обучения

Модификация алгоритма Ллойда

при наличии размеченных объектов $\{x_1, \ldots, x_k\}$

Вход: X^ℓ, $K = |Y|$. **Выход:** центры μ_a, $a \in Y$

1: $\mu_a :=$ начальное приближение центров, для всех $a \in Y$;
2: **повторять**
3: **E-шаг:**
отнести каждый $x_i \in U$ к ближайшему центру:
$a_i := \arg\min_{a \in Y} \rho(x_i, \mu_a)$, $i = k+1, \ldots, \ell$;
4: **M-шаг:**
вычислить новые положения центров:
$$\mu_{aj} := \frac{\sum_{i=1}^{\ell} [a_i = a] x_i}{\sum_{i=1}^{\ell} [a_i = a]}, \quad a \in Y;$$
5: **пока** a_i не перестанут изменяться;

Алгоритм КНП для кластеризации

Графовый алгоритм КНП (кратчайший незамкнутый путь)

1: Найти пару вершин $(x_i, x_j) \in X^\ell$ с наименьшим $\rho(x_i, x_j)$ 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;

Ограничения алгоритма:

- необходимость задавать число кластеров K
- высокая чувствительность к шуму

Алгоритм КНП для частичного обучения

Графовый алгоритм КНП (кратчайший незамкнутый путь)

1: Найти пару вершин $(x_i, x_j) \in X^\ell$ с наименьшим $\rho(x_i, x_j)$ 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.

Задача частичного обучения: земеняется только шаг 5...

Алгоритм кластеризации FOREL (ФОРмальные ЭЛементы)

1: $U := X^\ell$ — множество некластеризованных точек;
2: **пока** в выборке есть некластеризованные точки, $U \neq \varnothing$:
3: взять случайную точку $x_0 \in U$;
4: **повторять**
5: образовать кластер с центром в x_0 и радиусом R:
$K_0 := \{x_i \in U \mid \rho(x_i, x_0) \leqslant R\}$;
6: переместить центр x_0 в центр масс кластера:
$x_0 := \frac{1}{|K_0|} \sum_{x_i \in K_0} x_i$;
7: **пока** состав кластера K_0 не стабилизируется;
8: $U := U \setminus K_0$;
9: применить алгоритм КНП к множеству центров кластеров;
10: каждый $x_i \in X^\ell$ приписать кластеру с ближайшим центром;

Ёлкина В.Н., Ёлкин Е.А. Загоруйко Н.Г. О применении методики распознавания образов к решению задач палеонтологии. 1967.

Замечание к шагу 6:

если X не является линейным векторным пространством, то

$$x_0 := \frac{1}{|K_0|} \sum_{x_i \in K_0} x_i \quad \longrightarrow \quad x_0 := \arg\min_{x \in K_0} \sum_{x' \in K_0} \rho(x, x');$$

Преимущества FOREL:

- получаем двухуровневую структуру кластеров;
- кластеры могут быть произвольной формы;
- варьируя R, можно управлять детальностью кластеризации.

Недостаток FOREL:

- чувствительность к R и начальному выбору точки x_0.
  Способ устранения:
  сгенерировать несколько кластеризаций и
  выбрать лучшую по *критерию качества кластеризации*.

Агломеративная иерархическая кластеризация

Алгоритм иерархической кластеризации (Ланс, Уильямс, 1967): итеративный пересчёт расстояний R_{UV} между кластерами U, V.

1: $C_1 := \{\{x_1\}, \ldots, \{x_\ell\}\}$ — все кластеры 1-элементные;
$R_{\{x_i\}\{x_j\}} := \rho(x_i, x_j)$ — расстояния между ними;
2: **для всех** $t = 2, \ldots, \ell$ (t — номер итерации):
3: найти в C_{t-1} пару кластеров (U, V) с минимальным R_{UV};
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:
$R_{WS} := \alpha_U R_{US} + \alpha_V R_{VS} + \beta R_{UV} + \gamma \left| R_{US} - R_{VS} \right|$;

Алгоритм Ланса-Уильямса для частичного обучения

Алгоритм иерархической кластеризации (Ланс, Уильямс, 1967): итеративный пересчёт расстояний R_{UV} между кластерами U, V.

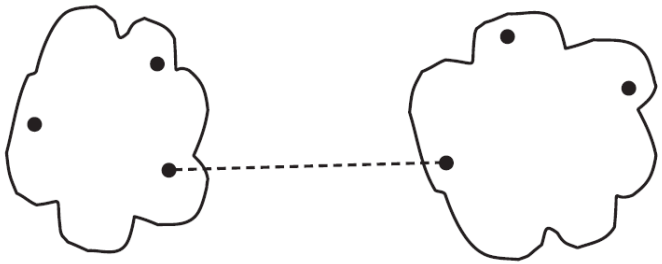
1: $C_1 := \{\{x_1\}, \ldots, \{x_\ell\}\}$ — все кластеры 1-элементные;
$R_{\{x_i\}\{x_j\}} := \rho(x_i, x_j)$ — расстояния между ними;
2: **для всех** $t = 2, \ldots, \ell$ (t — номер итерации):
3: найти в C_{t-1} пару кластеров (U, V) с минимальным R_{UV},
при условии, что в $U \cup V$ нет объектов 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:
$R_{WS} := \alpha_U R_{US} + \alpha_V R_{VS} + \beta R_{UV} + \gamma \left| R_{US} - R_{VS} \right|$;

Частные случаи формулы Ланса-Уильямса

1. Расстояние ближнего соседа:

$$R^{\text{б}}_{WS} = \min_{w\in W, s\in S} \rho(w,s);$$

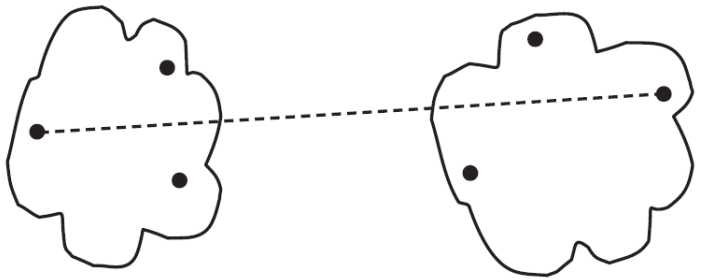
$$\alpha_U = \alpha_V = \tfrac{1}{2},\ \ \beta = 0,\ \ \gamma = -\tfrac{1}{2}.$$



2. Расстояние дальнего соседа:

$$R^{\text{д}}_{WS} = \max_{w\in W, s\in S} \rho(w,s);$$

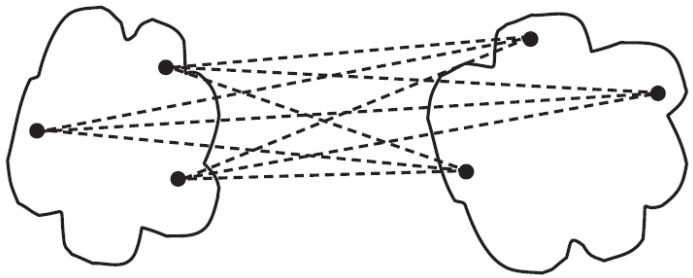
$$\alpha_U = \alpha_V = \tfrac{1}{2},\ \ \beta = 0,\ \ \gamma = \tfrac{1}{2}.$$



3. Групповое среднее расстояние:

$$R^{\text{г}}_{WS} = \frac{1}{|W||S|} \sum_{w\in W} \sum_{s\in S} \rho(w,s);$$

$$\alpha_U = \frac{|U|}{|W|},\ \ \alpha_V = \frac{|V|}{|W|},\ \ \beta = \gamma = 0.$$



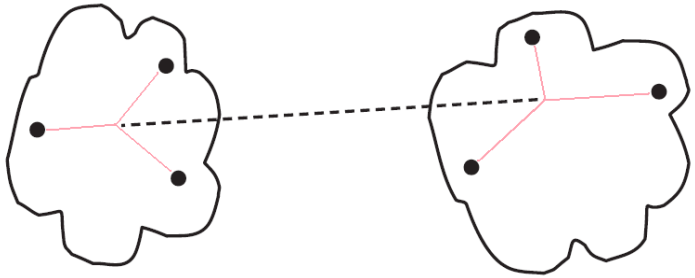
Частные случаи формулы Ланса-Уильямса

4. Расстояние между центрами:

$$R^{\text{ц}}_{WS} = \rho^2\Big(\sum_{w\in W}\frac{w}{|W|}, \sum_{s\in S}\frac{s}{|S|}\Big);$$

$$\alpha_U = \frac{|U|}{|W|}, \quad \alpha_V = \frac{|V|}{|W|},$$

$$\beta = -\alpha_U\alpha_V, \quad \gamma = 0.$$



5. Расстояние Уорда:

$$R^{\text{у}}_{WS} = \frac{|S||W|}{|S|+|W|}\,\rho^2\Big(\sum_{w\in W}\frac{w}{|W|}, \sum_{s\in S}\frac{s}{|S|}\Big);$$

$$\alpha_U = \frac{|S|+|U|}{|S|+|W|}, \quad \alpha_V = \frac{|S|+|V|}{|S|+|W|}, \quad \beta = \frac{-|S|}{|S|+|W|}, \quad \gamma = 0.$$

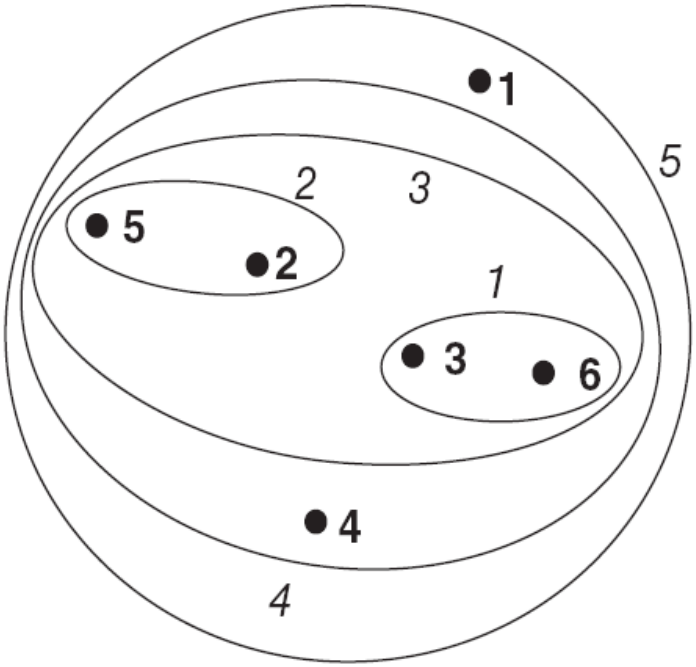
Проблема выбора

Какая функция расстояния лучше?

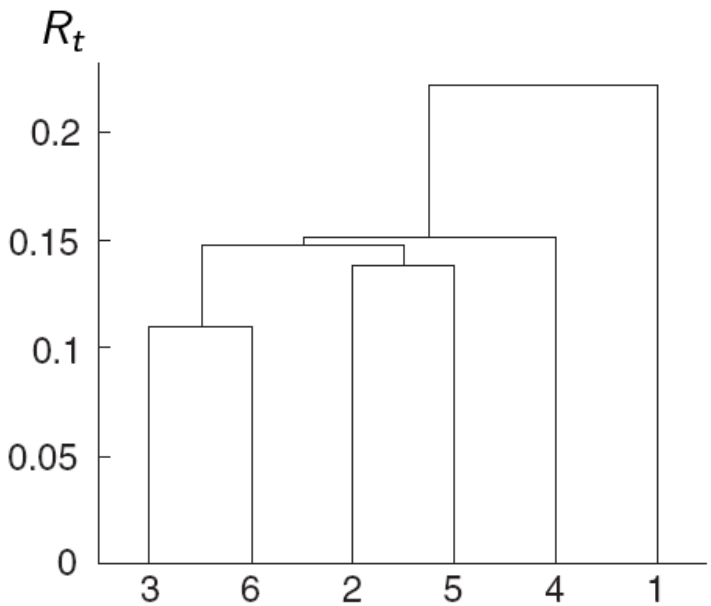
Визуализация кластерной структуры

1. Расстояние ближнего соседа:

Диаграмма вложения



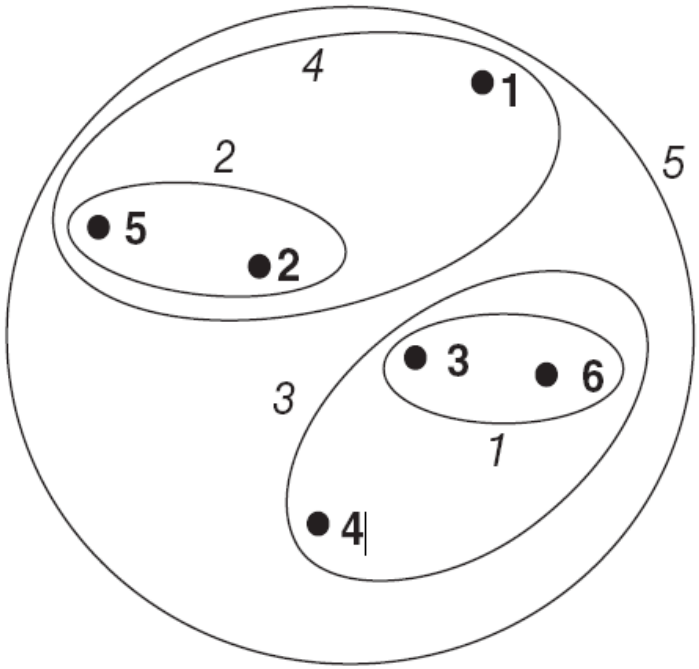
Дендрограмма



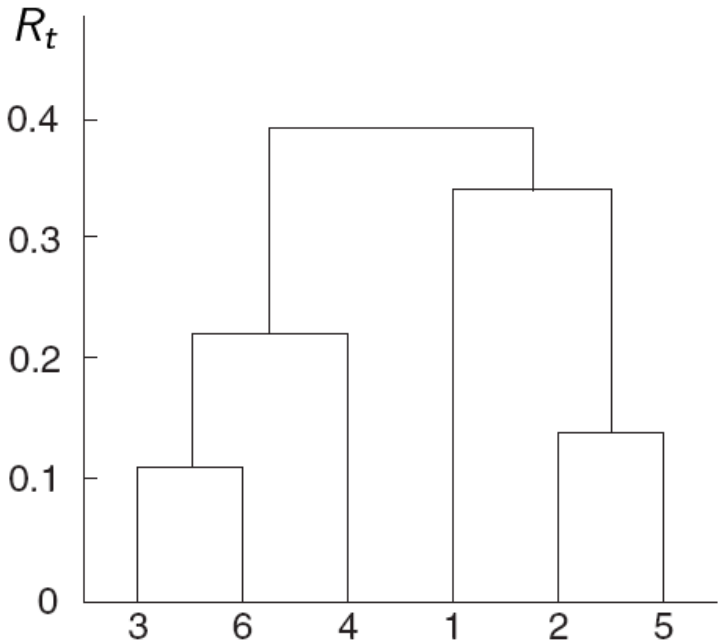
Визуализация кластерной структуры

2. Расстояние дальнего соседа:

Диаграмма вложения



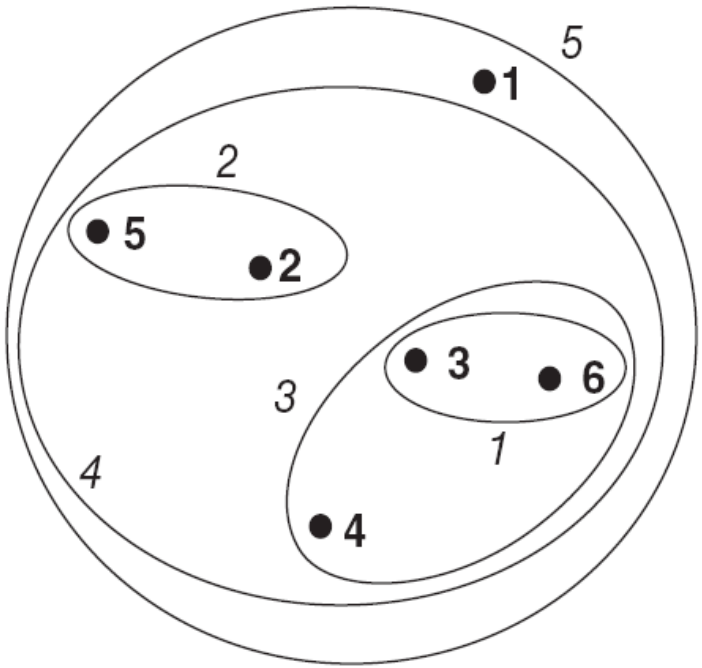
Дендрограмма



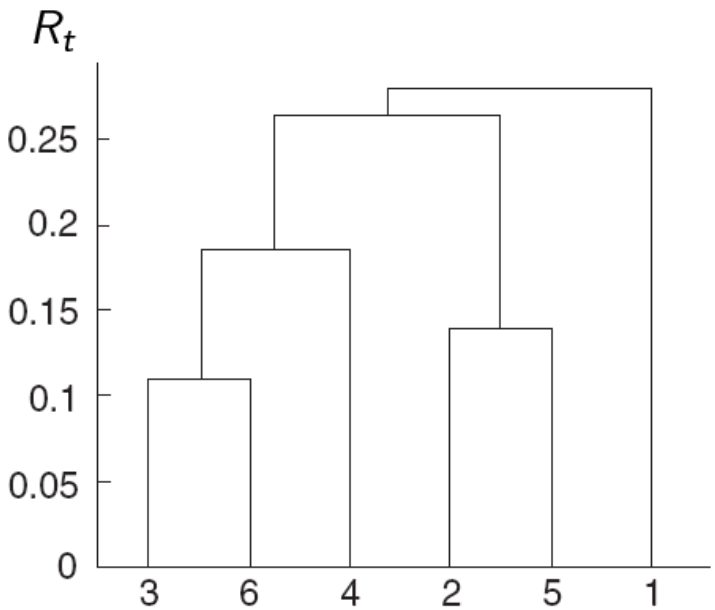
Визуализация кластерной структуры

3. Групповое среднее расстояние:

Диаграмма вложения



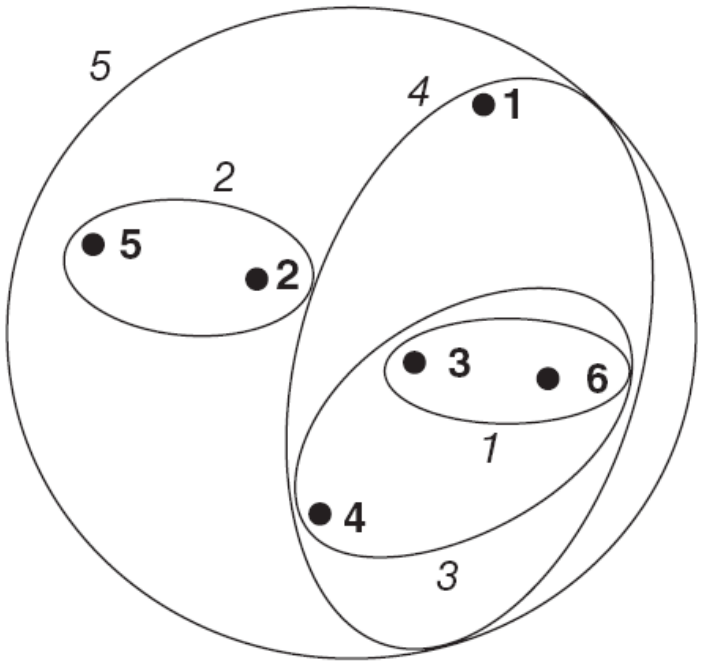
Дендрограмма



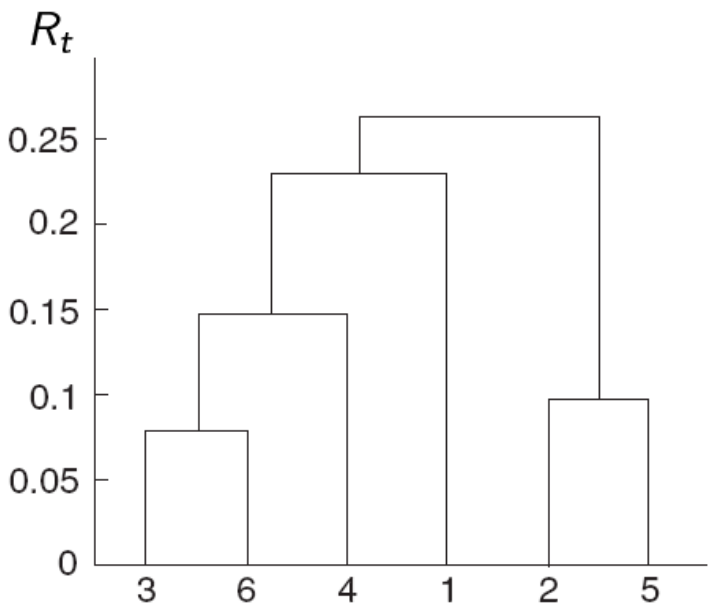
Визуализация кластерной структуры

5. Расстояние Уорда:

Диаграмма вложения



Дендрограмма



Основные свойства иерархической кластеризации

- *Монотонность*: дендрограмма не имеет самопересечений, при каждом слиянии расстояние между объединяемыми кластерами только увеличивается: $R_2 \leqslant R_3 \leqslant \ldots \leqslant R_\ell$.

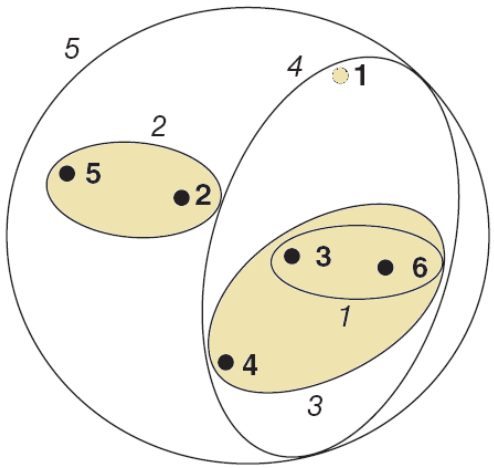
  Теорема (Миллиган, 1979)
  Достаточное условие монотонности:
  $\alpha_U \geqslant 0,\ \ \alpha_V \geqslant 0,\ \ \alpha_U + \alpha_V + \beta \geqslant 1,\ \ \min\{\alpha_U, \alpha_V\} + \gamma \geqslant 0.$
- *Сжимающее расстояние*: $R_t \leqslant \rho(\mu_U, \mu_V),\ \ \forall t.$
- *Растягивающее расстояние*: $R_t \geqslant \rho(\mu_U, \mu_V),\ \ \forall t$

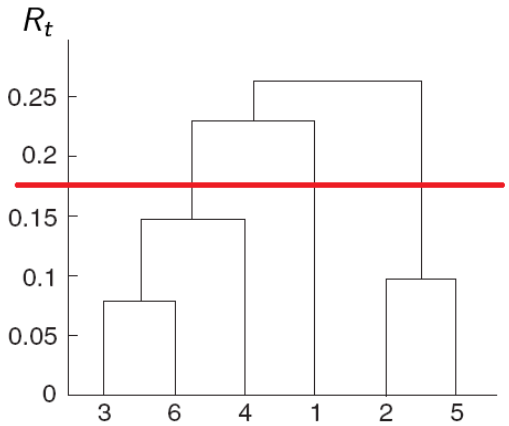
$R^{\text{ц}}$ не монотонно; $R^{\text{б}}$, $R^{\text{д}}$, $R^{\text{г}}$, $R^{\text{у}}$ — монотонны.

$R^{\text{б}}$ — сжимающее; $R^{\text{д}}$, $R^{\text{у}}$ — растягивающие;

Рекомендации и выводы

- рекомендуется пользоваться расстоянием Уорда $R^{\text{у}}$;
- обычно строят несколько вариантов и выбирают лучший визуально по дендрограмме;
- определение числа кластеров — по максимуму $|R_{t+1} - R_t|$, тогда результирующее множество кластеров $:= C_t$.





Метод self-training (1965-1970)

Пусть $\mu\colon X^k \to a$ — метод обучения классификации;
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: $Z := X^k$;
2: **пока** $|Z| < \ell$
3: $\quad a := \mu(Z)$;
4: $\quad \Delta := \left\{ x_i \in U \setminus Z \;\middle|\; M_i(a) \geqslant M_0 \right\}$;
5: $\quad a_i := a(x_i)$ для всех $x_i \in \Delta$;
6: $\quad Z := Z \cup \Delta$;

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

Метод co-training (Blum, Mitchell, 1998)

Пусть $\mu_1 \colon X^k \to a_1$, $\mu_2 \colon X^k \to a_2$ — два существенно различных метода обучения, использующих
— либо разные наборы признаков;
— либо разные парадигмы обучения (inductive bias);
— либо разные источники данных $X_1^{k_1}$, $X_2^{k_2}$.

$Z_1 := X_1^{k_1}$; $Z_2 := X_2^{k_2}$;
пока $|Z_1 \cup Z_2| < \ell$
  $a_1 := \mu_1(Z_1)$; $\Delta_1 := \left\{x_i \in U \backslash Z_1 \backslash Z_2 \;\middle|\; M_i(a_1) \geqslant M_{01}\right\}$;
  $a_i := a_1(x_i)$ для всех $x_i \in \Delta_1$;
  $Z_2 := Z_2 \cup \Delta_1$;
  $a_2 := \mu_2(Z_2)$; $\Delta_2 := \left\{x_i \in U \backslash Z_1 \backslash Z_2 \;\middle|\; M_i(a_2) \geqslant M_{02}\right\}$;
  $a_i := a_2(x_i)$ для всех $x_i \in \Delta_2$;
  $Z_1 := Z_1 \cup \Delta_2$;

Метод co-learning (deSa, 1993)

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

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

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

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

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

SVM: классификация

Линейный классификатор на два класса $Y = \{-1, 1\}$:

$$a(x) = \text{sign}\big(\langle w, x\rangle - w_0\big), \quad w, x \in \mathbb{R}^n,\ w_0 \in \mathbb{R}.$$

Отступ объекта x_i:

$$M_i(w, w_0) = \big(\langle w, x_i\rangle - w_0\big) y_i.$$

Задача обучения весов w, w_0 по размеченной выборке:

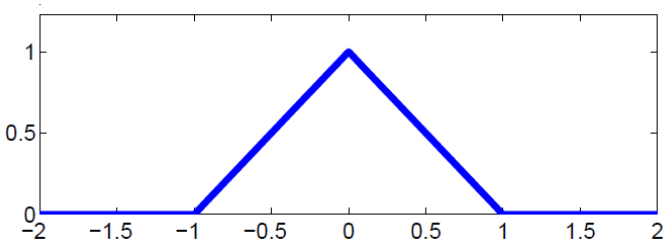
$$Q(w, w_0) = \sum_{i=1}^{k} \big(1 - M_i(w, w_0)\big)_+ + \frac{1}{2C}\|w\|^2 \ \to \ \min_{w, w_0}.$$

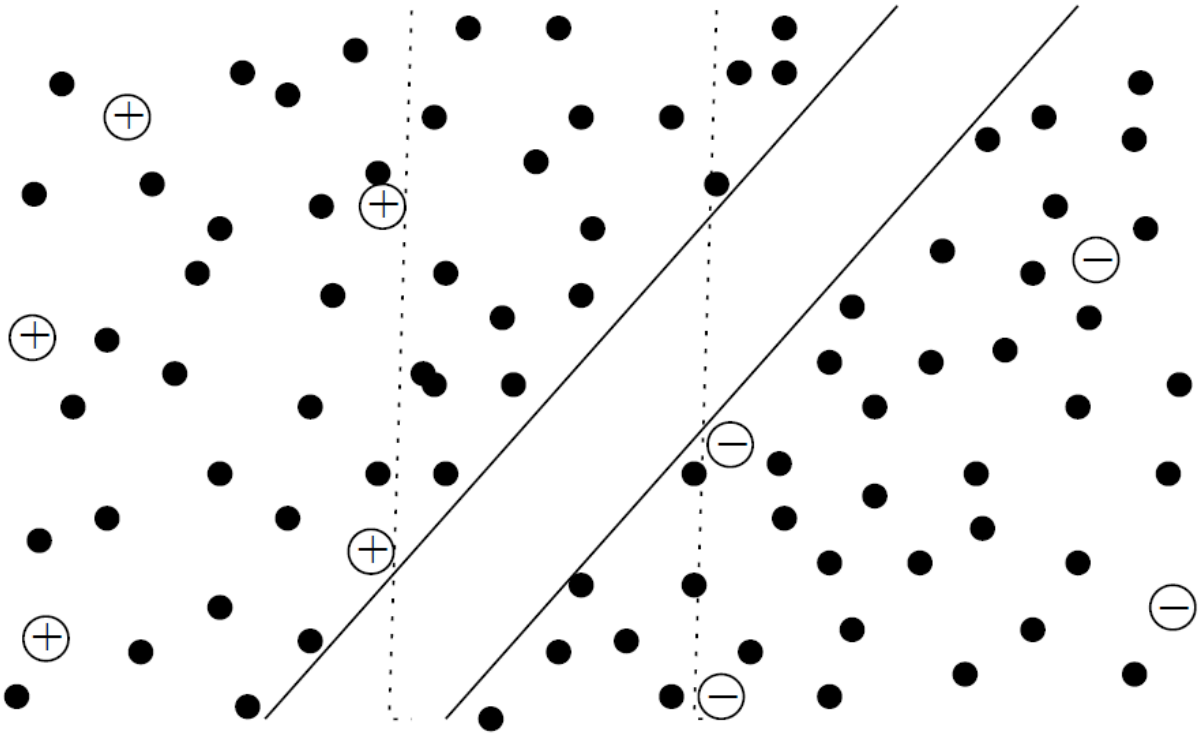
Функция $\mathscr{L}(M) = (1 - M)_+$ штрафует за уменьшение отступа.

Идея!
Функция $\mathscr{L}(M) = \big(1 - |M|\big)_+$ штрафует за попадание объекта внутрь разделяющей полосы.

Функция потерь для трансдуктивного SVM

Функция потерь $\mathscr{L}(M) = \big(1 - |M|\big)_+$ штрафует за попадание объекта внутрь разделяющей полосы.





Transductive SVM: частичное обучение

Обучение весов w, w_0 по частично размеченной выборке:

$$Q(w, w_0) = \sum_{i=1}^{k} \big(1 - M_i(w, w_0)\big)_+ + \frac{1}{2C}\|w\|^2 +$$

$$+ \gamma \sum_{i=k+1}^{\ell} \big(1 - |M_i(w, w_0)|\big)_+ \;\to\; \min_{w, w_0}.$$

Достоинства и недостатки TSVM:

- ⊕ как и в обычном SVM, можно использовать ядра;
- ⊕ имеются эффективные реализации для больших данных;
- ⊖ задача невыпуклая, методы оптимизации сложнее;
- ⊖ решение неустойчиво, если нет области разреженности;
- ⊖ требуется настройка двух параметров C, γ;

Sindhwani, Keerthi. Large scale semisupervised linear SVMs. SIGIR 2006.

Многоклассовая логистическая регрессия

Линейный классификатор по конечному множеству классов $|Y|$:

$$a(x) = \arg\max_{y \in Y} \langle w_y, x \rangle, \quad x, w_y \in \mathbb{R}^n.$$

Вероятность того, что объект x_i относится к классу y:

$$P(y|x_i, w) = \frac{\exp\langle w_y, x_i \rangle}{\sum_{c \in Y} \exp\langle w_c, x_i \rangle}.$$

Задача максимизации регуляризованного правдоподобия:

$$Q(w) = \sum_{i=1}^{k} \log P(y_i|x_i, w) - \frac{1}{2C} \sum_{y \in Y} \|w_y\|^2 \;\to\; \max_w,$$

Оптимизация $Q(w)$ — методом стохастического градиента.

Логистическая регрессия с частичным обучением

Теперь учтём неразмеченные данные $U = \{x_{k+1}, \ldots, x_\ell\}$.
Пусть $b_j(x)$ — бинарные признаки, $j = 1, \ldots, m$.

Оценим вероятности $\mathrm{P}(y|b_j(x) = 1)$ двумя способами:

1) эмпирическая оценка по размеченным данным X^k:

$$\hat{p}_j(y) = \frac{\sum_{i=1}^{k} b_j(x_i)[y_i = y]}{\sum_{i=1}^{k} b_j(x_i)};$$

2) оценка по неразмеченным данным U и линейной модели:

$$p_j(y|w) = \frac{\sum_{i=k+1}^{\ell} b_j(x_i) P(y|x_i, w)}{\sum_{i=k+1}^{\ell} b_j(x_i)}.$$

Будем минимизировать дивергенцию Кульбака–Лейблера между распределениями $\hat{p}_j(y)$ и $p_j(y|w)$.

Построение функционала качества

Минимизация KL-дивергенции между $\hat{p}_j(y)$ и $p_j(y|w)$:

$$\mathrm{KL}\big(\hat{p}_j(y) \,\|\, p_j(y|w)\big) = \sum_y \hat{p}_j(y) \log \frac{\hat{p}_j(y)}{p_j(y|w)} \;\to\; \min_w .$$

Вычтем сумму KL-дивергенций по всем признакам $j = 1, \ldots, m$ из функционала регуляризованного правдоподобия $Q(w)$, с коэффициентом регуляризации γ:

$$\tilde{Q}(w) = \sum_{i=1}^{k} \log P(y_i|x_i, w) - \frac{1}{2C} \sum_{y \in Y} \|w_y\|^2 +$$
$$+ \gamma \sum_{j=1}^{m} \sum_{y \in Y} \hat{p}_j(y) \log \left(\frac{\sum_{i=k+1}^{\ell} b_j(x_i) P(y|x_i, w)}{\sum_{i=k+1}^{\ell} b_j(x_i)} \right) \;\to\; \max_w .$$

Mann, McCallum. Simple, robust, scalable semi-supervised learning via expectation regularization. ICML 2007.

Особенности регуляризации для частичного обучения

1. Оптимизация $\tilde{Q}(w)$ — методом стохастического градиента.
2. Возможные варианты задания переменных b_j:
   - $b_j(x) \equiv 1$, тогда $\mathrm{P}(y|b_j(x)=1)$ — априорная вероятность класса y (label regularization) — хорошо подходит для задач с несбалансированными классами;
   - $b_j(x) = \left[\text{термин } j \text{ содержится в тексте } x\right]$ — для задач классификации и каталогизации текстов.
3. метод слабо чувствителен к выбору C и γ,
4. устойчив к погрешностям оценивания $\hat{p}_j(y)$,
5. не требует большого числа размеченных объектов k,
6. хорошо подходит для категоризации текстов.

Mann, McCallum. Simple, robust, scalable semi-supervised learning via expectation regularization. ICML 2007.

Резюме в конце лекции

- Кластеризация — это обучение без учителя, некорректно поставленная задача, существует много критериев и эвристических алгоритмов кластеризации
- Задача SSL занимает промежуточное положение между классификацией и кластеризацией, но не сводится к ним.
- Методы кластеризации легко адаптируются к SSL путём введения ограничений (constrained clustering).
- Адаптация методов классификации реализуется сложнее, но приводит к более эффективным методам.
- Регуляризация позволяет учитывать дополнительную информацию в постановке оптимизационной задачи.