Multiview Conditional Random Fields for Phone Recognition

Corbin Rosset, Xun Hu

Johns Hopkins University, Department of Computer Science

April 21st, 2016

Recognition in Time Series Data

Sequence Models

Given T time ordered examples from some stationary distribution $\{(x_t, y_t)\}_1^T$, predict the most likely label sequence y on test data. Because data is sparse, we must rely heavily on context

Generative Models

- e.g. HMMs model the joint $p(x, y)$ as $\prod_{t=1}^T p(y_t|y_{t-1})p(x_t|y_t)$
- Make two (crippling) independence assertions: $\forall y_k \notin nb(y_t)$, $y_t \perp\!\!\!\perp y_k|y_{nb}$ and $x_t \perp\!\!\!\perp x_k, y_k|y_t$.
- $p(y, x) = p(y)p(x|y)$ shows how to "generate" features x from a label y
- Bayes' theorem: given the true $p^*(x|y)$, compute exactly $p(y|x)$. But HMMs only model $p(x_t|y_t)$, a far cry from $p(x|y)$. We want richer features...

---

[1] $nb(y_t)$ is the neighbors of y_t

Recognition in Time Series Data

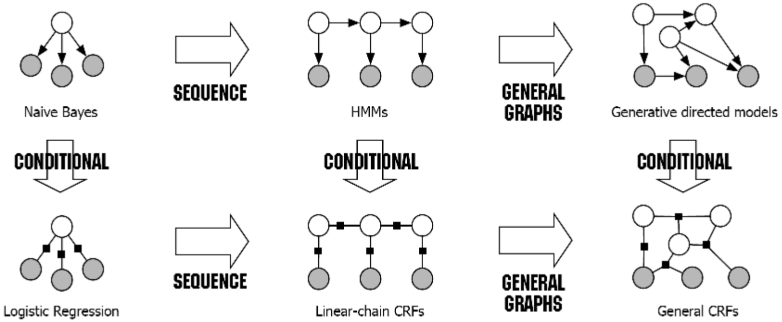
Log-linear Models

Logistic Regression models $p(y_i|x_i)$ by assuming $log(p(y|x))$ is linear in the features x: $p(y_i|x_i) = \frac{1}{Z(x)} exp\{b + \theta \cdot x\}$. Can also use a nonlinear feature function $f(y, x) \in \mathbb{R}^k$

Discriminative Models

- Don't bother wasting parameters to model $p(x)$, we only care about $p(y|x)$. But, the hypothesis space is the same
- More relaxed independence assertions: $\forall y_k \notin nb(y_t)$: $y_t \perp\!\!\!\perp y_k | y_{nb}, x$
- Express $p(y|x)$ as a log linear model, but now, feature functions can be derived from the whole input x
- If $p(y|x)$ factorizes according to some graphical model G, with small max cliques, we can do inference and parameter estimation efficiently

General CRFs



If $F = \{\Psi_a\}$ is the set of factors in G, then the conditional distribution for a CRF is $p(y|x) = \frac{1}{Z(x)} \prod_a \Psi_a(y_a, x_a)$ where y_a and x_a are the sets of variables in y and x that belong to clique Ψ_a. If we express Ψ_a in log-linear form...

Linear Chain CRF

Parameter Tying over Time

In general CRFs, each Ψ_a can have its own parameters θ_a and feature function $f_a(y_a, x_a)$. If our cliques tessellate through time, we can share parameters.

And throwing in the Markov property for linear chains:

$$p(y|x) = \frac{1}{Z(x)} \prod_a \Psi_a(y_a, x_a) \rightarrow \frac{1}{Z(x)} \prod_t^T \Psi_t(y_t, y_{t-1}, x_t) \qquad (1)$$

Force clique potential to be in the exponential family:

$$\Psi_a = exp\{\theta_a \cdot f_a(y_a, x_a)\}$$

Notice Z only depends on x and is more easily computed. However, Z is summed over all possible label sequences (use forward backward)

Linear Chain CRF

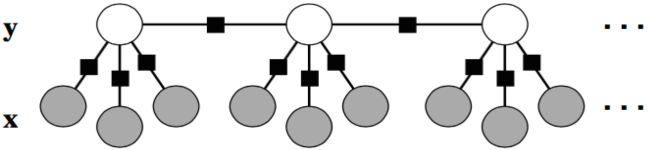
$p(y|x)$ can now be written as:

Definition 2.2. Let Y, X be random vectors, $\theta = \{\theta_k\} \in \Re^K$ be a parameter vector, and $\mathcal{F} = \{f_k(y, y', \mathbf{x}_t)\}_{k=1}^K$ be a set of real-valued feature functions. Then a *linear-chain conditional random field* is a distribution $p(\mathbf{y}|\mathbf{x})$ that takes the form:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{t=1}^{T} \exp\left\{ \sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right\}, \qquad (2.18)$$

where $Z(\mathbf{x})$ is an input-dependent normalization function

$$Z(\mathbf{x}) = \sum_{\mathbf{y}} \prod_{t=1}^{T} \exp\left\{ \sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right\}. \qquad (2.19)$$



Linear Chain CRF Tricks

Convenience of parameter tying

If each y_t can take on k values, then at each t we can define a matrix $M_t = \mathbb{R}^{k \times k}$ such that $M_t(y_i, y_j|x) = exp\{\theta_{ij} \cdot f(y_i, y_j, x)\}$

- Then $Z(x) = \prod_{t=1}^{T} M_i$
- and the probability of the label sequence becomes $p(y|x) = \frac{\prod_{t=1}^{T} M_i(y_{t-1}, y_t|x)}{Z(x)}$

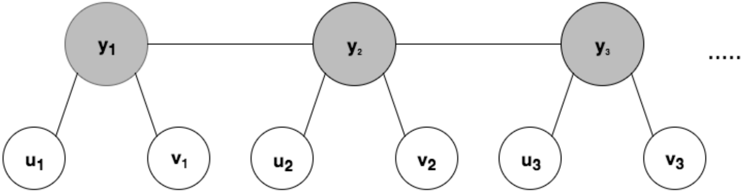
Sparsity in label space

Not every sequence of consecutive labels y_i, y_k may be valid, so the $O(k^2)$-time updates for each forward and backward step can be reduced drastically.

Our Project

Multiple views of x

We now TWO sequences of observations $u = \{u_t\}_{t=1}^{T}$ and $v = \{v_t\}_{t=1}^{T}$ $u_i \in \mathbb{R}^{d_1}$, $v_i \in \mathbb{R}^{d_2}$ corresponding to articulatory and acoustic measurements of natural speech.



Conditional Model for Multiview CRFs

HMM-like model

- Labels y_t take on k values, use one-hot vectors to simulate indicator function
- Three parameters: $\theta \in \mathbb{R}^{k \times k}$ (transition), $\phi_1 \in \mathbb{R}^{k \times d_1}$ (emission view 1), $\phi_1 \in \mathbb{R}^{k \times d_2}$ (emission view 2)
- emission features can easily be broadened
- clique potential: $\Psi_t = exp\{{y_t}^T \theta y_{t-1} + {y_t}^T \phi_1 u_t + {y_t}^T \phi_2 v_t\}$
- $Z(u, v) = \sum_y \prod_{t=1}^T \Psi_t$

$$p(y|u, v) = \frac{1}{Z(u, v)} \prod_{t=1}^{T} exp\{{y_t}^T \theta y_{t-1} + {y_t}^T \phi_1 u_t + {y_t}^T \phi_2 v_t\} \quad (2)$$

Inference in Multiview CRFs

Three main inference tasks, almost identical to those of HMMs:

- edge marginals: $p(y_t, y_{t-1}|x;\theta)$[1] For marginals, use forward-backward message passing: $p(y_t, y_{t-1}|x;\theta) = \frac{1}{Z(x)}\alpha_{t-1}(y_t - 1)\Psi_t(y_t, y_{t-1}, x_t)\beta_t(t)$
- node marginals: $p(y_t|x;\theta) = \frac{1}{Z(x)}\alpha_t(y_t)\beta_t(y_t)$
- sequence labels: $y^* = argmax_y\ p(y|u,v;\theta)$. Use the Viterbi algorithm
- finding the N-best sequence labels may also be useful

[1] use θ as shorthand for $\{\theta;\phi_1;\phi_2\}$, and x as shorthand for $\{u,v\}$

Parameter Estimation in Multiview CRFs

Training setup

We have N labeled time sequences[2] for training,
$\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^{N}$ each not necessarily of length T.[3]

Regularized MLE

- Recall for a single training sequence,
  $p(y|u,v) = \frac{1}{Z(u,v)} \prod_{t=1}^{T} exp\{{y_t}^T \theta y_{t-1} + {y_t}^T \phi_1 u_t + {y_t}^T \phi_2 v_t\}$[4]
- Conditional log-likelihood:
  $\ell(\theta) = \sum_{i=1}^{N} logp(y^{(i)}|u^{(i)}, v^{(i)}) - \sum_{i=1}^{N} logZ(u^{(i)}, v^{(i)}) - R$
- Regularizer $R = \lambda_1 \|\theta\|_2 + \lambda_2 \|\phi_1\|_2 + \lambda_3 \|\phi_2\|_2$. Analogous to zero mean Gaussian prior.
- Convex!

$\ell(\theta)$ cannot be maximized in closed form, use SGD...

[2]drawn i.i.d from the same stationary distribution

[3]parameter tying allows sequences of arbitrary length

[4]u, v, and y are assumed to have (i) superscripts

Parameter Estimation in Multiview CRFs

SGD: pick a training sequence at random, do $\theta \leftarrow \theta + \alpha \cdot \nabla \ell(\theta)$

Stochastic Gradient Updates

- $\frac{\partial \ell}{\partial \theta} = \sum_t^T y_t \cdot y_{t-1}^T - \sum_t \sum_{y',y''} y' \cdot y''^T p(y', y''|x) - \frac{\lambda_1}{2N}\theta$
- $\frac{\partial \ell}{\partial \phi_1} = \sum_t^T y_t \cdot u_t^T - \sum_t \sum_{y'} y' \cdot u_t^T p(y'|u, v) - \frac{\lambda_2}{2N}\phi_1$
- $\frac{\partial \ell}{\partial \phi_2} = \sum_t^T y_t \cdot v_t^T - \sum_t \sum_{y'} y' \cdot v_t^T p(y'|u, v) - \frac{\lambda_3}{2N}\phi_2$

REMEMBER: All y_t are one-hot vectors, so $\sum_t^T y_t \cdot y_{t-1}$ is a $k \times k$ matrix of counts of all label transitions in the sequence!
And $\sum_{y'} y' \cdot u_t^T p(y'|u, v)$ is a $k \times d_1$ matrix of v_t repeated and scaled row-wise by $p(y'|u, v)$

Parameter Estimation in Multiview CRFs

Runtime of SGD

Notice ALL edge and node marginals $p(y', y''|u, v)$, $p(y'|u, v)$ are needed for every step. Forward-Backward is $O(TK^2)$ per training sequence. Likelihood and gradients calculations $O(TK^2)$ per step. So runtime is $O(TK^2G)$, G number of steps.

2nd order approximation methods like BFGS would be require fewer steps than SGD...

SGD in CRFs requires O(T) inferences per step

At every SGD step, all edge and node marginals need to be inferred. Even though linear chains amenable to exact inference, b/ it needs to be done so frequently, use faster approximations like MCMC or variational methods, and leverage sparsity [5]

[5] If we did batch gradient descent, we would need to do inference N times per step, completely intractable

Parameter Estimation in Multiview CRFs

Algorithm for parameter estimation

repeat
  choose $(u^{(i)}, v^{(i)}, y^{(i)})$ from training sequences at random
  $\forall(y', y'')$, compute and store $p(y', y''|u^{(i)}, v^{(i)})$ via inference
  $\forall(y')$, compute and store $p(y'|u^{(i)}, v^{(i)})$ via inference
  for all all parameters θ **do**
    $\theta \leftarrow \theta + \alpha \cdot \nabla \ell(\theta)$
  end for
until convergence
Infer y^* the Viterbi tagging on held-out test data

Extensions: Hidden Conditional Random Fields

Subphones

CRF augmented with hidden states that model mixture components m_t and subphones s_t. We don't need to know phone boundaries.[6]

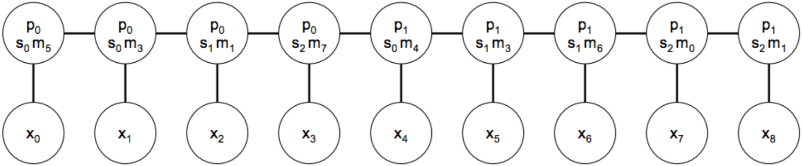


Figure 1: An instance of Viterbi labeling from an HCRF showing a phone sequence p_0, p_1 composed of a state sequence s together with mixture components m. s's and m's are hidden variables and must be marginalized out in learning and inference.

---

[6]see Sung and Jurafsky: https://web.stanford.edu/ jurafsky/asru09.pdf

Extensions: Multiview Hidden Conditional Random Fields

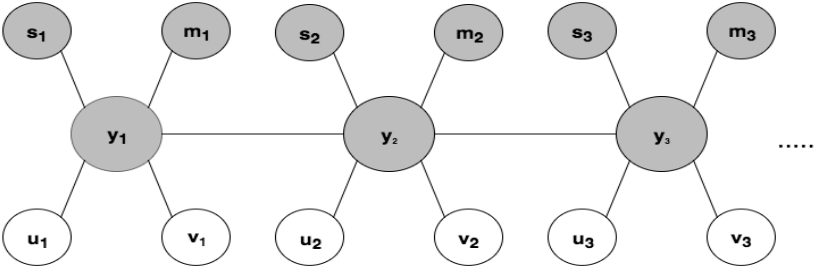


Figure 2: We propose augmenting HCRFs for phone recognition with multiview data: MFCC acoustic as well as articulatory data.

- Huang, Zhhiheng, Wei Xu, and Kai Yu. "Bidirectional LSTM-CRF Models for Sequence Tagging." ArXiv, n.d. Web. 21 Apr. 2016.
- Koller, Daphne, and Nir Friedman. Probabilistic Graphical Models: Principles and Techniques. Cambridge, MA: MIT, 2009. 684-94. Print.
- Lafferty, John and McCallum, Andrew, and Fernando C.N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", . June 2001.
- Sung, Yun-Hsuan, and Dan Jurafsky. "Hidden Conditional Random Fields for Phone Recognition." 2009 IEEE Workshop on Automatic Speech Recognition and Understanding (2009):
- Sutton, Charles. "An Introduction to Conditional Random Fields." FNT in Machine Learning Foundations and Trends® in Machine Learning 4.4 (2012): 267-373. Web.