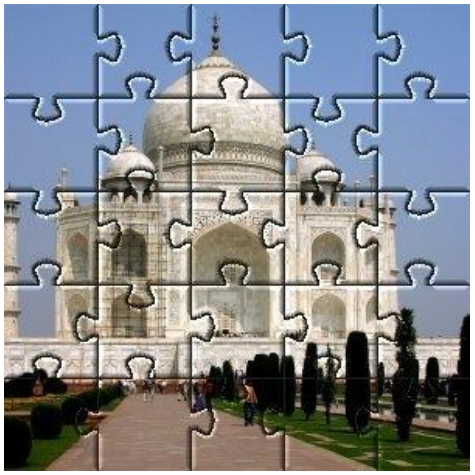
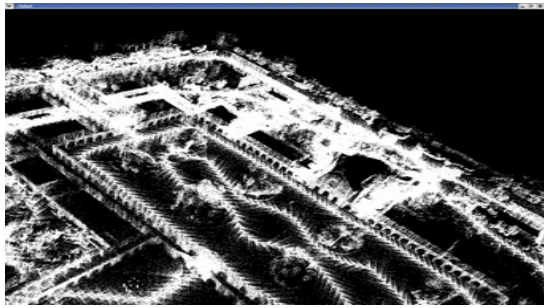
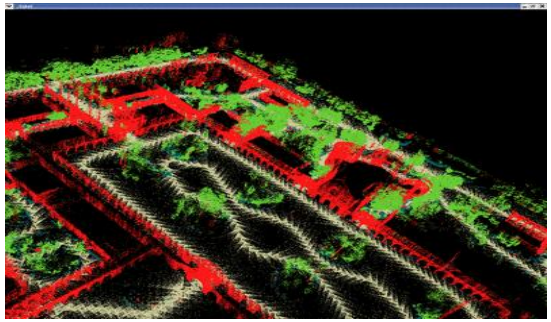
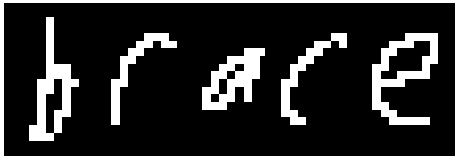
Структурное обучение. Структурный SVM









brace

Рома Шаповалов

29 апреля 2010

Задачи обучения с учителем

- Классификация:
  – подбор параметров функции $\mathbb{R}^m \to \{c_1, c_2, \ldots, c_k\}$

Задачи обучения с учителем

- Классификация:
  – подбор параметров функции $\mathbb{R}^m \to \{c_1, c_2, \ldots, c_k\}$
- Регрессия (упорядоченный выход):
  – подбор параметров функции $\mathbb{R}^m \to \mathbb{R}$

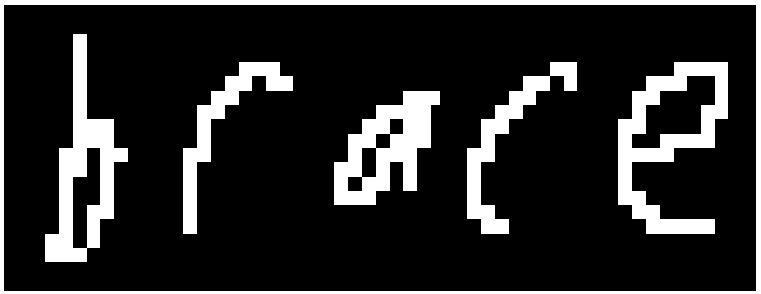
Задачи обучения с учителем

- Классификация:
  – подбор параметров функции $\mathbb{R}^m \to \{c_1, c_2, \ldots, c_k\}$
- Регрессия (упорядоченный выход):
  – подбор параметров функции $\mathbb{R}^m \to \mathbb{R}$
- Структурная классификация/регрессия (структурный выход):
  – подбор параметров функции $\mathbb{R}^{m \times l} \to \mathbb{R}^l$

- многомерный
- коррелированный
- ограниченный

Распознавание рукописного текста

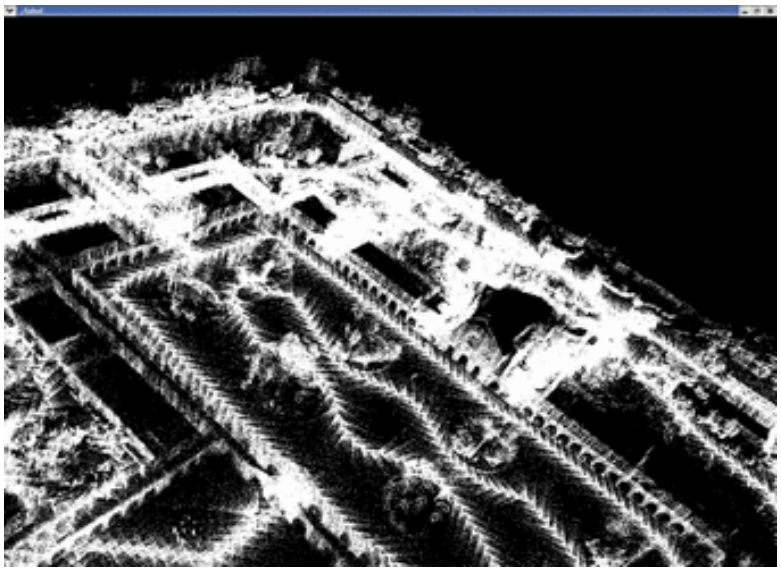
x **t**



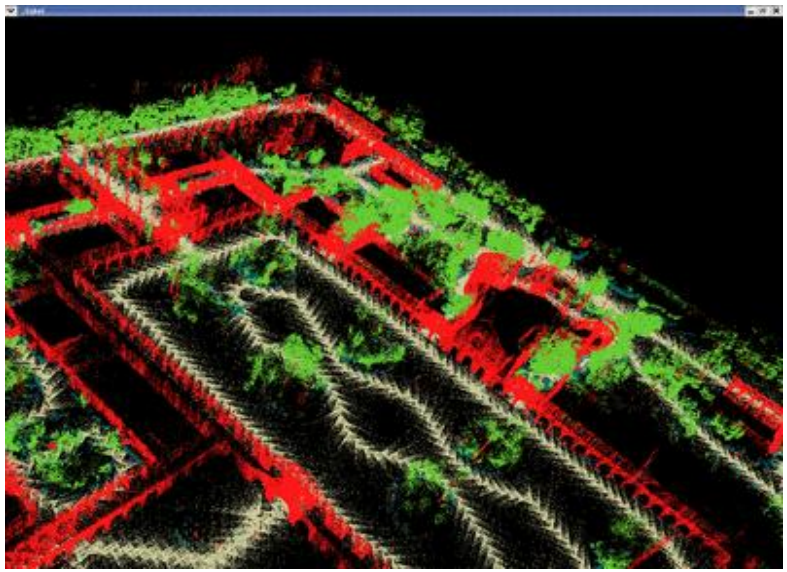
brace

Семантическая сегментация

x



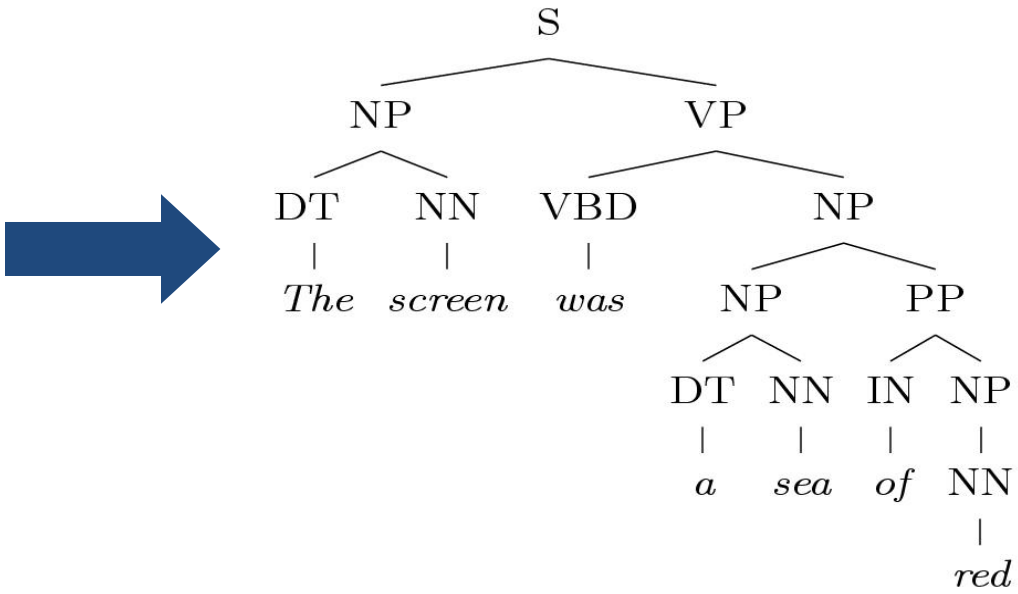
t



Разбор предложений естественных языков

x **t**

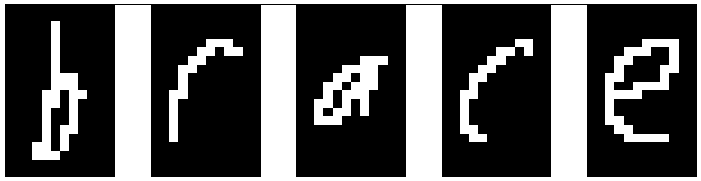
The screen was
a sea of red

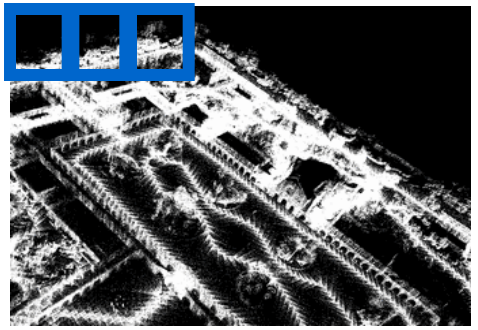


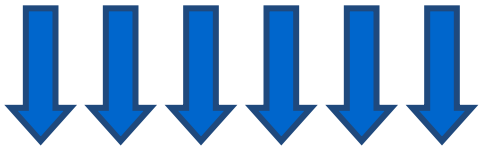
Другие применения

- Обучение MRF (зрение)
- Машинный перевод, парсинг (NLP)
- Выравнивание нуклеотидных последовательностей (биоинформатика)
- Ранжирование результатов поиска (IR)
- Транскрибирование речи (обработка речи)
- Планирование действий (роботика)

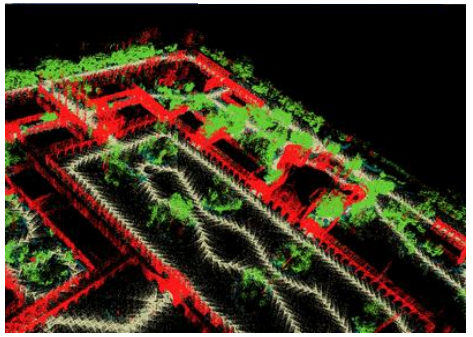
Локальное предсказание







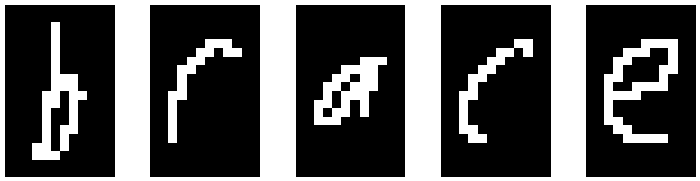
b r a c e

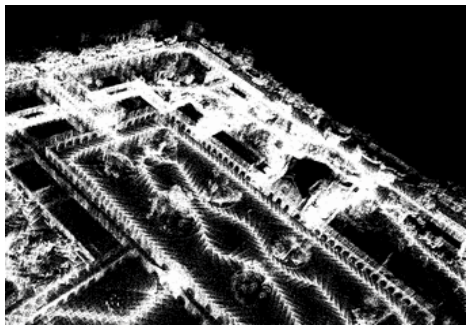


Классификация локальной информации

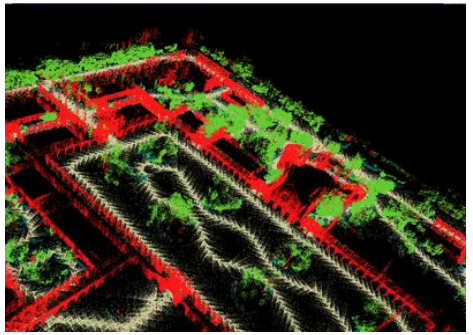
⇒ Игнорируются корреляции и ограничения!

Структурное предсказание





b r a c e



- Используется локальная информация
- Используются корреляции

Outline

- Структурное обучение
- Структурный SVM
- Оптимизация:
  - cutting-plane оптимизация

Max-margin estimation

- Необходимо:

$$\arg\max_{\mathbf{t}} \; \mathbf{w}^{\top}\Psi(\text{  }, \mathbf{t}) \;=\; \text{“brace”}$$

- Эквивалентно:

$$\mathbf{w}^{\top}\Psi(\text{  }, \text{“brace”}) \;>\; \mathbf{w}^{\top}\Psi(\text{  }, \text{“aaaaa”})$$

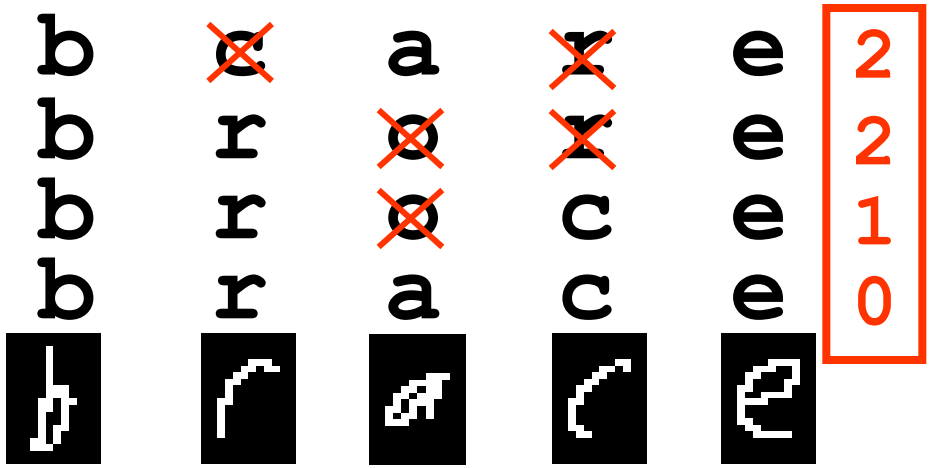
$$\mathbf{w}^{\top}\Psi(\text{  }, \text{“brace”}) \;>\; \mathbf{w}^{\top}\Psi(\text{  }, \text{“aaaab”})$$

$$\ldots$$

$$\mathbf{w}^{\top}\Psi(\text{  }, \text{“brace”}) \;>\; \mathbf{w}^{\top}\Psi(\text{  }, \text{“zzzzz”})$$

много!

Structured Loss



[1] B. Taskar, V. Chatalbashev, and D. Koller, "Learning associative Markov networks," *International Conference on Machine Learning*, Banff, Alberta, Canada: 2004, pp. 102-109.

Cutting-plane training

1: Input: $S = ((x_1, y_1), \ldots, (x_n, y_n))$, C, ε
2: $\mathcal{W}_i \leftarrow \emptyset$, $\xi_i \leftarrow 0$ for all $i = 1, \ldots, n$
3: **repeat**
4: **for** i=1,...,n **do**
5: $\hat{y} \leftarrow \operatorname{argmax}_{\hat{y} \in \mathcal{Y}} \{\Delta(y_i, \hat{y}) - \boldsymbol{w}^T [\Psi(x_i, y_i) - \Psi(x_i, \hat{y})]\}$
6: **if** $\Delta(y_i, \hat{y}) - \boldsymbol{w}^T [\Psi(x_i, y_i) - \Psi(x_i, \hat{y})] > \xi_i + \varepsilon$ **then**
7: $\mathcal{W}_i \leftarrow \mathcal{W}_i \cup \{\hat{y}\}$
8: $(\boldsymbol{w}, \boldsymbol{\xi}) \leftarrow \operatorname{argmin}_{\boldsymbol{w}, \boldsymbol{\xi} \geq \mathbf{0}} \frac{1}{2} \boldsymbol{w}^T \boldsymbol{w} + \frac{C}{n} \sum_{i=1}^{n} \xi_i$

$$\text{s.t. } \forall \bar{y}_1 \in \mathcal{W}_1 : \boldsymbol{w}^T [\Psi(x_1, y_1) - \Psi(x_1, \bar{y}_1)] \geq \Delta(y_1, \bar{y}_1) - \xi_1$$
$$\vdots$$
$$\forall \bar{y}_n \in \mathcal{W}_n : \boldsymbol{w}^T [\Psi(x_n, y_n) - \Psi(x_n, \bar{y}_n)] \geq \Delta(y_n, \bar{y}_n) - \xi_n$$

9: **end if**
10: **end for**
11: **until** no $\mathcal{W}_i$ has changed during iteration
12: return$(\boldsymbol{w}, \xi)$

- T. Joachims, T. Finley, and C. Yu, "Cutting-plane training of structural SVMs," *Machine Learning*, vol. 77, 2009, pp. 27-59.

Cutting-plane training: анализ

- Учитывает общий вид функции потерь
- Итерационный метод, время работы пропорционально числу рассмотренных ограничений
- Полиномиальное число ограничений должно быть рассмотрено, чтобы достичь наперёд заданной точности

Другие методы

- Переход к двойственной подзадаче

  B. Taskar, V. Chatalbashev, and D. Koller, "Learning associative Markov networks," *International Conference on Machine Learning*, Banff, Alberta, Canada: 2004, pp. 102-109.

- Субградиентный метод

  N. Ratliff, J. Bagnell, and M. Zinkevich, "(Online) Subgradient Methods for Structured Prediction," *International Conference on Artificial Intelligence and Statistics*, 2007.
  http://www.ri.cmu.edu/pub_files/pub4/ratliff_nathan_2007_3/ratliff_nathan_2007_3.pdf

- Функциональный градиентный бустинг

  D. Munoz, J. Bagnell, N. Vandapel, and M. Hebert, "Contextual classification with functional Max-Margin Markov Networks," *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL: 2009, pp. 975-982.
  N. Ratliff, J. Bagnell, and S. Srinivasa. Imitation learning for loco- motion and manipulation. In Humanoids, 2007.

- Структурный экстраградиент

  Корпелевич Г.М. Экстраградиентный метод для отыскания седловых точек и других задач // Эконом. и мат. методы. 1976. Т.12, N4. С.747–756.
  Y **Nesterov.** Gradient methods for minimizing composite objective function. 2007.
  http://www.uclouvain.be/cps/ucl/doc/core/documents/coredp2007_76.pdf
  B. Taskar, S. Lacoste-Julien, and M. Jordan. Structured prediction, dual extragradient and Bregman projections. JMLR, pages 1627-1653, 2006.

Спасибо за внимание!

- Ben Taskar. *NIPS 2007* Tutorial on Structured Prediction (спасибо за материал! :). http://media.nips.cc/Conferences/2007/Tutorials/Videos/Taskar/viewer.html
- Chris Lampert. Machine learning of structured outputs. *Computer Vision Winter Workshop* 2011. http://pub.ist.ac.at/~chl/talks/lampert-cvww2011.pdf
- SVM^{struct} library by Thorsten Joachims. http://www.cs.cornell.edu/People/tj/svm_light/svm_struct.html
- Обучение при неточном выводе

  O. Meshi, D. Sontag, T. Jaakkola, and A. Globerson, "Learning Efficiently with Approximate Inference via Dual Losses," *International Conference on Machine Learning*, 2010.

Оценка правдоподобия

$$P_w(y|x) = \frac{\prod_p \exp\left(w^T \varphi(x_p, y_p)\right)}{\sum_{y' \in Y} \prod_p \exp\left(w^T \varphi(x_p, y_p')\right)}$$

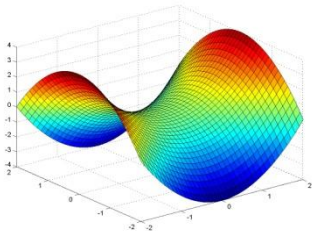
$$\rightarrow \max_w$$

- Функция выпуклая по весам, но знаменатель не считается

Другие методы: структурный экстраградиент

$$\min_{w,\xi} \frac{1}{2} w^T w + C \sum_i \xi_i$$

$$w^T \Psi(x_i, y_i) + \xi_i \geq \max_{\bar{y} \in Y} [w^T \Psi(x_i, \bar{y}) + \Delta(y_i, \bar{y})]$$



- Поиск седловой точки: линейная сходимость

Prediction:

$$\mathbf{w}^p = \pi_{\mathcal{W}}(\mathbf{w} - \beta \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \mathbf{z})) \qquad \mathbf{z}_i^p = \pi_{\mathcal{Z}_i}(\mathbf{z}_i + \beta \nabla_{\mathbf{z}_i} \mathcal{L}(\mathbf{w}, \mathbf{z}))$$

Correction:

$$\mathbf{w}^c = \pi_{\mathcal{W}}(\mathbf{w} - \beta \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}^p, \mathbf{z}^p)) \qquad \mathbf{z}_i^c = \pi_{\mathcal{Z}_i}(\mathbf{z}_i + \beta \nabla_{\mathbf{z}_i} \mathcal{L}(\mathbf{w}^p, \mathbf{z}^p))$$

Корпелевич Г.М. Экстраградиентный метод для отыскания седловых точек и других задач // Эконом. и мат. методы. 1976. Т.12, N4. С.747–756.

Y **Nesterov.** Gradient methods for minimizing composite objective function. 2007. http://www.uclouvain.be/cps/ucl/doc/core/documents/coredp2007_76.pdf

B. Taskar, S. Lacoste-Julien, and M. Jordan. Structured prediction, dual extragradient and Bregman projections. JMLR, pages 1627-1653, 2006.

Другие методы: субградиент

- Субградиентный метод

N. Ratliff, J. Bagnell, and M. Zinkevich, "(Online) Subgradient Methods for Structured Prediction," *International Conference on Artificial Intelligence and Statistics*, 2007. http://www.ri.cmu.edu/pub_files/pub4/ratliff_nathan_2007_3/ratliff_nathan_2007_3.pdf

Algorithm 1 MMSC subgradient calculation

procedure SUBGRADMMSC((x_i, y_i, v_i), $\mathcal{L}_i(y)$, $f_i : \mathcal{X} \rightarrow \mathbb{R}^d$, $w \in \mathcal{W}$)
  $y^* = \arg\max_{y \in \mathcal{Y}} w^T f_i(y) + \mathcal{L}_i(y)$
  $g \leftarrow g + f_i(y^*) - f_i(y_i)$
  return g
end procedure

Algorithm 3 Subgradient algorithm

procedure SUBGRAD($h = \sum_{i=1}^{m} h_i$, $w_0 \in \mathcal{W}$)
  $w_\alpha \leftarrow 0, \quad \alpha_s \leftarrow 0$
  for $t = 1, \ldots, T-1$ **do**
    Compute $g_t \leftarrow \nabla h(w_t)$
    Update $w_{t+1} \leftarrow \mathcal{P}_{\mathcal{W}}[w_t - \alpha_t g_t]$
  end for
  return $\arg\min_t h(w_t)$
end procedure

- Функциональный градиентный бустинг

D. Munoz, J. Bagnell, N. Vandapel, and M. Hebert, "Contextual classification with functional Max-Margin Markov Networks," *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL: 2009, pp. 975-982.

N. Ratliff, J. Bagnell, and S. Srinivasa. Imitation learning for loco- motion and manipulation. In Humanoids, 2007.

Что мы обучаем?

- Structured prediction:

$$h: X \to Y$$

- Вывод:

$$h(x) = \underset{y \in Y}{\operatorname{argmax}} f(x, y)$$

Линейная модель

- Structured prediction:

$$h: X \to Y$$

- Вывод:

$$h(x) = \underset{y \in Y}{\operatorname{argmax}} f(x, y)$$

- Предположение линейности:

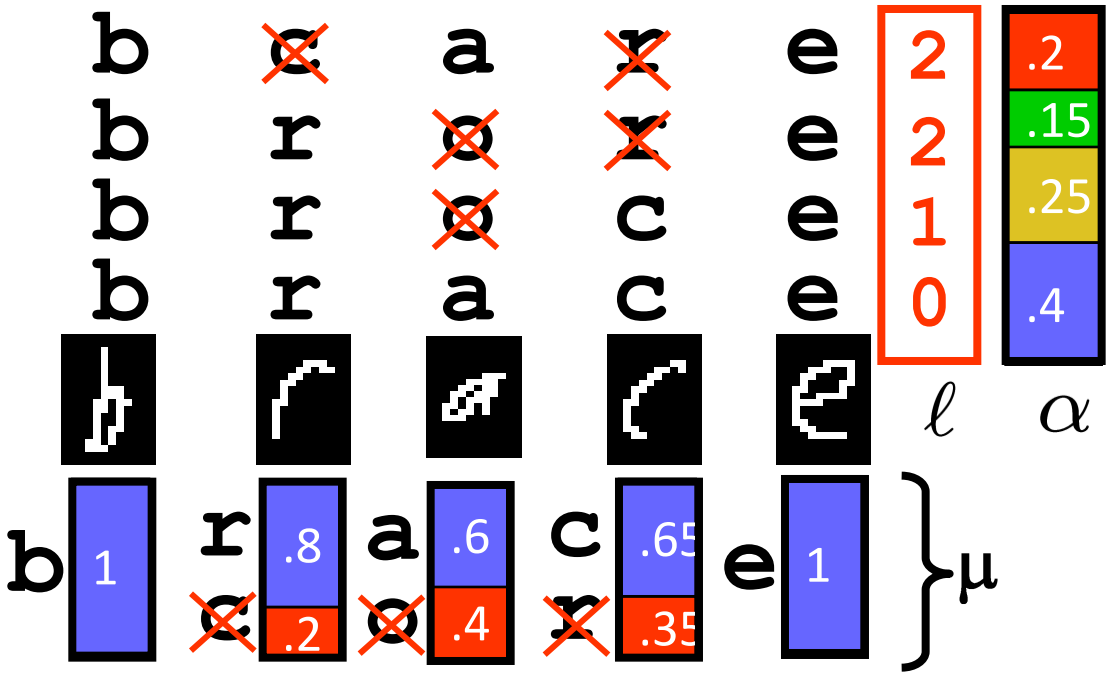
$$f(x, y) = \log P_w(y|x) = w^T \Psi(x, y)$$
$$= \sum_c w^T \varphi(x_c, y_c)$$

Оценка весов

- Цель: результат на обучающей выборке должен быть хорошим

- Локально (не учитывает структуру)
- Максимум правдоподобия (часто intractable)
- Max-margin

Связь двойственных



Формулировка структурного SVM

$$\min_{\mathbf{w},\boldsymbol{\xi}\geq\mathbf{0}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{C}{n}\sum_{i=1}^{n}\xi_i$$

$$s.t. \quad \forall \bar{y_1} \in \mathscr{Y} : \mathbf{w}^T[\Psi(x_1,y_1) - \Psi(x_1,\bar{y_1})] \geq \Delta(y_1,\bar{y_1}) - \xi_1$$

$$\vdots$$

$$s.t. \quad \forall \bar{y_n} \in \mathscr{Y} : \mathbf{w}^T[\Psi(x_n,y_n) - \Psi(x_n,\bar{y_n})] \geq \Delta(y_n,\bar{y_n}) - \xi_n$$

- Ограничения можно переформулировать:

$$w^T\Psi(x_i,y_i) + \xi_i \geq \max_{\bar{y}\in Y}[w^T\Psi(x_i,\bar{y}) + \Delta(y_i,\bar{y})]$$