# 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.

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

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

Отступ объекта $x_i$:

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

Задача обучения весов $w, w_0$ по размеченной выборке:

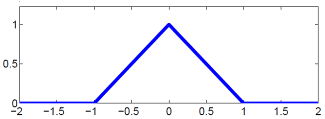
$$Q(w, w_0) = \sum_{i=1}^{\ell} \bigl(1 - M_i(w, w_0)\bigr)_+ + \frac{1}{2C}\|w\|^2 \;\to\; \min_{w, w_0}.$$

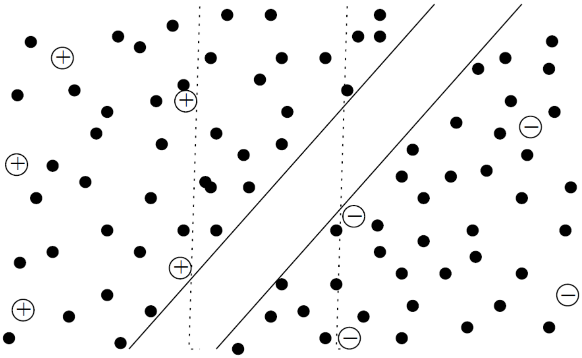
Функция $\mathscr{L}(M) = (1 - M)_+$ штрафует за уменьшение отступа.

**Идея!**

Функция $\mathscr{L}(M) = \bigl(1 - |M|\bigr)_+$ штрафует за попадание в зазор,
$\bigl|M_i\bigr|$ не зависит от $y_i$ и определён на неразмеченных объектах.

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



Обучение весов $w, w_0$ по частично размеченной выборке:

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

**Достоинства и недостатки TSVM:**

- ⊕ как и в обычном SVM, можно использовать ядра;
- ⊕ имеются эффективные реализации для больших данных;
- ⊖ решение неустойчиво, если нет области разреженности;
- ⊖ требуется настройка двух параметров $C$, $\gamma$;

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

# Многоклассовая логистическая регрессия

Линейный классификатор на конечное множество 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 $y$:

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

Задача максимизации регуляризованного правдоподобия:

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

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

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

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

1) эмпирическая оценка по размеченным данным $X^\ell$:

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

2) оценка по неразмеченным данным $X^k$ и линейной модели:

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

Будем минимизировать расстояние между $\hat{p}_j(y)$ и $p_j(y, w)$, в качестве расстояния между распределениями возьмём дивергенцию Кульбака–Лейблера.

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

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

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

$$\tilde{Q}(w) = \sum_{i=1}^{\ell} \log P(y_i | x_i, w) - \frac{1}{2C} \sum_{y \in Y} \|w_y\|^2 +$$

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

где $\gamma$ — коэффициент регуляризации.

---

*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 $\mathsf{P}(y|b_j(x) = 1)$ — априорная вероятность класса $y$ (label regularization) — хорошо подходит для задач с несбалансированными классами;
   - $b_j(x) = \bigl[$термин $j$ содержится в тексте $x\bigr]$ — для задач классификации и каталогизации текстов.
3. метод слабо чувствителен к выбору $C$ и $\gamma$,
4. устойчив к погрешностям оценивания $\hat{p}_j(y)$,
5. не требует большого числа размеченных объектов $\ell$,
6. хорошо подходит для категоризации текстов,
7. в экспериментах показывает высокую точность.

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

# 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.