CS 446/ECE 449: Machine Learning

A. G. Schwing

University of Illinois at Urbana-Champaign, 2020

L24: Structured Latent Variable Models

Goals of this lecture

- Getting to know Structured Latent Variable Models
- Learning about Hidden Markov Models (HMMs)

Reading material:

- K. Murphy; Machine Learning: A Probabilistic Perspective; Chapter 17

Recap:

Generative:

$$\ln p_\theta(x^{(i)}) = \ln \sum_{\boldsymbol{z}} p_\theta(x^{(i)}, \boldsymbol{z})$$

Discriminative:

$$\ln p_{\boldsymbol{w}}(\boldsymbol{y}|x^{(i)})$$

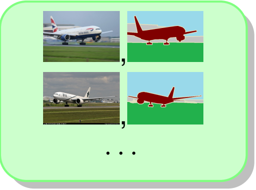
Observations:

- $\boldsymbol{z}$ never observed (assignment to cluster center in GMM)
- $\boldsymbol{z}$ very simple categorical variable
- $\boldsymbol{y}$ always completely observed

Questions: structure in $\boldsymbol{z}$ or parts of $\boldsymbol{y}$ unobserved

Recap: Learning with full observations

- Training set of data pairs $(x^{(i)}, \boldsymbol{y}^{(i)})$



- Inference

$$\arg\max_{\hat{\boldsymbol{y}}} F(\boldsymbol{w}, x^{(i)}, \hat{\boldsymbol{y}}) = \arg\max_{\hat{\boldsymbol{y}}} \sum_r f_r(\boldsymbol{w}, x^{(i)}, \hat{\boldsymbol{y}}_r)$$

- Learning target

$$\forall \hat{\boldsymbol{y}} \quad F(\boldsymbol{w}, x^{(i)}, \hat{\boldsymbol{y}}) \leq F(\boldsymbol{w}, x^{(i)}, \boldsymbol{y}^{(i)})$$

$$\max_{\hat{\boldsymbol{y}}} \quad F(\boldsymbol{w}, x^{(i)}, \hat{\boldsymbol{y}}) \leq F(\boldsymbol{w}, x^{(i)}, \boldsymbol{y}^{(i)})$$

Recap: Learning with full observations

Hinge loss: penalize whenever maximum is within a margin $L(\hat{\boldsymbol{y}}, \boldsymbol{y}^{(i)})$ of the data $(x^{(i)}, \boldsymbol{y}^{(i)})$ score:

$$\max_{\hat{\boldsymbol{y}}} \left(F(\boldsymbol{w}, x^{(i)}, \hat{\boldsymbol{y}}) + L(\hat{\boldsymbol{y}}, \boldsymbol{y}^{(i)}) \right) \geq F(\boldsymbol{w}, x^{(i)}, \boldsymbol{y}^{(i)})$$

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i \in \mathcal{D}} \left(\underbrace{\max_{\hat{\boldsymbol{y}}} \left(F(\boldsymbol{w}, x^{(i)}, \hat{\boldsymbol{y}}) + L(\hat{\boldsymbol{y}}, \boldsymbol{y}^{(i)}) \right)}_{\text{Loss-augmented inference}} - F(\boldsymbol{w}, x^{(i)}, \boldsymbol{y}^{(i)}) \right)$$

Parts of $\boldsymbol{y}$ unobserved/not annotated:

Fully labeled:

Weakly labeled:

Complete Data:

$$\boldsymbol{y} = (\boldsymbol{s}, \boldsymbol{z})$$

Parts of $\boldsymbol{y}$ unobserved/not annotated:

- Complete Data:

$$\boldsymbol{y} = (\boldsymbol{s}, \boldsymbol{z})$$

- Weakly labeled hinge loss: penalize whenever best overall prediction exceeds best prediction with annotation being clamped

$$\max_{\hat{\boldsymbol{s}}, \hat{\boldsymbol{z}}} F(\boldsymbol{w}, x^{(i)}, \hat{\boldsymbol{s}}, \hat{\boldsymbol{z}}) \geq \max_{\hat{\boldsymbol{z}}} F(\boldsymbol{w}, x^{(i)}, \boldsymbol{s}^{(i)}, \hat{\boldsymbol{z}})$$

- Latent SSVM (LSSVM) [Yu and Joachims 2009]:

$$\frac{C}{2}\|w\|_2^2 + \sum_{i \in \mathcal{D}} \max_{\hat{\boldsymbol{s}}, \hat{\boldsymbol{z}}} F(\boldsymbol{w}, x^{(i)}, \hat{\boldsymbol{s}}, \hat{\boldsymbol{z}}) - \sum_{i \in \mathcal{D}} \max_{\hat{\boldsymbol{z}}} F(\boldsymbol{w}, x^{(i)}, \boldsymbol{s}^{(i)}, \hat{\boldsymbol{z}})$$

Structured Prediction with Latent Variables

$$\min_w \frac{C}{2}\|w\|_2^2 + \sum_i \epsilon \ln \sum_{\hat{\mathbf{s}},\hat{\mathbf{z}}} \exp \frac{F(\mathbf{w}, x^{(i)}, \hat{\mathbf{s}}, \hat{\mathbf{z}})) + L_i(\hat{\mathbf{s}}, \hat{\mathbf{z}})}{\epsilon} - $$
$$- \sum_i \epsilon \ln \sum_{\hat{\mathbf{z}}} \exp \frac{F(\mathbf{w}, x^{(i)}, \mathbf{s}^{(i)}, \hat{\mathbf{z}}) + L_i^c(\mathbf{s}^{(i)}, \hat{\mathbf{z}})}{\epsilon}$$

- Soft-max function

$$\epsilon \ln \sum_{\hat{\mathbf{s}},\hat{\mathbf{z}}} \exp \frac{F(\mathbf{w}, x^{(i)}, \hat{\mathbf{s}}, \hat{\mathbf{z}})}{\epsilon} \quad \xrightarrow{\epsilon\to 0} \quad \max_{\hat{\mathbf{s}},\hat{\mathbf{z}}} F(\mathbf{w}, x^{(i)}, \hat{\mathbf{s}}, \hat{\mathbf{z}})$$

- LSSVM & hidden CRFs (HCRFs) [Lafferty et al. 2001]

$\epsilon = 0$	hinge-loss	max-margin	LSSVM
$\epsilon = 1$	log-loss	max-likelihood	HCRF

- Margin L

Recall: variational expression for partition function

$$\epsilon \ln \sum_{\boldsymbol{z}} \exp \frac{F(\boldsymbol{w}, x^{(i)}, \boldsymbol{s}^{(i)}, \boldsymbol{z})}{\epsilon} = \max_{q(\boldsymbol{z})\in\Delta} \left(\sum_{\boldsymbol{z}} q(\boldsymbol{z}) F(\boldsymbol{w}, x^{(i)}, \boldsymbol{s}^{(i)}, \boldsymbol{z}) + \epsilon H(q(\boldsymbol{z})) \right)$$

- Similarity to structured inference
- Similar algorithms can be employed

$$\begin{aligned} \min_{w} \quad & \frac{C}{2}\|w\|_2^2 + \sum_i \left(\epsilon \ln \sum_{\hat{\boldsymbol{s}},\hat{\boldsymbol{z}}} \exp \frac{F(\boldsymbol{w}, x^{(i)}, \hat{\boldsymbol{s}}, \hat{\boldsymbol{z}})) + L_i(\hat{\boldsymbol{s}}, \hat{\boldsymbol{z}})}{\epsilon} \right. \\ & \left. - \max_{q_i(\boldsymbol{z})\in\Delta} \left(\sum_{\boldsymbol{z}} q_i(\boldsymbol{z}) \left(F(\boldsymbol{w}, x^{(i)}, \boldsymbol{s}^{(i)}, \boldsymbol{z}) + L_i^c(\boldsymbol{s}^{(i)}, \hat{\boldsymbol{z}}) \right) + \epsilon H(q_i(\boldsymbol{z})) \right) \right) \end{aligned}$$

Alternating optimization between q and $\boldsymbol{w}$

Algorithmic structure

HCRF, LSSVM:

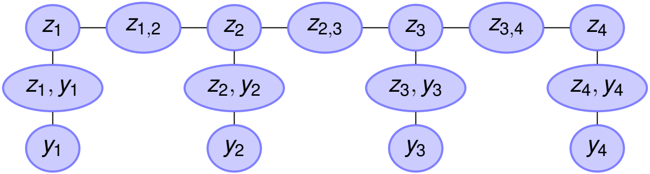
repeat
- repeat
  - latent variable prediction to obtain $q(\boldsymbol{z})$
- until convergence
- repeat
  - update w
- until convergence

until convergence

Irrespective of whether $\boldsymbol{z}$ sometimes observed or never observed

Example of model with latent variables:

Hidden Markov Model



Algorithm:

- Inferring $q(\boldsymbol{z})$: forward/backward pass on a chain graph
- Updating $\boldsymbol{w}$: loss-augmented inference on a tree graph

Applications of HMMs:

- Time-series data in general
- Video data
- Language models

Directly applicable to generative modeling with structured $\boldsymbol{z}$

Quiz:

- Variational expression of partition function?
- Structured prediction with latent variables?
- Algorithmic structure of general framework?

Important topics of this lecture

- General framework for structured prediction with latent variables
- Similarity between generative and discriminative techniques

What's next

- Variational Auto-Encoders