Learning Structured Predictors

Xavier Carreras

Universitat Politècnica de Catalunya

thanks to: M. Collins, A. Globerson, T. Koo, A. Quattoni

Supervised (Structured) Prediction

- Learning to predict: given training data

$$\{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), (\mathbf{x}^{(2)}, \mathbf{y}^{(2)}), \ldots, (\mathbf{x}^{(m)}, \mathbf{y}^{(m)})\}$$

  learn a predictor $\mathbf{x} \to \mathbf{y}$ that *works well* on unseen inputs $\mathbf{x}$

- Non-Structured Prediction: outputs $\mathbf{y}$ are atomic
  - Binary prediction: $\mathbf{y} \in \{-1, +1\}$
  - Multiclass prediction: $\mathbf{y} \in \{1, 2, \ldots, L\}$

- Structured Prediction: outputs $\mathbf{y}$ are structured
  - Sequence prediction: $\mathbf{y}$ are sequences
  - Parsing: $\mathbf{y}$ are trees
  - ...

Named Entity Recognition

$\mathbf{y}$	PER	-	QNT	-	-	ORG	ORG	-	TIME
$\mathbf{x}$	Jim	bought	300	shares	of	Acme	Corp.	in	2006

Named Entity Recognition

$\mathbf{y}$	PER	-	QNT	-	-	ORG	ORG	-	TIME
$\mathbf{x}$	Jim	bought	300	shares	of	Acme	Corp.	in	2006

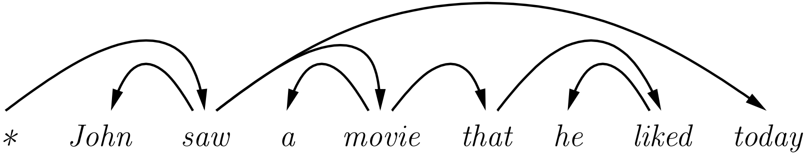
$\mathbf{y}$	PER	PER	-	-	LOC
$\mathbf{x}$	Jack	London	went	to	Paris

$\mathbf{y}$	PER	PER	-	-	LOC
$\mathbf{x}$	Paris	Hilton	went	to	London

Part-of-speech Tagging

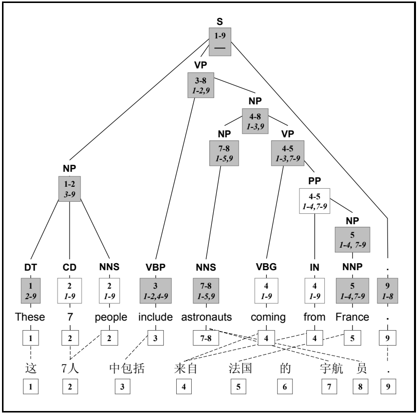
$\mathbf{y}$	NNP	NNP	VBZ	NNP	.
$\mathbf{x}$	Ms.	Haag	plays	Elianti	.

Syntactic Parsing



$\mathbf{x}$ are sentences
$\mathbf{y}$ are syntactic dependency trees

Machine Translation



(Galley et al 2006)

$\mathbf{x}$ are sentences in Chinese

$\mathbf{y}$ are sentences in English aligned to $\mathbf{x}$

Object Detection

(image removed)

(Kumar and Hebert 2003)

$\mathbf{x}$ are images

$\mathbf{y}$ are grids labeled with object types

Today's Goals

- Introduce basic tools for structure prediction
  - We will restrict to sequence prediction
- Understand what tools we can use from standard classification
  - Learning paradigms and algorithms, in essence, work here too
  - However, computations behind algorithms are prohibitive
- Understand what tools can we use from grammatical formalisms
  - We will borrow inference algorithms for tractable computations
  - E.g., algorithms for HMMs (Viterbi, forward-backward) will play a major role in today's methods

Today's Goals

- Introduce basic tools for structure prediction
  - We will restrict to sequence prediction
- Understand what tools we can use from standard classification
  - Learning paradigms and algorithms, in essence, work here too
  - However, computations behind algorithms are prohibitive
- Understand what tools can we use from grammatical formalisms
  - We will borrow inference algorithms for tractable computations
  - E.g., algorithms for HMMs (Viterbi, forward-backward) will play a major role in today's methods

Conditional Random Fields

for sequence prediction

$\mathbf{y}$	PER	PER	-	-	LOC
$\mathbf{x}$	Jack	London	went	to	Paris

Conditional Random Fields

(Lafferty, McCallum, Pereira 2001)

- Model the conditional distribution:

$$P(\mathbf{y}|\mathbf{x};\mathbf{w})$$

  where
  - $\mathbf{x} = \mathbf{x}_1\mathbf{x}_2\ldots\mathbf{x}_n \in \mathcal{X}^*$
  - $\mathbf{y} = \mathbf{y}_1\mathbf{y}_2\ldots\mathbf{y}_n \in \mathcal{Y}^*$ and $\mathcal{Y} = \{1,\ldots,L\}$
  - $\mathbf{w}$ are model parameters

- To predict the best sequence

$$\hat{\mathbf{y}} = \operatorname*{argmax}_{\mathbf{y}\in\mathcal{Y}^*} P(\mathbf{y}|\mathbf{x})$$

Conditional Random Fields

(Lafferty, McCallum, Pereira 2001)

- Model the conditional distribution:

$$P(\mathbf{y}|\mathbf{x}; \mathbf{w})$$

  where

  - $\mathbf{x} = \mathbf{x}_1\mathbf{x}_2 \ldots \mathbf{x}_n \in \mathcal{X}^*$
  - $\mathbf{y} = \mathbf{y}_1\mathbf{y}_2 \ldots \mathbf{y}_n \in \mathcal{Y}^*$ and $\mathcal{Y} = \{1, \ldots, L\}$
  - $\mathbf{w}$ are model parameters

- To predict the best sequence

$$\hat{\mathbf{y}} = \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}^*} P(\mathbf{y}|\mathbf{x})$$

- Problem: exponentially many $\mathbf{y}$'s for a given input $\mathbf{x}$

Compatibility Function

- Think of a compatibility function

$$\phi(\mathbf{x}, \mathbf{y}; \mathbf{w}) \to \mathbb{R}$$

that gives high positive scores to compatible $(\mathbf{x}, \mathbf{y})$ pairs

- Using ϕ we define:

$$P(\mathbf{y}|\mathbf{x}; \mathbf{w}) = \frac{\exp\{\phi(\mathbf{x}, \mathbf{y}; \mathbf{w})\}}{\sum_{\mathbf{z} \in \mathcal{Y}^*} \exp\{\phi(\mathbf{x}, \mathbf{z}; \mathbf{w})\}}$$

- Predict: $\hat{\mathbf{y}} = \mathrm{argmax}_{\mathbf{y} \in \mathcal{Y}^*} P(\mathbf{y}|\mathbf{x}; \mathbf{w})$
- Choose ϕ so that $\hat{\mathbf{y}}$ can be computed efficiently.

Towards Efficient Compatibility Functions

- How can we define $\phi(\mathbf{x}, \mathbf{y}; \mathbf{w})$?
- How do we represent a pair $\langle \mathbf{x}, \mathbf{y} \rangle$?

$\mathbf{y}$	PER	PER	-	-	LOC
$\mathbf{x}$	Jack	London	went	to	Paris

- Approach: compute features of $\mathbf{x}$ and $\mathbf{y}$. How?
  - Look at individual assignments $\mathbf{y}_i$ (standard classification)
  - Look at the full assignment $\mathbf{y}$ (unnatural)
  - Look at bigrams of outputs labels $\langle \mathbf{y}_{i-1}, \mathbf{y}_i \rangle$

(higher-order n-grams of the output are also possible)

Towards Efficient Compatibility Functions

- How can we define $\phi(\mathbf{x}, \mathbf{y}; \mathbf{w})$?
- How do we represent a pair $\langle \mathbf{x}, \mathbf{y} \rangle$?

$\mathbf{y}$	PER	PER	-	-	LOC
$\mathbf{x}$	Jack	London	went	to	Paris

- Approach: compute features of $\mathbf{x}$ and $\mathbf{y}$. How?
  - Look at individual assignments $\mathbf{y}_i$ (standard classification)
  - Look at the full assignment $\mathbf{y}$ (unnatural)
  - Look at bigrams of outputs labels $(\mathbf{y}_{i-1}, \mathbf{y}_i)$

(higher-order n-grams of the output are also possible)

Towards Efficient Compatibility Functions

- How can we define $\phi(\mathbf{x}, \mathbf{y}; \mathbf{w})$?
- How do we represent a pair $\langle \mathbf{x}, \mathbf{y} \rangle$?

$\mathbf{y}$	PER	PER	-	-	LOC
$\mathbf{x}$	Jack	London	went	to	Paris

- Approach: compute features of $\mathbf{x}$ and $\mathbf{y}$. How?
  - Look at individual assignments $\mathbf{y}_i$ (standard classification)
  - Look at the full assignment $\mathbf{y}$ (unnatural)
  - Look at bigrams of outputs labels $(\mathbf{y}_{i-1}, \mathbf{y}_i)$

(higher-order n-grams of the output are also possible)

Bigram "Indicator" Features

	1	2	3	4	5
$\mathbf{y}$	PER	PER	-	-	LOC
$\mathbf{x}$	Jack	London	went	to	Paris

- Indicator features:

$$\mathbf{f}_j(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i) = \begin{cases} 1 & \text{if } \mathbf{x}_i = \text{"London" and} \\ & \quad \mathbf{y}_{i-1} = \text{PER and } \mathbf{y}_i = \text{PER} \\ 0 & \text{otherwise} \end{cases}$$

e.g., $\mathbf{f}_j(\mathbf{x}, 2, \text{PER}, \text{PER}) = 1, \quad \mathbf{f}_j(\mathbf{x}, 3, \text{PER}, \text{-}) = 0$

More Bigram "Indicator" Features

	1	2	3	4	5
$\mathbf{x}$	Jack	London	went	to	Paris
$\mathbf{y}$	PER	PER	-	-	LOC
$\mathbf{y}'$	PER	LOC	-	-	LOC
$\mathbf{y}''$	-	-	-	LOC	-
$\mathbf{x}'$	My	trip	to	London	...

$\mathbf{f}_1(\ldots) = 1$ iff $\mathbf{x}_i =$"London" and $\mathbf{y}_{i-1} = \text{PER}$ and $\mathbf{y}_i = \text{PER}$

$\mathbf{f}_2(\ldots) = 1$ iff $\mathbf{x}_i =$"London" and $\mathbf{y}_{i-1} = \text{PER}$ and $\mathbf{y}_i = \text{LOC}$

$\mathbf{f}_3(\ldots) = 1$ iff $\mathbf{x}_{i-1} \sim$/(in|to|at)/ and $\mathbf{x}_i \sim$/[A-Z]/ and $\mathbf{y}_i = \text{LOC}$

$\mathbf{f}_4(\ldots) = 1$ iff $\mathbf{y}_i = \text{LOC}$ and $\text{WORLD-CITIES}(\mathbf{x}_i) = 1$

$\mathbf{f}_5(\ldots) = 1$ iff $\mathbf{y}_i = \text{PER}$ and $\text{FIRST-NAMES}(\mathbf{x}_i) = 1$

Factored Compatibility Functions

- Define $\mathbf{f}(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i)$ as a vector of D features:

$$(\ \mathbf{f}_1(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i), \ldots, \mathbf{f}_j(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i), \ldots, \mathbf{f}_D(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i) \ )$$

- Let $\mathbf{w} \in \mathbb{R}^D$ and $\mathbf{f}(\cdot,\cdot,\cdot,\cdot) \in \mathbb{R}^D$. Let $\mathbf{y}_0 = \text{NULL}$.

$$\begin{aligned}
\phi(\mathbf{x}, \mathbf{y}, \mathbf{w}) &= \sum_{i=1}^{n} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{D} \mathbf{w}_j \mathbf{f}_j(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i)
\end{aligned}$$

- $\phi(\mathbf{x}, \mathbf{y}, \mathbf{w})$ factors into scores for label bigrams $(\mathbf{y}_{i-1}, \mathbf{y}_i)$
- This factorization will allow efficient algorithms
(intuitively, if $\mathbf{y} \neq \mathbf{y}'$ share bigrams, they will share scores)

Conditional Random Fields (CRFs)

- The model form is:

$$P(\mathbf{y}|\mathbf{x};\mathbf{w}) = \frac{\exp\{\phi(\mathbf{x},\mathbf{y},\mathbf{w})\}}{Z(\mathbf{x},\mathbf{w})} = \frac{\exp\{\sum_{i=1}^{n} \mathbf{w}\cdot\mathbf{f}(\mathbf{x},i,\mathbf{y}_{i-1},\mathbf{y}_i)\}}{Z(\mathbf{x},\mathbf{w})}$$

  where

$$Z(\mathbf{x},\mathbf{w}) = \sum_{\mathbf{z}\in\mathcal{Y}^*} \exp\left\{\sum_{i=1}^{n} \mathbf{w}\cdot\mathbf{f}(\mathbf{x},i,\mathbf{z}_{i-1},\mathbf{z}_i)\right\}$$

- Features $\mathbf{f}(\ldots)$ are given (they are problem-dependent)
- $\mathbf{w}\in\mathbb{R}^D$ are the parameters of the model
- CRFs are log-linear models on the feature functions

Conditional Random Fields: Three Problems

- Compute the probability of an output sequence $\mathbf{y}$ for $\mathbf{x}$

$$P(\mathbf{y}|\mathbf{x};\mathbf{w})$$

- Decoding: predict the best output sequence for $\mathbf{x}$

$$\underset{\mathbf{y}\in\mathcal{Y}^*}{\operatorname{argmax}} P(\mathbf{y}|\mathbf{x};\mathbf{w})$$

- Parameter estimation: given training data

$$\left\{(\mathbf{x}^{(1)},\mathbf{y}^{(1)}),(\mathbf{x}^{(2)},\mathbf{y}^{(2)}),\ldots,(\mathbf{x}^{(m)},\mathbf{y}^{(m)})\right\} \quad ,$$

learn parameters $\mathbf{w}$

Decoding with CRFs

- Given $\mathbf{w}$, given $\mathbf{x}$, find:

$$\begin{aligned}
\operatorname*{argmax}_{\mathbf{y}\in\mathcal{Y}^*} P(\mathbf{y}|\mathbf{x};\mathbf{w}) &= \frac{\exp\left\{\sum_{i=1}^{n} \mathbf{w}\cdot\mathbf{f}(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i)\right\}}{Z(\mathbf{x};\mathbf{w})} \\
&= \exp\left\{\sum_{i=1}^{n} \mathbf{w}\cdot\mathbf{f}(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i)\right\} \\
&= \sum_{i=1}^{n} \mathbf{w}\cdot\mathbf{f}(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i)
\end{aligned}$$

- We can use the Viterbi algorithm

Viterbi for CRFs

- Calculate in $O(nL^2)$:

$$\hat{\mathbf{y}} = \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}^n} \sum_{i=1}^{n} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i)$$

- Define (score of optimal sequence for $\mathbf{x}_{1:i}$ ending with $a \in \mathcal{Y}$):

$$\delta_i(a) = \max_{\mathbf{y} \in \mathcal{Y}^i : \mathbf{y}_i = a} \sum_{j=1}^{i} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, j, \mathbf{y}_{j-1}, \mathbf{y}_j)$$

- Use the following recursions, for all $a \in \mathcal{Y}$:

$$\begin{aligned} \delta_1(a) &= \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, 1, \mathbf{y}_0 = \text{NULL}, a) \\ \delta_i(a) &= \max_{b \in \mathcal{Y}} \delta_{i-1}(b) + \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, i, b, a) \end{aligned}$$

- The optimal score for $\mathbf{x}$ is $\max_{a \in \mathcal{Y}} \delta_n(a)$
- The optimal sequence $\hat{\mathbf{y}}$ can be recovered through *pointers*

Parameter Estimation in CRFs

- Given a training set

$$\{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), (\mathbf{x}^{(2)}, \mathbf{y}^{(2)}), \ldots, (\mathbf{x}^{(m)}, \mathbf{y}^{(m)})\} \quad ,$$

  estimate $\mathbf{w}$

- Define the conditional log-likelihood of the data:

$$L(\mathbf{w}) = \sum_{k=1}^{m} \log P(\mathbf{y}^{(k)} | \mathbf{x}^{(k)}; \mathbf{w})$$

- $L(\mathbf{w})$ measures how well $\mathbf{w}$ explains the data. A good value for $\mathbf{w}$ will give a high value for $P(\mathbf{y}^{(k)} | \mathbf{x}^{(k)}; \mathbf{w})$ for all $k = 1 \ldots m$.
- We want $\mathbf{w}$ that maximizes $L(\mathbf{w})$

Learning the Parameters of a CRF

- Recall first lecture on log-linear / maximum-entropy models
- Find:

$$\mathbf{w}^* = \operatorname*{argmax}_{\mathbf{w} \in \mathbb{R}^D} L(\mathbf{w}) - \frac{\lambda}{2} ||\mathbf{w}||^2$$

  where
  - The first term is the log-likelihood of the data
  - The second term is a regularization term, it penalizes solutions with large norm
  - λ is a parameter to control the trade-off between fitting the data and model complexity

Learning the Parameters of a CRF

- Find
$$\mathbf{w}^* = \underset{\mathbf{w} \in \mathbb{R}^D}{\operatorname{argmax}} \, L(\mathbf{w}) - \frac{\lambda}{2} ||\mathbf{w}||^2$$
- In general there is no analytical solution to this optimization
- We use iterative techniques, i.e. gradient-based optimization
  1. Initialize $\mathbf{w} = \mathbf{0}$
  2. Take derivatives of $L(\mathbf{w}) - \frac{\lambda}{2}||\mathbf{w}||^2$, compute gradient
  3. Move $\mathbf{w}$ in steps proportional to the gradient
  4. Repeat steps 2 and 3 until convergence

Computing the gradient

$$
\begin{aligned}
\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}_j} &= \frac{1}{m}\sum_{k=1}^{m} \mathbf{f}_j(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}) \\
&\quad - \sum_{k=1}^{m} \sum_{\mathbf{y}\in\mathcal{Y}^*} P(\mathbf{y}|\mathbf{x}^{(k)}; \mathbf{w})\ \mathbf{f}_j(\mathbf{x}^{(k)}, \mathbf{y})
\end{aligned}
$$

where

$$
\mathbf{f}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \mathbf{f}_j(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i)
$$

- First term: observed mean feature value
- Second term: expected feature value under current $\mathbf{w}$

Computing the gradient

- The first term is easy to compute, by counting explicitly

$$\frac{1}{m}\sum_{k=1}^{m}\sum_{i}\mathbf{f}_j(\mathbf{x}, i, \mathbf{y}_{i-1}^{(k)}, \mathbf{y}_i^{(k)})$$

- The second term is more involved,

$$\sum_{k=1}^{m}\sum_{\mathbf{y}\in\mathcal{Y}^*} P(\mathbf{y}|\mathbf{x}^{(k)};\mathbf{w})\sum_{i}\mathbf{f}_j(\mathbf{x}^{(k)}, i, \mathbf{y}_{i-1}, \mathbf{y}_i)$$

because it sums over all sequences $\mathbf{y} \in \mathcal{Y}^*$

Computing the gradient

- For an example $(\mathbf{x}^{(k)}, \mathbf{y}^{(k)})$:

$$\sum_{\mathbf{y} \in \mathcal{Y}^n} P(\mathbf{y}|\mathbf{x}^{(k)}; \mathbf{w}) \sum_{i=1}^{n} \mathbf{f}_j(\mathbf{x}^{(k)}, i, \mathbf{y}_{i-1}, \mathbf{y}_i) =$$

$$\sum_{i=1}^{n} \sum_{a,b \in \mathcal{Y}} \mu_i^k(a,b) \mathbf{f}_j(\mathbf{x}^{(k)}, i, a, b)$$

where

$$\mu_i^k(a,b) = \sum_{\mathbf{y} \in \mathcal{Y}^n \;:\; \mathbf{y}_{i-1}=a, \; \mathbf{y}_i=b} P(\mathbf{y}|\mathbf{x}^{(k)}; \mathbf{w})$$

- The quantities μ_i^k can be computed efficiently in $O(nL^2)$ using the forward-backward algorithm

Forward-Backward for CRFs

- Assume fixed $\mathbf{x}$. Calculate in $O(nL^2)$

$$\mu_i(a,b) = \sum_{\mathbf{y} \in \mathcal{Y}^n : \mathbf{y}_{i-1}=a, \mathbf{y}_i = b} P(\mathbf{y}|\mathbf{x};\mathbf{w}) \quad , \ 1 \le i \le n; \ a, b \in \mathcal{Y}$$

- Define (forward and backward quantities):

$$\begin{aligned} \alpha_i(a) &= \sum_{\mathbf{y} \in \mathcal{Y}^i : \mathbf{y}_i = a} \exp\left\{ \sum_{j=1}^{i} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, j, \mathbf{y}_{j-1}, \mathbf{y}_j) \right\} \\ \beta_i(b) &= \sum_{\mathbf{y} \in \mathcal{Y}^{(n-i+1)} : \mathbf{y}_1 = b} \exp\left\{ \sum_{j=2}^{n-i+1} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, i+j-1, \mathbf{y}_{j-1}, \mathbf{y}_j) \right\} \end{aligned}$$

- Compute recursively $\alpha_i(a)$ and $\beta_i(b)$ (similar to Viterbi)
- $Z = \sum_a \alpha_n(a)$
- $\mu_i(a,b) = \{\alpha_{i-1}(a) * \exp\{\mathbf{w} \cdot \mathbf{f}(\mathbf{x}, i, a, b)\} * \beta_i(b) * Z^{-1}\}$

Compute the probability of a label sequence

$$P(\mathbf{y}|\mathbf{x}, \mathbf{w}) = \frac{1}{Z(\mathbf{x}; \mathbf{w})} \exp\left\{\sum_i \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i)\right\}$$

where

$$Z(\mathbf{x}; \mathbf{w}) = \sum_{\mathbf{z} \in \mathcal{Y}^n} \exp\left\{\sum_i \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, i, \mathbf{z}_{i-1}, \mathbf{z}_i)\right\}$$

- Compute $Z(\mathbf{x}; \mathbf{w})$ efficiently, using the forward algorithm

CRFs: summary so far

- Log-linear models for sequence prediction, $P(\mathbf{y}|\mathbf{x};\mathbf{w})$
- Computations factorize on label bigrams
- Model form:

$$\operatorname*{argmax}_{\mathbf{y}\in\mathcal{Y}^*} \sum_i \mathbf{w}\cdot\mathbf{f}(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i)$$

- Decoding: uses Viterbi (from HMMs)
- Parameter estimation:
  - Gradient-based methods, in practice L-BFGS
  - Computation of gradient uses forward-backward (from HMMs)

CRFs: summary so far

- Log-linear models for sequence prediction, $P(\mathbf{y}|\mathbf{x};\mathbf{w})$
- Computations factorize on label bigrams
- Model form:

$$\underset{\mathbf{y}\in\mathcal{Y}^*}{\operatorname{argmax}} \sum_i \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i)$$

- Decoding: uses Viterbi (from HMMs)
- Parameter estimation:
  - Gradient-based methods, in practice L-BFGS
  - Computation of gradient uses forward-backward (from HMMs)
- Next Question: HMMs or CRFs?

HMMs for sequence prediction

- $\mathbf{x}$ are the observations, $\mathbf{y}$ are the (un)hidden states
- HMMs model the joint distributon $P(\mathbf{x}, \mathbf{y})$
- Parameters: (assume $\mathcal{X} = \{1, \ldots, k\}$ and $\mathcal{Y} = \{1, \ldots, l\}$)
  - $\pi \in \mathbb{R}^l$, $\pi_a = \Pr(\mathbf{y}_1 = a)$
  - $T \in \mathbb{R}^{l \times l}$, $T_{a,b} = \Pr(\mathbf{y}_i = b | \mathbf{y}_{i-1} = a)$
  - $O \in \mathbb{R}^{l \times k}$, $O_{a,c} = \Pr(\mathbf{x}_i = c | \mathbf{y}_i = a)$
- Model form

$$P(\mathbf{x}, \mathbf{y}) = \pi_{\mathbf{y}_1} O_{\mathbf{y}_1, \mathbf{x}_1} \prod_{i=2}^{n} T_{\mathbf{y}_{i-1}, \mathbf{y}_i} O_{\mathbf{y}_i, \mathbf{x}_i}$$

- Parameter Estimation: maximum likelihood by counting events and normalizing

HMMs and CRFs

- In CRFs: $\hat{\mathbf{y}} = \text{amax}_{\mathbf{y}} \sum_i \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i)$
- In HMMs:
  $$\begin{aligned} \hat{\mathbf{y}} &= \text{amax}_{\mathbf{y}}\, \pi_{\mathbf{y}_1} O_{\mathbf{y}_1,\mathbf{x}_1} \prod_{i=2}^{n} T_{\mathbf{y}_{i-1},\mathbf{y}_i} O_{\mathbf{y}_i,\mathbf{x}_i} \\ &= \text{amax}_{\mathbf{y}} \log(\pi_{\mathbf{y}_1} O_{\mathbf{y}_1,\mathbf{x}_1}) + \sum_{i=2}^{n} \log(T_{\mathbf{y}_{i-1},\mathbf{y}_i} O_{\mathbf{y}_i,\mathbf{x}_i}) \end{aligned}$$
- An HMM can be ported into a CRF by setting:

$\mathbf{f}_j(\mathbf{x}, i, y, y')$	$\mathbf{w}_j$
$i = 1 \;\&\; y' = a$	$\log(\pi_a)$
$i > 1 \;\&\; y = a \;\&\; y' = b$	$\log(T_{a,b})$
$y' = a \;\&\; \mathbf{x}_i = c$	$\log(O_{a,b})$

- Hence, HMM parameters $\subset$ CRF parameters

HMMs and CRFs: main differences

- Representation:
  - HMM "features" are tied to the generative process.
  - CRF features are **very** flexible. They can look at the whole input $\mathbf{x}$ paired with a label bigram (y, y').
  - In practice, for prediction tasks, "good" discriminative features can improve accuracy **a lot**.
- Parameter estimation:
  - HMMs focus on explaining the data, both $\mathbf{x}$ and $\mathbf{y}$.
  - CRFs focus on the mapping from $\mathbf{x}$ to $\mathbf{y}$.
  - A priori, it is hard to say which paradigm is better.
  - Same dilemma as Naive Bayes vs. Maximum Entropy.

Structured Prediction

Perceptron, SVMs, CRFs

Learning Structured Predictors

- Goal: given training data
$$\{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), (\mathbf{x}^{(2)}, \mathbf{y}^{(2)}), \ldots, (\mathbf{x}^{(m)}, \mathbf{y}^{(m)})\}$$
learn a predictor $\mathbf{x} \to \mathbf{y}$ with small error on unseen inputs

- In a CRF:
$$\begin{aligned} \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}^*} P(\mathbf{y}|\mathbf{x}; \mathbf{w}) &= \frac{\exp\left\{\sum_{i=1}^{n} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i)\right\}}{Z(\mathbf{x}; \mathbf{w})} \\ &= \sum_{i=1}^{n} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i) \end{aligned}$$
  - To predict new values, $Z(\mathbf{x}; \mathbf{w})$ is not relevant
  - Parameter estimation: $\mathbf{w}$ is set to maximize likelihood

- Can we learn $\mathbf{w}$ more directly, focusing on errors?

Learning Structured Predictors

- Goal: given training data
$$\left\{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), (\mathbf{x}^{(2)}, \mathbf{y}^{(2)}), \ldots, (\mathbf{x}^{(m)}, \mathbf{y}^{(m)})\right\}$$
learn a predictor $\mathbf{x} \rightarrow \mathbf{y}$ with small error on unseen inputs

- In a CRF:
$$\begin{aligned} \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}^*} P(\mathbf{y}|\mathbf{x}; \mathbf{w}) &= \frac{\exp\left\{\sum_{i=1}^{n} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i)\right\}}{Z(\mathbf{x}; \mathbf{w})} \\ &= \sum_{i=1}^{n} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i) \end{aligned}$$
  - To predict new values, $Z(\mathbf{x}; \mathbf{w})$ is not relevant
  - Parameter estimation: $\mathbf{w}$ is set to maximize likelihood

- Can we learn $\mathbf{w}$ more directly, focusing on errors?

The Structured Perceptron

(Collins, 2002)

- Set $\mathbf{w} = \mathbf{0}$
- For $t = 1 \ldots T$
  - For each training example $(\mathbf{x}, \mathbf{y})$
    1. Compute $\mathbf{z} = \text{argmax}_{\mathbf{z}} \sum_{i=1}^{n} \mathbf{f}(\mathbf{x}, i, \mathbf{z}_{i-1}, \mathbf{z}_i)$
    2. If $\mathbf{z} \neq \mathbf{y}$

$$\mathbf{w} \leftarrow \mathbf{w} + \sum_i \mathbf{f}(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i) - \sum_i \mathbf{f}(\mathbf{x}, i, \mathbf{z}_{i-1}, \mathbf{z}_i)$$

- Return $\mathbf{w}$

The Structured Perceptron + Averaging

(Freund and Schapire, 1998)

- Set $\mathbf{w} = \mathbf{0}$, $\mathbf{w}^a = \mathbf{0}$
- For $t = 1 \ldots T$
  - For each training example $(\mathbf{x}, \mathbf{y})$
    1. Compute $\mathbf{z} = \mathrm{argmax}_{\mathbf{z}} \sum_{i=1}^{n} \mathbf{f}(\mathbf{x}, i, \mathbf{z}_{i-1}, \mathbf{z}_i)$
    2. If $\mathbf{z} \neq \mathbf{y}$

$$\mathbf{w} \leftarrow \mathbf{w} + \sum_i \mathbf{f}(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i) - \sum_i \mathbf{f}(\mathbf{x}, i, \mathbf{z}_{i-1}, \mathbf{z}_i)$$

    3. $\mathbf{w^a} = \mathbf{w^a} + \mathbf{w}$
- Return $\mathbf{w^a}/NT$, where N is the number of training examples

Properties of the Perceptron

- Online algorithm. Often much more efficient than "batch" algorithms
- If the data is separable, it will converge to parameter values with 0 errors
- Number of errors before convergence is related to a definition of *margin*. Can also relate margin to generalization properties
- In practice:
  1. Averaging improves performance **a lot**
  2. Typically reaches a good solution after only a few (say 5) iterations over the training set
  3. Often performs nearly as well as CRFs, or SVMs

Averaged Perceptron Convergence

Iteration	Accuracy
1	90.79
2	91.20
3	91.32
4	91.47
5	91.58
6	91.78
7	91.76
8	91.82
9	91.88
10	91.91
11	91.92
12	91.96
...	

(results on validation set for a parsing task)

Margin-based Structured Prediction

- Let $\mathbf{f}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \mathbf{f}(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i)$

- Model: $\mathrm{argmax}_{\mathbf{y} \in \mathcal{Y}^*} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, \mathbf{y})$

- Consider an example $(\mathbf{x}^{(k)}, \mathbf{y}^{(k)})$:
  $\exists \mathbf{y} \neq \mathbf{y}^{(k)} \;:\; \mathbf{w} \cdot \mathbf{f}(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}) < \mathbf{w} \cdot \mathbf{f}(\mathbf{x}^{(k)}, \mathbf{y}) \Longrightarrow$ error

- Let $\mathbf{y}' = \mathrm{argmax}_{\mathbf{y} \in \mathcal{Y}^* : \mathbf{y} \neq \mathbf{y}^{(k)}} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}^{(k)}, \mathbf{y})$
  Define $\gamma_k = \mathbf{w} \cdot (\mathbf{f}(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}) - \mathbf{f}(\mathbf{x}^{(k)}, \mathbf{y}'))$

- The quantity γ_k is a notion of margin on example k:
  $\gamma_k > 0 \Longleftrightarrow$ no mistakes in the example
  high $\gamma_k \Longleftrightarrow$ high confidence

Margin-based Structured Prediction

- Let $\mathbf{f}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \mathbf{f}(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i)$

- Model: $\operatorname{argmax}_{\mathbf{y} \in \mathcal{Y}^*} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, \mathbf{y})$

- Consider an example $(\mathbf{x}^{(k)}, \mathbf{y}^{(k)})$:
  $\exists \mathbf{y} \neq \mathbf{y}^{(k)} \ : \ \mathbf{w} \cdot \mathbf{f}(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}) < \mathbf{w} \cdot \mathbf{f}(\mathbf{x}^{(k)}, \mathbf{y}) \Longrightarrow$ error

- Let $\mathbf{y}' = \operatorname{argmax}_{\mathbf{y} \in \mathcal{Y}^* : \mathbf{y} \neq \mathbf{y}^{(k)}} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}^{(k)}, \mathbf{y})$
  Define $\gamma_k = \mathbf{w} \cdot (\mathbf{f}(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}) - \mathbf{f}(\mathbf{x}^{(k)}, \mathbf{y}'))$

- The quantity γ_k is a notion of margin on example k:
  $\gamma_k > 0 \Longleftrightarrow$ no mistakes in the example
  high $\gamma_k \Longleftrightarrow$ high confidence

Margin-based Structured Prediction

- Let $\mathbf{f}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \mathbf{f}(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i)$

- Model: $\text{argmax}_{\mathbf{y} \in \mathcal{Y}^*} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, \mathbf{y})$

- Consider an example $(\mathbf{x}^{(k)}, \mathbf{y}^{(k)})$:
  $\exists \mathbf{y} \neq \mathbf{y}^{(k)} \;\; : \;\; \mathbf{w} \cdot \mathbf{f}(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}) < \mathbf{w} \cdot \mathbf{f}(\mathbf{x}^{(k)}, \mathbf{y}) \Longrightarrow$ error

- Let $\mathbf{y}' = \text{argmax}_{\mathbf{y} \in \mathcal{Y}^* : \mathbf{y} \neq \mathbf{y}^{(k)}} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}^{(k)}, \mathbf{y})$
  Define $\gamma_k = \mathbf{w} \cdot (\mathbf{f}(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}) - \mathbf{f}(\mathbf{x}^{(k)}, \mathbf{y}'))$

- The quantity γ_k is a notion of margin on example k:
  $\gamma_k > 0 \iff$ no mistakes in the example
  high $\gamma_k \iff$ high confidence

Mistake-augmented Margins

(Taskar et al, 2004)

$\mathbf{x}^{(k)}$	Jack	London	went	to	Paris
$\mathbf{y}^{(k)}$	PER	PER	-	-	LOC
$\mathbf{y}'$	PER	LOC	-	-	LOC
$\mathbf{y}''$	PER	-	-	-	-
$\mathbf{y}'''$	-	-	PER	PER	-

- Def: $e(\mathbf{y}, \mathbf{y}') = \sum_{i=1}^{n} [\mathbf{y}_i \neq \mathbf{y}'_i]$
  e.g., $e(\mathbf{y}^{(k)}, \mathbf{y}^{(k)}) = 0$, $e(\mathbf{y}^{(k)}, \mathbf{y}') = 1$, $e(\mathbf{y}^{(k)}, \mathbf{y}''') = 5$
- Def: $\gamma_{k,\mathbf{y}} = \mathbf{w} \cdot (\mathbf{f}(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}) - \mathbf{f}(\mathbf{x}^{(k)}, \mathbf{y})) - e(\mathbf{y}^{(k)}, \mathbf{y})$
- Def: $\gamma_k = \min_{\mathbf{y} \neq \mathbf{y}^{(k)}} \gamma_{k,\mathbf{y}}$

Structured Hinge Loss

- Define loss function on example k as:

$$L(\mathbf{w}, \mathbf{x}^{(k)}, \mathbf{y}^{(k)}) = \max_{\mathbf{y} \in \mathcal{Y}^*} \Big(e(\mathbf{y}^{(k)}, \mathbf{y}) - \mathbf{w} \cdot (\mathbf{f}(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}) - \mathbf{f}(\mathbf{x}^{(k)}, \mathbf{y})) \Big)$$

- Leads to an SVM for structured prediction
- Given a training set, find:

$$\operatorname*{argmin}_{\mathbf{w} \in \mathbb{R}^D} \sum_{k=1}^{m} L(\mathbf{w}, \mathbf{x}^{(k)}, \mathbf{y}^{(k)}) + \frac{\lambda}{2} \|\mathbf{w}\|^2$$

Regularized Loss Minimization

- Given a training set $\left\{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \ldots, (\mathbf{x}^{(m)}, \mathbf{y}^{(m)})\right\}$. Find:

$$\operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^D} \sum_{k=1}^{m} L(\mathbf{w}, \mathbf{x}^{(k)}, \mathbf{y}^{(k)}) + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

- Two common loss functions $L(\mathbf{w}, \mathbf{x}^{(k)}, \mathbf{y}^{(k)})$:
  - Log-likelihood loss (CRFs)

$$-\log P(\mathbf{y}^{(k)} \mid \mathbf{x}^{(k)}; \mathbf{w})$$

  - Hinge loss (SVMs)

$$\max_{\mathbf{y}\in\mathcal{Y}^*} \Big(e(\mathbf{y}^{(k)}, \mathbf{y}) - \mathbf{w} \cdot (\mathbf{f}(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}) - \mathbf{f}(\mathbf{x}^{(k)}, \mathbf{y}))\Big)$$

Learning Structure Predictors: summary so far

- Linear models for sequence prediction

$$\underset{\mathbf{y} \in \mathcal{Y}^*}{\operatorname{argmax}} \sum_i \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i)$$

- Computations factorize on label bigrams
  - Decoding: using Viterbi
  - Marginals: using forward-backward
- Parameter estimation:
  - Perceptron, Log-likelihood, SVMs
  - Extensions from classification to the structured case
  - Optimization methods:
    - Stochastic (sub)gradient methods (LeCun et al 98) (Shalev-Shwartz et al. 07)
    - Exponentiated Gradient (Collins et al 08)
    - SVM Struct (Tsochantaridis et al. 04)
    - Structured MIRA (McDonald et al 05)

Structure Prediction

abstractions

Sequence Prediction, Beyond Bigrams

- It is easy to extend the scope of features to k-grams

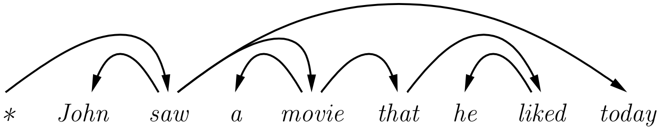
$$\mathbf{f}(\mathbf{x}, i, \mathbf{y}_{i-k+1:i-1}, \mathbf{y}_i)$$

- In general, think of state σ_i remembering relevant history
  - $\sigma_i = \mathbf{y}_{i-1}$ for bigrams
  - $\sigma_i = \mathbf{y}_{i-k+1:i-1}$ for k-grams
  - σ_i can be the state at time i of a deterministic automaton generating $\mathbf{y}$
- The structured predictor is

$$\operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}^*} \sum_i \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, i, \sigma_i, \mathbf{y}_i)$$

- Viterbi and forward-backward extend naturally, in $O(nL^k)$

Dependency Structures



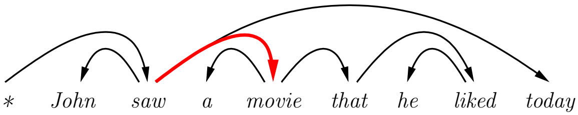
- Directed arcs represent dependencies between a head word and a modifier word.

- E.g.:

  movie *modifies* saw,
  John *modifies* saw,
  today *modifies* saw

Dependency Parsing: arc-factored models

(McDonald et al. 2005)



- Parse trees decompose into single dependencies $\langle h, m \rangle$

$$\underset{\mathbf{y} \in \mathcal{Y}(\mathbf{x})}{\operatorname{argmax}} \sum_{\langle h,m \rangle \in y} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, h, m)$$

- Some features: $\mathbf{f}_1(\mathbf{x}, h, m) = [\ \text{"saw"} \ \rightarrow \text{"movie"}\ ]$
  $\mathbf{f}_2(\mathbf{x}, h, m) = [\ \text{distance} = +2\ ]$
- Tractable inference algorithms exist (tomorrow's lecture)

Linear Structured Prediction

- Sequence prediction (bigram factorization)

$$\operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} \sum_{i} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, i, \mathbf{y}_{i-1}, \mathbf{y}_i)$$

- Dependency parsing (arc-factored)

$$\operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} \sum_{\langle h, m \rangle \in y} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, h, m)$$

- In general, we can enumerate parts $r \in \mathbf{y}$

$$\operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} \sum_{r \in \mathbf{y}} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, r)$$

Linear Structured Prediction Framework

- Abstract models of structures
  - Input domain $\mathcal{X}$, output domain $\mathcal{Y}$
  - A choice of factorization, $r \in \mathbf{y}$
  - Features: $\mathbf{f}(\mathbf{x}, r) \rightarrow \mathbb{R}^D$
- The linear prediction model, with $\mathbf{w} \in \mathbb{R}^D$

$$\operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} \sum_{r \in y} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, r)$$

- Generic algorithms for Perceptron, CRF, SVM
  - Require tractable inference algorithms