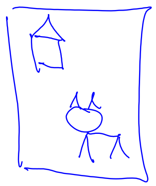
Lecture 22 - scribbles - latent

Wednesday, April 5, 2017
10:27

today:
- latent SVMstruct
- kernels
- deep learning?

I) Latent variables

motivation: semantic segmentation → finding boundary of different objects



segmentation is expensive → z "latent variable"
only have class labels → y

also: [Felzenszwalb & al. TPAMI 2010 "Deformable Part models" for object recognition]
↳ z there was an object configuration

before, we had $S(x,y;w) = \langle w, \varphi(x,y) \rangle$

now, consider $S(x,y,z;w) = \langle w, \varphi(x,y,\underline{z}) \rangle$

as before, could predict with $\underset{y \in \mathcal{Y},\, z \in \mathcal{Z}}{\arg\max}\ S(x,y,z;w)$

* CRF $(p(y|x)) \rightsquigarrow$ hidden CRF

similar to latent variable modeling with graphical model

ML $\rightsquigarrow$ Expectation-Maximization
↳ analog for SVMstruct is CCCP

latent SVMstruct:

$\ell(y, (\tilde{y}, \tilde{z}))$

generalization of structured hinge loss:

$$f(x,y,w) \triangleq \max_{(\tilde{y},\tilde{z})} \left(\langle w, \varphi(x,\tilde{y},\tilde{z})\rangle + \ell(y,(\tilde{y},\tilde{z})) \right) \underbrace{- \max_{\tilde{z}'} \langle w, \varphi(x,y,\tilde{z}')\rangle}_{\text{concave function of } w} \geq \ell(y, \underset{(\hat{y},\hat{z})}{h_w(x)})$$

recall: sup of convex function $\Rightarrow$ always convex

inf of jointly convex function is convex i.e. $\inf_z g(w,z)$ where g is jointly convex in w & z

$\triangleq \tilde{f}(w)$ is convex in w

here $f(x,y,w) = u(w) - v(w)$

where u & v are convex functions of w "difference of convex functions"

$\hookrightarrow$ CCCP procedure is used to minimize this

CCCP procedure:

- linearize $v(w)$ at w_t to get upper bound
- minimize the upper bound
- repeat

$\rightarrow$ majorization-minimization procedure (EM is another example of that)

$$\left[\begin{array}{l} f_t(w) = u(w) - [v(w_t) + \langle \nabla v(w_t), w - w_t\rangle] \\ \qquad\qquad (\text{or subgradient}) \\ w_{t+1} = \arg\min_w f_t(w) \end{array}\right.$$

properties of this procedure:

- like EM, descent procedure i.e. $f(w_{t+1}) \leq f(w_t)$

$$f(w_t) = f_t(w_t) \geq f_t(w_{t+1}) \geq f(w_{t+1})$$

- local linear convergence to stationary point

• local linear convergence to stationary point for latent SVMstruct [see NIPS OPT 2012 paper]

for SVMstruct $v(w) = \max_{\tilde{z}} \langle w, \varphi(x,y,\tilde{z}) \rangle$

$\partial v(w_t) = \varphi(x,y,\hat{z}(x,w_t))$ argmax$_{\tilde{z}} \langle w, \varphi(x,y,\tilde{z}) \rangle$

$\Rightarrow f_t(x,y,w) \doteq \max_{(\tilde{y},\tilde{z})} \langle w, \varphi(x,\tilde{y},\tilde{z}) \rangle + \ell(y,(\tilde{y},\tilde{z})) - \langle w, \varphi(x,y,\hat{z}_t) \rangle + cst.$

→ like SVMstruct objective

CCCP for latent SVM struct repeat:
- fill in $\hat{z}_t^{(i)}$ for all ground truth $y^{(i)}$ using w_t
- solve standard SVMstruct to get w_{t+1}
- repeat

Kernels:

so far $s(x,y;w) = \langle w, \varphi(x,y) \rangle$

recall for both SVMstruct, CRF $w(\alpha) = \frac{1}{\lambda n} \sum_{i,\tilde{y}} \alpha_i(\tilde{y}) \psi_i(\tilde{y})$ → $\varphi(x^{(i)},y^{(i)}) - \varphi(x^{(i)},\tilde{y})$

$\langle w, \varphi(x,y) \rangle = \frac{1}{\lambda n} \sum_{i,\tilde{y}} \alpha_i(\tilde{y}) \underbrace{\langle \psi_i(\tilde{y}), \varphi(x,y) \rangle}_{K((x^{(i)},y^{(i)});(x,y)) - K((x^{(i)},\tilde{y});(x,y))}$

$K((x,y);(x',y')) \triangleq \langle \varphi(x,y), \varphi(x',y') \rangle$

examples: $K(\ldots) = k_x(x,x') \cdot k_y(y,y')$

OCR $\quad \varphi(x,y) = \sum_p \varphi_n(x_p, y_p) + \sum_p \varphi_e(y_p, y_{p+1})$

$$\langle \varphi(x,y), \varphi(x',y')\rangle = \sum_{p,p'} \left(\langle \varphi_n(x_p,y_p), \varphi_n(x'_{p'}, y'_{p'})\rangle + \langle \varphi_e(y_p,y_{p+1}), \varphi_e(y'_{p'}, y'_{p'+1})\rangle\right)$$

$$[\varphi_n \perp \varphi_e]$$

for OCR, we had used $\varphi_n(x_p,y_p) = \begin{pmatrix}0\\ \mathrm{vec}(x_p)\\ 0\end{pmatrix} \leftarrow y_p$ position

$$\Rightarrow K\left((x_p,y_p),(x'_{p'},y'_{p'})\right) = \mathbb{1}\{y_p = y'_{p'}\}\langle x_p, x'_{p'}\rangle$$

but instead, could use $\longrightarrow$ $\mathbb{1}\{y_p=y'_{p'}\} \underbrace{\exp\left(-\frac{\|x_p - x'_{p'}\|^2}{2\sigma^2}\right)}_{\text{RBF kernel}}$

* binary SVM with RBF kernel $\rightsquigarrow$ similar to nearest neighbor classifier

Computation:

BCFW: stored $w_i \rightarrow O(d)$ for each i

$$\langle w, \varphi(x,y)\rangle \rightsquigarrow O(d)$$

with kernel: cannot store w, so maintain $\alpha_i(\tilde y)$ instead...

$$w = \frac{1}{\lambda n}\sum_{i,\tilde y} \alpha_i(\tilde y)\psi_i(\tilde y)$$ $\leftarrow$ (super) sparse when using FW

after t iterations of BCFW, only t non-zero variables

$$\langle w, \varphi(x,y)\rangle \rightarrow O(t)$$

kernel methods $\rightsquigarrow O(n^2)$

primal $\rightsquigarrow O(d \cdot n)$

Deep learning

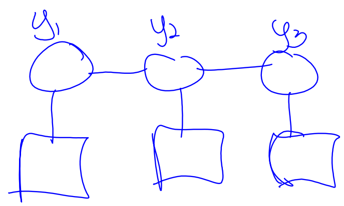
go from $\langle w, \varphi(x,y) \rangle$ to $\langle w, \varphi(x,y,\theta) \rangle$

can learn θ

I) plug in "deep learning" features in a structured prediction model

example OCR:

images:



example:
[Vu et al. ICCV 2015] "context-aware CNNs for person head detection"

II) recurrent neural network (RNN)

$$p(y|x) = \prod_{t \le T} p(y_t | y_{1:t-1}, x) \xrightarrow{\text{graphical model approach}} \prod_t p(y_t | \underbrace{y_{\pi_t}}_{\text{conditional independence assumptions}}, x) \quad \text{(directed)}$$

$$\frac{1}{Z(x)} \prod_c \psi_c(y_c, x) \quad \text{(undirected)}$$

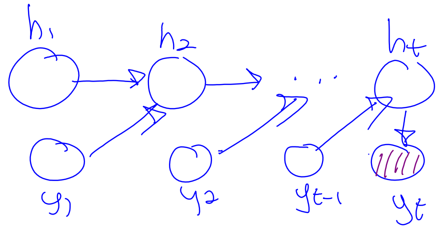
chain rule

RNN $\rightarrow$ structured parametrization of $p(y_t | y_{1:t-1}, x)$

$$h_{t+1} = f(h_t, x, y_t, W)$$

$$p(y_t | y_{1:t-1}, x) \propto \exp\left(c(y_t)^T \tilde{W} h_t \right)$$

$$f(f(\ldots\ldots y_{t-2}), x, y_{t-1}, W)$$



NN architecture for $p(y_t | y_{1:t-1}, x)$

learning: maximum likelihood $\log p(y|x) = \sum_t \log p(y_t | y_{1:t-1}, x)$ "teacher forcing"

("easy" apart the non-convexity)

decoding: $\arg\max_y \sum_t \log p(y_t | y_{1:t-1}, x)$ → need approximation

e.g. "beam search" (see Friday)

Pointers

- latent variable SVMstruct:
  - Chun-Nam Yu, Thorsten Joachims. "Learning Structural SVMs with Latent Variables". ICML 2009

    others:
    - hidden CRF: A. Quattoni, S. Wang, L. Morency, M. Collins, and T. Darrell, "Hidden Conditional Random Fields," TPAMI 2007.
    - deformable part models for object recognition (highly cited paper): Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, "Object Detection with Discriminatively Trained Part Based Models" TPAMI, Vol. 32, No. 9, September 2010
- CCCP procedure convergence rate:
  - Ian E.H. Yen, Nanyun Peng, Po-Wei Wang and Shou-de Lin, "On Convergence Rate of Concave-Convex Procedure", NIPS 2012 OPT Workshop (not considered a publication by the way)
- kernels:
  - example of early paper presenting kernels for structured SVM: Juho Rousu, Craig Saunders, Sandor Szedmak, John Shawe-Taylor, "Kernel-Based Learning of Hierarchical Multilabel Classification Models", JMLR 2006
  - example of application: L. Bertelli, T. Yu, D. Vu, and B. Gokturk, "Kernelized structural SVM learning for supervised object segmentation," CVPR 2011
  - for computation in BCFW, see Appendix B.5 in the usual: S. Lacoste-Julien, M. Jaggi, M. Schmidt and P. Pletscher, "Block-Coordinate Frank-Wolfe Optimization for Structural SVMs", ICML 2013
  - the standard book for kernels: Bernhard Schölkopf and Alexander J. Smola, Learning with Kernels, MIT Press 2001
- deep learning:

- see chapter 10 of the "Deep learning book" for RNNs
- head detection plug in example mentioned in class: Tuan-Hung Vu, Anton Osokin, and Ivan Laptev, "Context-aware CNNs for person head detection", ICCV 2015