Структурное обучение и S-SVM

Петр Ромов

14 декабря 2011

Суть

"Normal" Machine Learning

$$f : \mathcal{X} \to \mathbb{R}$$

- вход $\mathcal{X}$ — объекты произвольной природы
- выход y — вещественное число

Structured Output Learning

$$f : \mathcal{X} \to \mathcal{Y}$$

- вход $\mathcal{X}$ — объекты произвольной природы
- выход $y \in \mathcal{Y}$ — сложный (имеющий внутреннюю структуру) объект

Outline

Структурное предсказание

$$f : \mathcal{X} \to \mathcal{Y}$$

$$f(x) := \arg\max_{y \in \mathcal{Y}} g(x, y)$$

$g(x, y)$ — функция совместности (compatibility function)

Структурное предсказание

$$f : \mathcal{X} \to \mathcal{Y}$$

$$f(x) := \arg\max_{y \in \mathcal{Y}} g(x, y)$$

- Линейное решающее правило
  $\mathcal{Y} = \{-1, +1\}, \quad g(x, y) = y \cdot w^\top \varphi(x)$
- Много классов (one-vs-all)
  $\mathcal{Y} = \{1, \ldots, K\}, \quad g(x, y) = \sum_{i=1}^{K} [y = i] \cdot w_i^\top \varphi(x)$
- Много меток
  $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_n, \quad \mathcal{Y}_i = \{1, \ldots, K\}$

Структурные выходы

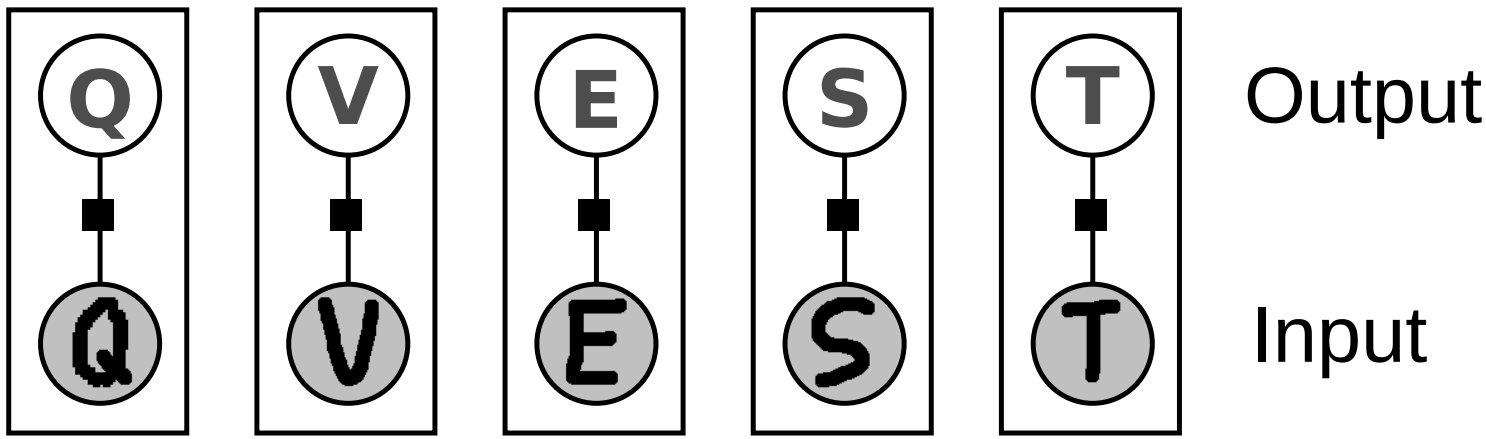
В качестве “сложных” объектов будем рассматривать:

$$\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_n,$$

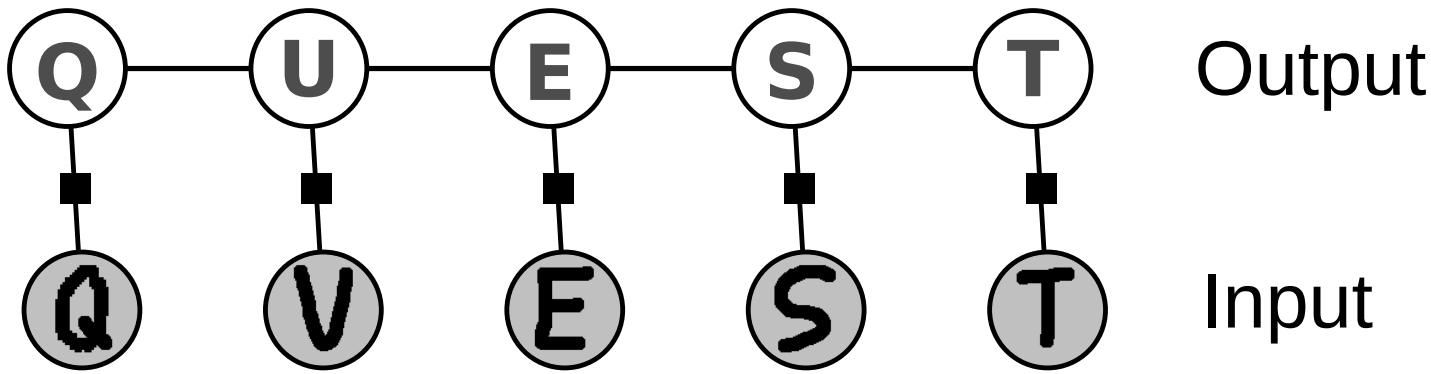
$$y \in \mathcal{Y}, \quad y = (y_1, \ldots, y_n).$$

- y_i имеет конечное множество значений $\mathcal{Y}_i$;
- существуют зависимости между переменными y_i;
- во многих задачах зрения $\mathcal{Y}_i = \mathcal{L}$,
  y_i называют метками;

Пример структурного выхода



Пример структурного выхода

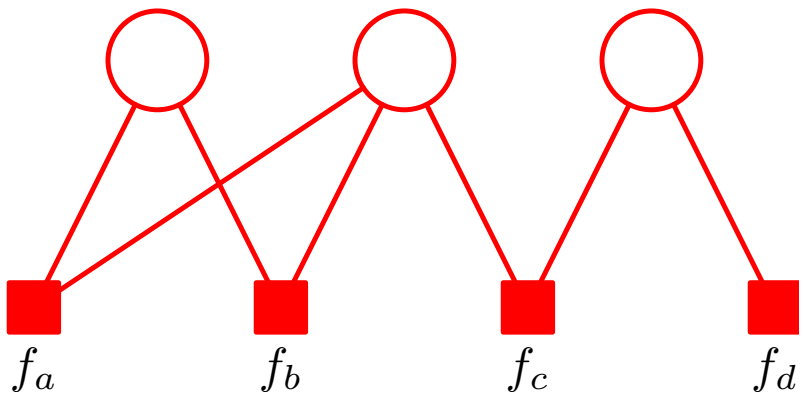


Описание структуры: CRF

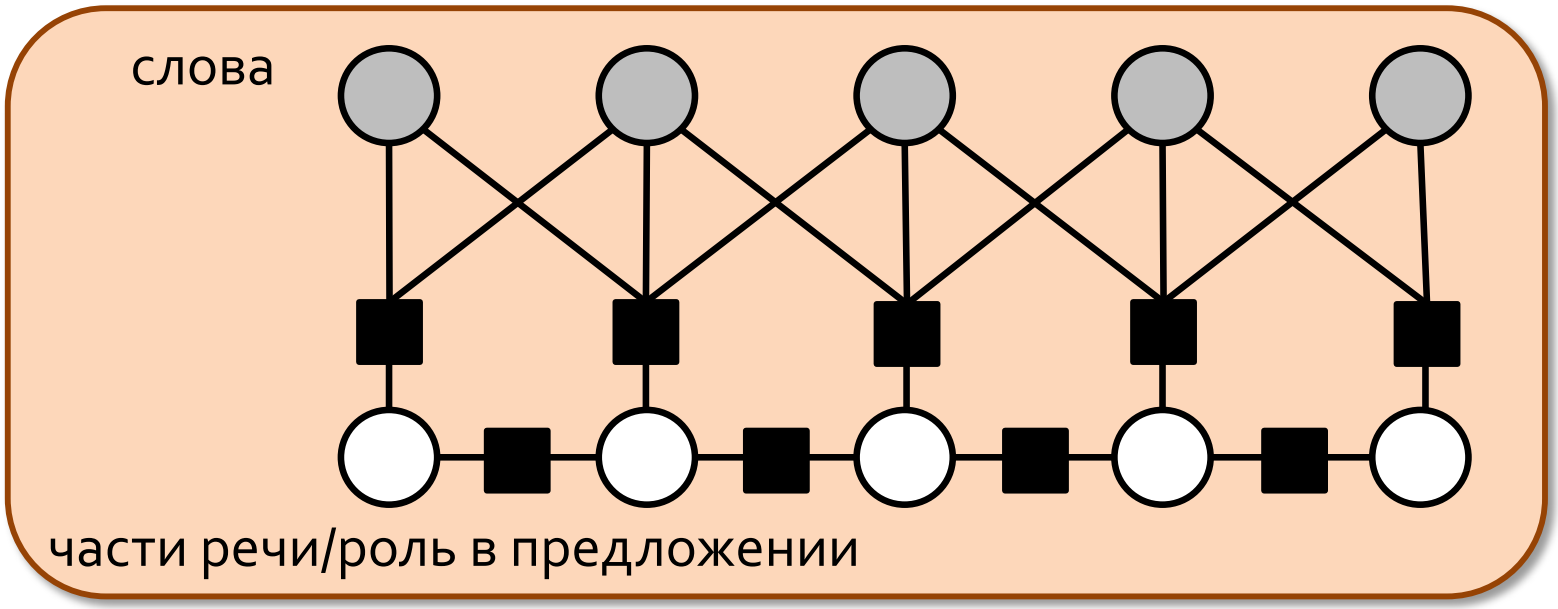
Conditional Random Field

$$p(y|x) = \frac{1}{Z(x)} \prod_{f \in \mathcal{F}} \psi_f(y_{N(f)}; x) = \frac{1}{Z(x)} \exp\{-E(y|x)\}$$

$$E(y|x) = \sum_{f \in \mathcal{F}} E_f(y_{N(f)}|x)$$

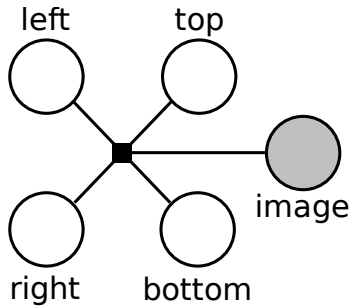


Пример CRF для описания структуры



Пример CRF для описания структуры

$(x, y) \hat{=} ($  ,  $)$

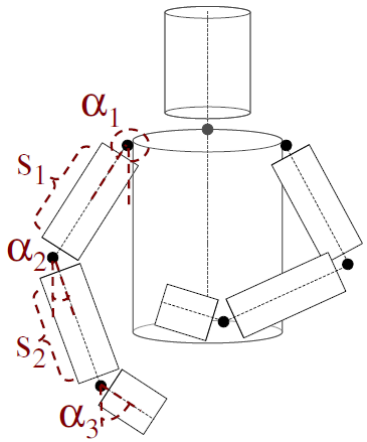


$$E(x, y) = w^\top \varphi(x, y)$$

$\varphi(x, y)$ — гистограмма визуальных слов в окне y.

Пример CRF для описания структуры

input: image



body model

output: model fit

$$E(x,y) = \sum_i w_i^\top \varphi_{fit}(x_i, y_i) + \sum_{ij} w_{ij}^\top \varphi_{pose}(y_i, y_j)$$

Вывод в CRF

Пусть $\Delta(y, \hat{y})$ — функция потерь.

$$\hat{y}(x) = \arg\min_{\hat{y}} \mathbb{E}_{y \sim p(y|x)} \Delta(y, \hat{y})$$

- Maximum aposteriori (MAP)

$$\Delta(y, \hat{y}) = [y \neq \hat{y}] \qquad g(x, y) = p(y|x)$$

$$\hat{y}(x) = \arg\max_{\hat{y}} p(y|x)$$

- Max-marginals

$$\Delta(y, \hat{y}) = \sum_i [y_i \neq \hat{y}_i] \qquad g(x, y) = \sum_i p(y_i|x)$$

$$\hat{y}_i(x) = \arg\max_{\hat{y}_i} p(y_i|x)$$

Вывод в CRF

Пусть $\Delta(y, \hat{y})$ — функция потерь.

$$\hat{y}(x) = \arg\min_{\hat{y}} \mathbb{E}_{y \sim p(y|x)} \Delta(y, \hat{y})$$

- **Maximum aposteriori (MAP)**

$$\Delta(y, \hat{y}) = [y \neq \hat{y}] \qquad g(x, y) = p(y|x)$$
$$\hat{y}(x) = \arg\max_{\hat{y}} p(y|x)$$

- Max-marginals

$$\Delta(y, \hat{y}) = \sum_i [y_i \neq \hat{y}_i] \qquad g(x, y) = \sum_i p(y_i|x)$$
$$\hat{y}_i(x) = \arg\max_{\hat{y}_i} p(y_i|x)$$

Вывод в CRF

Пусть $\Delta(y, \hat{y})$ — функция потерь.

$$\hat{y}(x) = \arg\min_{\hat{y}} \mathbb{E}_{y \sim p(y|x)} \Delta(y, \hat{y})$$

- **Maximum aposteriori (MAP)**

$$\Delta(y, \hat{y}) = [y \neq \hat{y}] \qquad g(x, y) = p(y|x)$$
$$\hat{y}(x) = \arg\max_{\hat{y}} p(y|x)$$

- **Max-marginals**

$$\Delta(y, \hat{y}) = \sum_i [y_i \neq \hat{y}_i] \qquad g(x, y) = \sum_i p(y_i|x)$$
$$\hat{y}_i(x) = \arg\max_{\hat{y}_i} p(y_i|x)$$

Параметризация CRF

Энергия линейна по параметру w.

$$E_f(y_f; x, w) = \langle w(y_f), \varphi_f(x_f) \rangle$$

$$\begin{aligned} E(x, y, w) &= w^\top \varphi(x, y) \\ &= \sum_{i=1}^{n} w_i^\top \varphi_i(x, y) \quad \textbf{unary terms} \\ &\quad + \sum_{i,j=1}^{n} w_{ij}^\top \varphi_{ij}(x, y) \quad \textbf{pairwise terms} \\ &\quad + \ldots \quad \textbf{higher-order terms} \end{aligned}$$

Обучение CRF

Имеется:

- модель $p(y|x,w) = \frac{1}{Z(x,w)} \exp\{-E(x,y,w)\}$
- набор данных $(x^1, y^1), \ldots, (x^N, y^N)$

Требуется: настроить параметр w.

Подходы

- Вероятностное обучение:
  - регуляризованный максимум условного правдоподобия $\mathcal{R}(w) + \log p(y|x,w) \to \max_w$
- Максимизация зазора (Структурный SVM)

Линейная модель функции совместности

$$g(x, y, w) = w^\top \psi(x, y)$$

- Линейное решающее правило
  $\mathcal{Y} = \{-1, +1\}$,
  $g(x, y, w) = w^\top (y \cdot \varphi(x))$
- Много меток
  $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_n$,
  $g(x, y, w) = -E(y|x) = w^\top (-\varphi(x, y))$

Outline

Структурное обучение

Задача (Loss-minimizing parameter learning)

Реальное распределение данных $d(x, y)$ *неизвестно.*
С точностью до параметра w *задана модель предиктора*

$$f_w(x) = \arg\max_y g(x, y, w).$$

Дано:

- $\mathcal{D} = \{(x^1, y^1), \ldots, (x^n, y^n)\} \subset \mathcal{X} \times \mathcal{Y}$ — *однородная независимая выборка,*
- $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ — *функция потерь.*

Требуется найти: w^*, *т.ч. Байесовский риск*

$$\mathbb{E}_{(x,y)\sim d(x,y)} \Delta(y, f_{w^*}(x))$$

насколько возможно мал.

Максимизация зазора

Обучение, методом максимизации зазора[1]:

$$\begin{cases} \gamma \to \max\limits_{w,\gamma} \\ \|w\| \leq 1 \\ g(x^n, y^n, w) - g(x^n, y, w) \geq \gamma \\ \qquad \forall n = 1, \ldots, N,\ \forall y \in \mathcal{Y} : y \neq y^n \end{cases}$$

γ — величина зазора
$g(x^n, y^n, w) - g(x^n, y, w)$ — отступ между разметками

[1] слишком много значков, пора рисовать картиночки на доске

Максимизация зазора

Обучение, методом максимизации зазора[1]:

$$\begin{cases} \gamma \to \max\limits_{w,\gamma} \\ \|w\| \le 1 \\ \boxed{g(x^n, y^n, w) - g(x^n, y, w)} \ge \gamma \\ \qquad \forall n = 1, \ldots, N,\ \forall y \in \mathcal{Y} : y \ne y^n \end{cases}$$

γ — величина зазора

$\boxed{g(x^n, y^n, w) - g(x^n, y, w)}$ — отступ между разметками

Физический смысл максимизации γ: помешать тому, чтобы значения функции совместности g были одинаково большими для правильного (y^n) и неправильного $(y \ne y^n)$ ответов.

[1] слишком много значков, пора рисовать картиночки на доске

Максимизация зазора

Обучение, методом максимизации зазора[1]:

$$\begin{cases} \gamma \to \max\limits_{w,\gamma} \\ \|w\| \le 1 \\ g(x^n, y^n, w) - g(x^n, y, w) \ge \gamma \\ \qquad \forall n = 1, \ldots, N,\ \forall y \in \mathcal{Y} : y \ne y^n \end{cases}$$

Линейный SVM:

$$g(x, y, w) = y \cdot w^\top x$$

$$\begin{cases} \gamma \to \max_{w,\gamma} \\ \|w\| = 1 \\ y^n \cdot w^\top x^n \ge \frac{\gamma}{2}, \quad n = 1, \ldots, N \end{cases}$$

---

[1] слишком много значков, пора рисовать картиночки на доске

Максимизация зазора: функция потерь

Обобщим понятие зазора на случай произвольной функции потерь. Пусть задана $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, $\Delta(y, y) = 0$:

$$\begin{cases} \gamma \to \max\limits_{w,\gamma} \\ \|w\| \leq 1 \\ g(x^n, y^n, w) - g(x^n, y, w) \geq \gamma \Delta(y^n, y) \\ \qquad \forall n = 1, \ldots, N, \ \forall y \in \mathcal{Y} \end{cases}$$

Чем больше потери от неправильной классификации — тем больший отступ требуем.

Максимизация зазора: QP

В случае линейной модели $g(x, y, w) = w^\top \psi(x, y)$, метод можно сформулировать в виде задачи квадратичного программирования:

$$\begin{cases} \|w\|^2 \to \min\limits_w \\ w^\top(\psi(x^n, y^n) - \psi(x^n, y)) \geq \Delta(y, y^n) \\ \qquad \forall n = 1, \ldots, N,\ \forall y \in \mathcal{Y} \end{cases}$$

Максимизация зазора: переменные невязки

Не все выборки (в данном случае объектов из $\mathcal{X} \times \mathcal{Y}$) линейно разделимы, введем *переменные невязки* (slack variables):

$$\begin{cases} \|w\|^2 + C \cdot \dfrac{1}{N} \displaystyle\sum_{n=1}^{N} \xi_n \to \min\limits_{w, \xi \geq 0} \\ w^\top (\psi(x^n, y^n) - \psi(x^n, y)) \geq \Delta(y, y^n) - \xi_n \\ \qquad \forall n = 1, \ldots, N, \ \forall y \in \mathcal{Y} \end{cases}$$

Физический смысл ξ_n: за определенную плату, разрешаем функции совместности $g(x, y, w)$ отличать разные разметки *“не так сильно”*, как того требует зазор.

Обобщение hinge-loss

$$l_{\text{hinge}}(x^n, y^n, w) = \max_y \left\{ \Delta(y^n, y) + w^\top (\psi(x^n, y) - \psi(x^n, y^n)) \right\}$$

- выпуклая по w
- кусочно линейная по w
- является верхней гранью для функции потерь:
  $\Delta(y^n, f_w(x^n)) \leq l_{\text{hinge}}(x^n, y^n, w) \quad \forall w$

Минимизация структурного риска

$$\mathcal{R}(f_w) + C \cdot \underbrace{\frac{1}{N} \sum_{n=1}^{N} \Delta(y^n, f_w(x^n))}_{\approx \mathbb{E}_{d(x,y)} \Delta(y, f(x))} \to \min_w$$

- Байесовский риск приближается ошибкой на обучении (эмпирический риск);
- регуляризатор $\mathcal{R}(f)$ — емкость модели.

Структурный метод опорных векторов

Кусочно-постоянная (по w) функция потерь $\Delta(y^n, f_w(x^n))$ заменяется выпуклой верхней гранью:

$$\frac{1}{2}\|w\|^2 + C \cdot \frac{1}{N}\sum_{n=1}^{N} l_{\text{hinge}}(x^n, y^n, w) \to \min_w$$

$$l_{\text{hinge}} = \max_y \{\Delta(y^n, y) + g(x^n, y, w) - g(x^n, y^n, w)\}$$

Двойственная формулировка

$$\sum_{\substack{y\in\mathcal{Y}\\ n=1,\dots,N}} \alpha_{ny}\Delta(y^n,y) - \frac{1}{2}\sum_{\substack{y\in\mathcal{Y}\\ n=1,\dots,N}}\sum_{\substack{y'\in\mathcal{Y}\\ n'=1,\dots,N}} \alpha_{ny}\alpha_{n'y'}\bar{K}^{nn'}_{yy'} \to \max_{\alpha\geq 0}$$

$$\text{s.t.:}\quad \sum_{y\in\mathcal{Y}}\alpha_{ny} \leq \frac{C}{N},\quad n=1,\dots,N$$

$$\bar{K}^{nn'}_{yy'} = K^{nn'}_{y^n y^{n'}} - K^{nn'}_{y^n y'} - K^{nn'}_{y y^{n'}} + K^{nn'}_{yy'}$$

$$K^{nn'}_{yy'} = \psi(x^n,y)^\top \psi(x^{n'},y')$$

Двойственная формулировка

$$\sum_{\substack{y\in\mathcal{Y}\\ n=1,\dots,N}} \alpha_{ny}\Delta(y^n,y) - \frac{1}{2}\sum_{\substack{y\in\mathcal{Y}\\ n=1,\dots,N}}\sum_{\substack{y'\in\mathcal{Y}\\ n'=1,\dots,N}} \alpha_{ny}\alpha_{n'y'}\bar{K}^{nn'}_{yy'} \to \max_{\alpha\geq 0}$$

$$\text{s.t.:}\quad \sum_{y\in\mathcal{Y}}\alpha_{ny} \leq \frac{C}{N},\quad n=1,\dots,N$$

$$\bar{K}^{nn'}_{yy'} = K^{nn'}_{y^n y^{n'}} - K^{nn'}_{y^n y'} - K^{nn'}_{y y^{n'}} + K^{nn'}_{yy'}$$

$$K^{nn'}_{yy'} = \underbrace{k(\,(x^n,y),(x^{n'},y')\,)}_{\text{kernel trick}}$$

Двойственная формулировка

$$\sum_{\substack{y\in\mathcal{Y}\\ n=1,\dots,N}} \alpha_{ny}\Delta(y^n,y) - \frac{1}{2}\sum_{\substack{y\in\mathcal{Y}\\ n=1,\dots,N}}\sum_{\substack{y'\in\mathcal{Y}\\ n'=1,\dots,N}} \alpha_{ny}\alpha_{n'y'}\bar{K}^{nn'}_{yy'} \to \max_{\alpha\geq 0}$$

$$\text{s.t.:}\quad \sum_{y\in\mathcal{Y}}\alpha_{ny} \leq \frac{C}{N},\quad n=1,\dots,N$$

$$\bar{K}^{nn'}_{yy'} = K^{nn'}_{y^n y^{n'}} - K^{nn'}_{y^n y'} - K^{nn'}_{y y^{n'}} + K^{nn'}_{yy'}$$

$$K^{nn'}_{yy'} = k(\,(x^n,y),(x^{n'},y')\,)$$

$$g(x,y) = \sum_{\substack{y'\in\mathcal{Y}\\ n=1,\dots,N}} \alpha_{ny'}k(\,(x^n,y'),(x,y)\,)$$

Отложенная генерация ограничений (Cutting-Plane)

$$\frac{1}{2}\|w\|^2 + C \cdot \frac{1}{N}\sum_{n=1}^{N}\xi^n \to \min_{x,\xi}$$

$$w^\top\left(\psi(x^n, y^n) - \psi(x^n, y)\right) \geq \Delta(y^n, y) - \xi^n$$

$$\forall y \in \mathcal{Y}, n = 1, \ldots, N$$

- Проблема: экспоненциально много ограничений ($|\mathcal{Y}| = |\mathcal{Y}_i|^M$)
- Предположение: много неактивных ограничений
- Идея: оптимизировать, используя не все ограничения, итеративно добавляя наиболее нарушаемые

Алгоритм Cutting-Plane

Пусть S — полный набор ограничений.

- $S_{use} \leftarrow \emptyset$
- Повторять до сходимости
  1. Решаем QP, используя ограничения S_{use}. Получаем w^*.
  2. Находим конфигурацию (значение y), соответствующую "максимально нарушенному ограничению":
     $y^* = \arg\max_y \{\Delta(y^n, y) + w^\top \psi(x^n, y)\}$
  3. Если не выполняется (с учетом точности)
     ${w^*}^\top (\psi(x^n, y^n) - \psi(x^n, y^*)) \geq \Delta(y^n, y) - \xi^n$
     то добавляем это ограничение в S_{use}
- Обученный параметр: последнее w^*.

Theorem (Tsochantaridis et al, 05)

Алгоритм сходится за $O(\frac{1}{\varepsilon^2})$ итераций, при этом значение функционала будет не более чем на ε больше оптимального значения.

Алгоритм Cutting-Plane

Алгоритм секущих плоскостей использует т.н. **отделяющий оракул**, решающий задачу:

$$\Delta(y^n, y) + w^\top \psi(x^n, y) \to \max_y$$

Обучение CRF:

- Оракул решает задачу вывода
- Ограничение на функцию потерь

$$\Delta(y^n, y) = \sum_i \sum_p y_{ip} \Delta_{ip}(y^n) + \sum_{ij} \sum_{pq} y_{ip} y_{jq} \Delta_{ij,pq}(y^n) + \dots$$

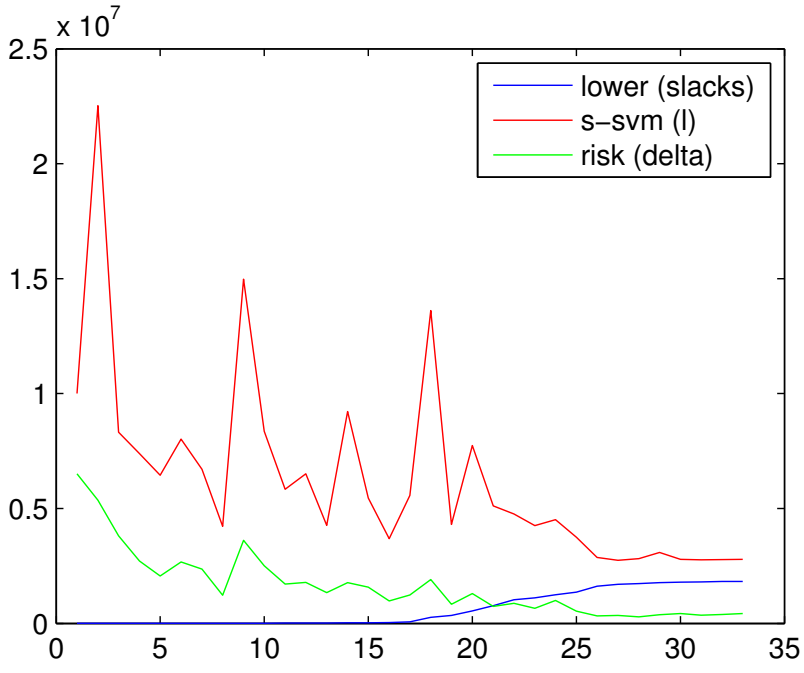
- Вывод в CRF: NP-трудная задача, во многих случаях решается приближенно

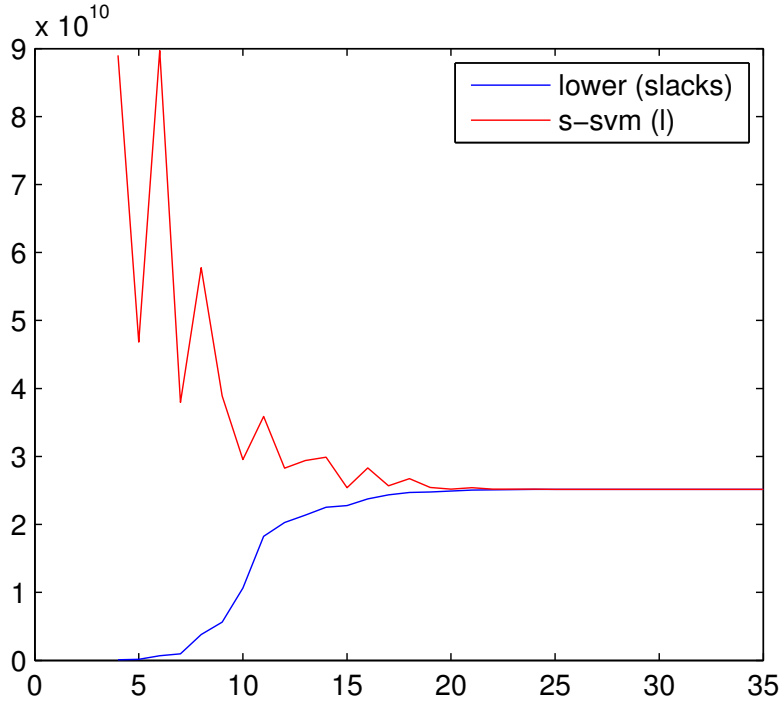
Приближенный оракул

- **undergenerating**: $\mathcal{Y}_{under} \subset \mathcal{Y}$
  Алгоритмы вывода в графических моделях, которые находят разметку (Loopy-BP, покоординатный подъем, α-expansion).
- **overgenerating**: $\mathcal{Y} \subset \mathcal{Y}_{over}$
  LP-релаксация, $\mathcal{Y}_{over} = [0, 1]^M$.

Результаты экспериментов Finley and Joachims [2008]:
Подход overgenerating работает лучше.

Алгоритм Cutting-Plane





Одна невязка

$$\frac{1}{2}\|w\|^2 + C\xi \to \min_{w,\xi}$$

$$\frac{1}{N}\sum_{n=1}^{N} w^\top \Big(\psi(x^n, y^n) - \psi(x^n, y^{(n)})\Big) \geq \frac{1}{N}\sum_{i=1}^{N} \Delta(y^n, y^{(n)}) - \xi$$

$$\forall\ (y^{(1)}, \ldots, y^{(n)}) \in \mathcal{Y}^N$$

Theorem (Joachims et al, 09)

Алгоритм Cutting-Plane сходится за $O(\frac{1}{\varepsilon})$ итераций, при этом значение функционала будет не более чем на ε больше оптимального значения.