Augmented Statistical Models: Exploiting Generative Models in Discriminative Classifiers

Martin Layton & Mark Gales

9 December 2005

Cambridge University Engineering Department

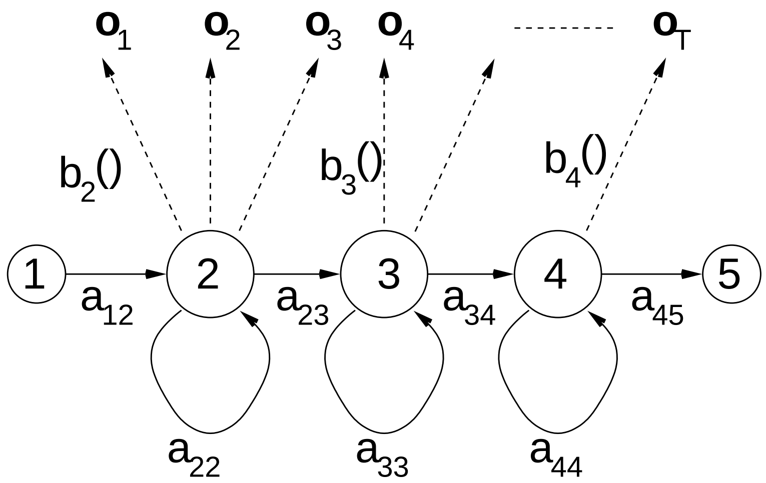
NIPS 2005

Overview

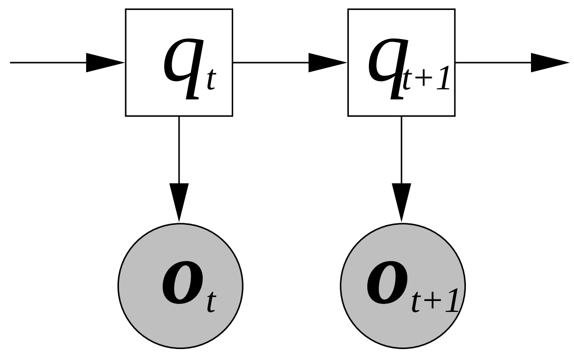
- Generative models in discriminative classifiers
  - Fisher score-space
  - Generative score-space
- Augmented Statistical Models
  - extension of standard models, e.g. GMMs and HMMs
  - allows additional dependencies to be represented
- Discriminative training
  - maximum margin
  - conditional maximum likelihood
- TIMIT results

Generative Models in Discriminative Classifiers

The Hidden Markov Model



(a) Standard HMM phone topology



(b) HMM Dynamic Bayesian Network

- Observations conditionally independent of other observations given state.
- States conditionally independent of other states given previous states.
- Poor model of the speech process - piecewise constant state-space.

Fisher Score-spaces

- Jaakkola & Haussler (1999)
- Method of incorporating generative models within a discriminative framework
- Define a base generative model $\hat{p}(\boldsymbol{O};\boldsymbol{\lambda})$
  - 1-dimensional log-likelihood
  - not enough information for good classification
- Instead use a score-space $\boldsymbol{\phi}^{\mathrm{F}}(\boldsymbol{O};\boldsymbol{\lambda})$
  - tangent-space captures essence of generative process

$$\boldsymbol{\phi}^{\mathrm{F}}(\boldsymbol{O};\boldsymbol{\lambda}) = \left[\nabla_{\boldsymbol{\lambda}} \ln \hat{p}(\boldsymbol{O};\boldsymbol{\lambda})\right]$$

  - dimensionality of score-space: `parameters` $\boldsymbol{\lambda}$
  - suitable for discriminative training (SVMs, etc)
  - has been applied to many tasks, e.g. comp. biology and speech recognition

Generative Score-spaces

- Smith & Gales (2002)
- Extension for supervised binary classification tasks
- Define class-conditional base models $\hat{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(1)})$ and $\hat{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(2)})$
  - includes log-likelihood ratio to improve discrimination
  - avoids wrap-around (different $\boldsymbol{O}$'s map to the same point in score-space)
- Score-space $\boldsymbol{\phi}^{\text{LL}}(\boldsymbol{O};\boldsymbol{\lambda})$

$$\boldsymbol{\phi}^{\text{LL}}(\boldsymbol{O};\boldsymbol{\lambda}) = \left[\begin{array}{c} \ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(1)}) - \ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(2)}) \\ \nabla_{\boldsymbol{\lambda}^{(1)}}\ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(1)}) \\ -\nabla_{\boldsymbol{\lambda}^{(2)}}\ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(2)}) \end{array}\right]$$

  - suitable for discriminative training — SVMs
  - no probabilistic interpretation
  - restricted to binary problems

Augmented Statistical Models

Dependency Modelling

- Speech data is dynamic — observations are not of a fixed length
- Dependency modelling essential part of speech recognition

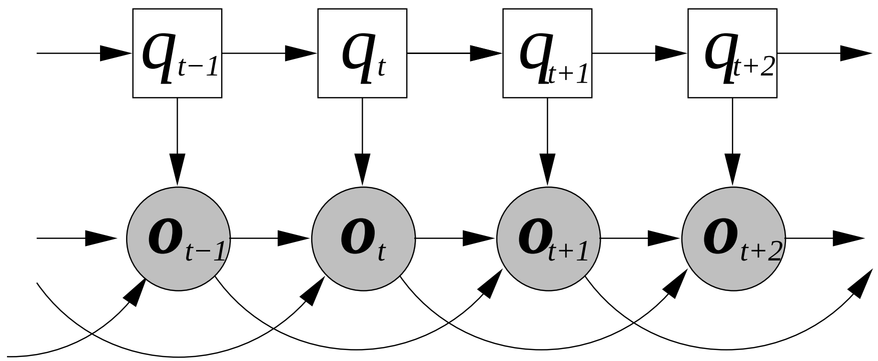
$$p(\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T; \boldsymbol{\lambda}) = p(\boldsymbol{o}_1; \boldsymbol{\lambda})p(\boldsymbol{o}_2|\boldsymbol{o}_1; \boldsymbol{\lambda}) \ldots p(\boldsymbol{o}_T|\boldsymbol{o}_1, \ldots, \boldsymbol{o}_{T-1}; \boldsymbol{\lambda})$$

  - impractical to directly model in this form
  - make extensive use of conditional independence

- Two possible forms of conditional independence
  - latent (unobserved) variables
  - observed variables

- Even if given a set of dependencies (form of Bayesian Network)
  - need to determine how dependencies interact

Dependency Modelling



- Commonly use a member (or mixture) of the exponential family

$$p(\boldsymbol{O};\boldsymbol{\alpha}) = \frac{1}{\tau(\boldsymbol{\alpha})} h(\boldsymbol{O}) \exp\left(\boldsymbol{\alpha}^T \boldsymbol{T}(\boldsymbol{O})\right)$$

$h(\boldsymbol{O})$ is the reference distribution
τ is the normalisation term
$\boldsymbol{\alpha}$ are the natural parameters
$\boldsymbol{T}(\boldsymbol{O})$ are sufficient statistics

- What is the appropriate form of statistics ($\boldsymbol{T}(\boldsymbol{O})$)?
  - for diagram above, $\boldsymbol{T}(\boldsymbol{O}) = \sum_{t=1}^{T-2} \boldsymbol{o}_t \boldsymbol{o}_{t+1} \boldsymbol{o}_{t+2}$

Augmented Statistical Models

- Augmented statistical models (related to fibre bundles)

$$p(\boldsymbol{O};\boldsymbol{\lambda},\boldsymbol{\alpha}) = \frac{1}{\tau(\boldsymbol{\lambda},\boldsymbol{\alpha})}\hat{p}(\boldsymbol{O};\boldsymbol{\lambda})\exp\left(\boldsymbol{\alpha}^T\left[\begin{array}{c}\nabla_{\boldsymbol{\lambda}}\ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda})\\ \frac{1}{2!}\text{vec}\left(\nabla^2_{\boldsymbol{\lambda}}\ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda})\right)\\ \vdots \\ \frac{1}{\rho!}\text{vec}\left(\nabla^\rho_{\boldsymbol{\lambda}}\ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda})\right)\end{array}\right]\right)$$

- Two sets of parameters:
  - $\boldsymbol{\lambda}$ - parameters of base distribution ($\hat{p}(\boldsymbol{O};\boldsymbol{\lambda})$)
  - $\boldsymbol{\alpha}$ - natural parameters of local exponential model

- Normalisation term $\tau(\boldsymbol{\lambda},\boldsymbol{\alpha})$ ensures valid PDF

$$\int p(\boldsymbol{O};\boldsymbol{\lambda},\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{O} = 1; \qquad p(\boldsymbol{O};\boldsymbol{\lambda},\boldsymbol{\alpha}) = \frac{\bar{p}(\boldsymbol{O};\boldsymbol{\lambda},\boldsymbol{\alpha})}{\tau(\boldsymbol{\lambda},\boldsymbol{\alpha})}$$

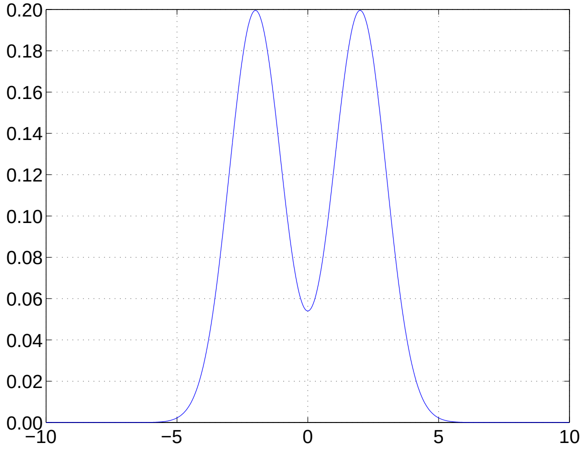
  - can be very difficult to estimate

Example: Augmented GMM

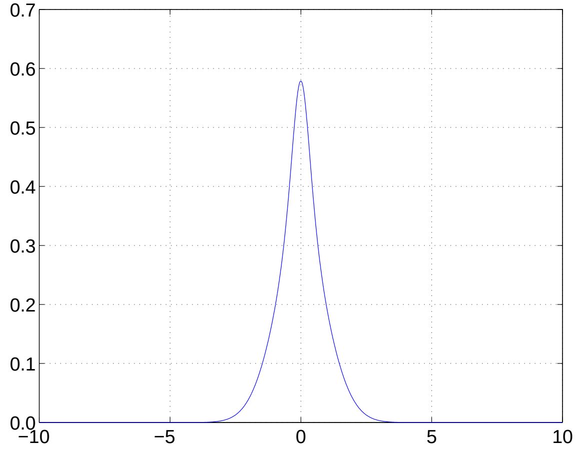
- Use a GMM as the base distribution: $\hat{p}(\boldsymbol{o};\boldsymbol{\lambda}) = \sum_{m=1}^{M} c_m \mathcal{N}(\boldsymbol{o};\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m)$

$$p(\boldsymbol{o};\boldsymbol{\lambda},\boldsymbol{\alpha}) = \frac{1}{\tau}\sum_{m=1}^{M} c_m \mathcal{N}(\boldsymbol{o};\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m)\exp\left(\sum_{n=1}^{M} P(n|\boldsymbol{o};\boldsymbol{\lambda})\boldsymbol{\alpha}_n^T\boldsymbol{\Sigma}_n^{-1}(\boldsymbol{o}-\boldsymbol{\mu}_n)\right)$$

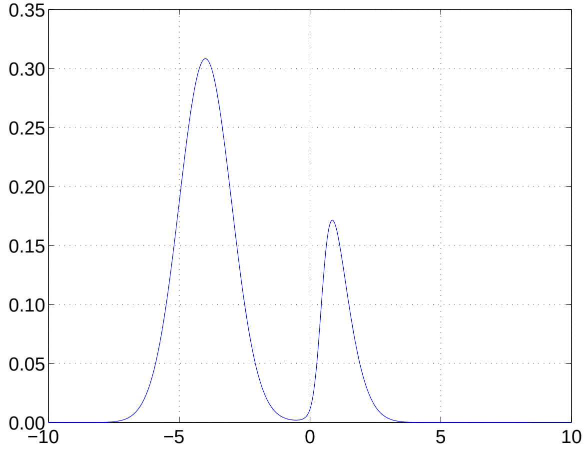
- Simple two component one-dimensional example:



$\boldsymbol{\alpha} = [0.0, 0.0]^T$



$\boldsymbol{\alpha} = [-1.0, -1.0]^T$



$\boldsymbol{\alpha} = [1.0, -1.0]^T$

Augmented Model Dependencies

- If the base distribution is a latent-variable model — GMM,HMM,...
  - Sufficient statistics contain a first-order differential

$$\nabla_{\boldsymbol{\mu}_{jm}} \ln \hat{p}(\boldsymbol{O};\boldsymbol{\lambda}) = \sum_{t=1}^{T} P(\theta_t = \{s_j, m\}|\boldsymbol{O};\boldsymbol{\lambda}) \boldsymbol{\Sigma}_{jm}^{-1}(\boldsymbol{O}_t - \boldsymbol{\mu}_{jm})$$

  - depends on a posterior
  - compact representation of effects of all observations

- Augmented models of this form:
  - retain independence assumptions of the base model
  - remove conditional independence assumptions of the base model...
    ... since the local exponential model depends on a posterior

- For HMM base models,
  - observations are dependent on all observations and all latent states
  - higher-order derivatives create increasingly powerful models

Discriminative Training

Maximum Margin Estimation

- Consider the simplified two-class problem
- Bayes' decision rule (consider $\boldsymbol{\lambda}$ fixed)

$$\frac{P(\omega_1|\boldsymbol{O})}{P(\omega_2|\boldsymbol{O})} = \frac{P(\omega_1)\,\tau(\boldsymbol{\lambda}^{(2)},\boldsymbol{\alpha}^{(2)})\,\bar{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(1)},\boldsymbol{\alpha}^{(1)})}{P(\omega_2)\tau(\boldsymbol{\lambda}^{(1)},\boldsymbol{\alpha}^{(1)})\,\bar{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(2)},\boldsymbol{\alpha}^{(2)})} \mathop{\gtrless}_{\omega_2}^{\omega_1} 1$$

  – class priors $P(\omega_1)$ and $P(\omega_2)$

- Can be rewritten as a linear decision boundary in a generative score-space,

$$\underbrace{\frac{1}{T}\ln\left(\frac{\bar{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(1)},\boldsymbol{\alpha}^{(1)})}{\bar{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(2)},\boldsymbol{\alpha}^{(2)})}\right)}_{\boldsymbol{w}^T\boldsymbol{\phi}^{\text{LL}}(\boldsymbol{O};\boldsymbol{\lambda})} + \underbrace{\frac{1}{T}\ln\left(\frac{P(\omega_1)\tau(\boldsymbol{\lambda}^{(2)},\boldsymbol{\alpha}^{(2)})}{P(\omega_2)\tau(\boldsymbol{\lambda}^{(1)},\boldsymbol{\alpha}^{(1)})}\right)}_{b} \mathop{\gtrless}_{\omega_2}^{\omega_1} 0$$

  – no need to explicitly calculate $\tau(\boldsymbol{\lambda}^{(1)},\boldsymbol{\alpha}^{(1)})$ or $\tau(\boldsymbol{\lambda}^{(2)},\boldsymbol{\alpha}^{(2)})$

- Note: restrictions on $\boldsymbol{\alpha}$'s required to ensure a valid PDF

Maximum Margin Estimation (cont.)

- First-order Generative score-space given by

$$\phi^{\text{LL}}(\boldsymbol{O};\boldsymbol{\lambda}) = \frac{1}{T}\left[\begin{array}{c} \ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(1)}) - \ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(2)}) \\ \nabla_{\boldsymbol{\lambda}^{(1)}}\ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(1)}) \\ -\nabla_{\boldsymbol{\lambda}^{(2)}}\ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(2)}) \end{array}\right]$$

  – independent of augmented parameters $\boldsymbol{\alpha}$

- Linear decision boundary specified by

$$\boldsymbol{w}^T = \left[\begin{array}{ccc} 1 & \boldsymbol{\alpha}^{(1)T} & \boldsymbol{\alpha}^{(2)T} \end{array}\right]^T$$

  – only a function of the exponential model parameters $\boldsymbol{\alpha}$

- Bias calculated as a by-product of training — depends on both $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$
- Potentially many parameters to estimate:
  – maximum margin estimation (MME) good choice — SVM training

Conditional Augmented Models

- Often impossible to calculate normalisation term for generative augmented models
  - restricted to binary tasks
  - cannot use direct training

- Instead, consider conditional augmented models

$$p(\omega_j|\boldsymbol{O};\boldsymbol{\lambda},\boldsymbol{\alpha}) = \frac{1}{Z(\boldsymbol{\lambda},\boldsymbol{\alpha})}\hat{p}(\boldsymbol{O};\boldsymbol{\lambda})\exp\left(\boldsymbol{\alpha}^T\begin{bmatrix}\nabla_{\boldsymbol{\lambda}}\ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda})\\ \frac{1}{2!}\text{vec}\left(\nabla^2_{\boldsymbol{\lambda}}\ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda})\right)\\ \vdots\\ \frac{1}{\rho!}\text{vec}\left(\nabla^\rho_{\boldsymbol{\lambda}}\ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda})\right)\end{bmatrix}\right)$$

  - directly model decision surfaces between classes
  - normalisation calculated as expectation over classes — easy to calculate

Conditional Maximum Likelihood Estimation

- Maximum likelihood of conditional model

$$\{\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\alpha}}\} = \underset{\boldsymbol{\lambda},\boldsymbol{\alpha}}{\operatorname{argmax}} \sum_{i=1}^{n} \ln P(y_i|\boldsymbol{O}_i; \boldsymbol{\lambda}, \boldsymbol{\alpha})$$

  - $\boldsymbol{O}_i$ are training examples; y_i are class labels
  - No closed-form solution

- Use stochastic gradient descent
  - use noisy estimates of conditional log-likelihood gradient

$$\nabla_{\boldsymbol{\alpha}} \ln P(y_i|\boldsymbol{O}_i; \boldsymbol{\lambda}, \boldsymbol{\alpha}) = \boldsymbol{T}(y_i, \boldsymbol{O}_i; \boldsymbol{\lambda}) - \sum_{\omega \in \Omega} p(\omega|\boldsymbol{O}_i; \boldsymbol{\lambda}, \boldsymbol{\alpha}) \boldsymbol{T}(\omega, \boldsymbol{O}_i; \boldsymbol{\lambda})$$

  - $\Omega = \{\omega_1, \ldots, \}$ is the set of all class labels
  - $\boldsymbol{T}(y_i, \boldsymbol{O}_i; \boldsymbol{\lambda})$ are the augmented model sufficient statistics
  - optimisation is convex

TIMIT Results

TIMIT

- Phone classification task
- Training
  - 462 speakers: 3,696 sentances
  - 48 possible phones (classes)
- Testing
  - 24 speakers: 192 sentances
  - 48 phones mapped to a 39-class set for scoring purposes
- Data encoded using standard features: `MFCC_0_D_A`
  - 3 emitting state HMMs with 10 or 20 mixture components
  - first-order score-space: means, variances and component priors

TIMIT

Classifier	Criterion		Components	
	λ	α	10	20
HMM	ML	–	29.4	27.3
C-Aug	ML	CML	25.6	–
HMM	MMI	–	25.3	24.8
C-Aug	MMI	CML	24.1	–

- Conditional augmented models outperform HMMs
  - given a base model, it is better to augment it instead of increasing the number of mixture components
- Maximum-margin outperforms Conditional MLE (results not shown)
  - restricted to binary tasks
  - partly due to CML overtraining — regularisation required

Summary

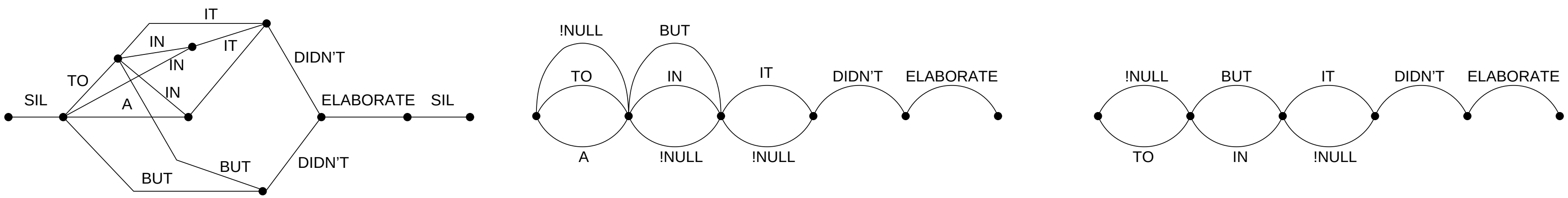
- Augmented statistical models
  - allow complex dependencies to be added in a systematic fashion
  - breaks conditional independence assumptions of base model
  - simple to train using MM or CML estimation
- Preliminary results positive
  - outperform ML and MMI HMMs with similar numbers of parameters
  - CML optimisation is simple and easy to extend...
- Current work
  - Regularisation of CML
  - Updates of base model $\boldsymbol{\lambda}$
  - Recognition

Extra Slides

Binary Classifiers and LVCSR

- Many classifiers (e.g. SVMs) are inherently binary:
  - speech recognition has a vast number of possible classes
  - how to map to a simple binary problem?

- Use pruned confusion networks (Venkataramani et al ASRU 2003):



Word lattice | Confusion Network | Pruned confusion network

  - use standard HMM decoder to generate word lattice
  - generate confusion networks (CN) from word lattice
    - gives posterior for each arc being correct;
  - prune CN to a maximum of two arcs (based on posteriors).

8-Fold Cross-Validation LVCSR Results

Word Pair (Examples/class)	Classifier	Training		WER (%)
		Base (λ)	Aug (α)	
CAN/CAN'T (3761)	HMM	ML	—	11.0
		MMI	—	10.4
	A-HMM	ML	MM	9.5
KNOW/NO (4475)	HMM	ML	—	27.7
		MMI	—	27.1
	A-HMM	ML	MM	23.8

- A-HMM outperforms both ML and MMI HMM
  - also outperforms using "equivalent" number of parameters
  - difficult to split dependency modelling gains from change in training criterion

Evaluation Data LVCSR Results

- Baseline performance using Viterbi and Confusion Network decoding

Decoding	trigram LM
Viterbi	30.8
Confusion Network	30.1

- Rescore word-pairs using 3-state/4-component A-HMM+βCN

SVM Rescoring	#corrected/#pairs	% corrected
10 SVMs	56/1250	4.5%

  - only 1.6% of 76157 hypothesised words rescored - more SVMs required!

- More suitable to smaller tasks, e.g. digit recognition in low SNR conditions