Probability-Oriented Neural Networks and Hybrid Connectionist/Stochastic Networks

as presented in the similarly called chapter 8 by Jean-Paul Haton

Vesa Siivola, February 28, 2002

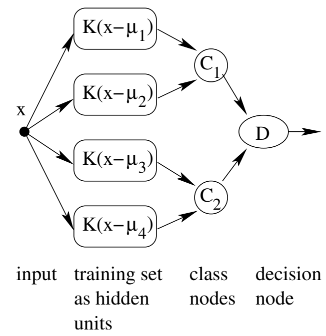
1. Types of Probability Oriented Networks

1.1 Probabilistic Neural Networks

- Nonparametric method, very much like Parzen window estimation
- All training samples are put in hidden layer and respond locally.

\+ Fast and incremental training. Asymptotically Bayesian.

\- Requires lot of storage space, all patterns equally important



1.1.1 Probabilistic neural network variants:

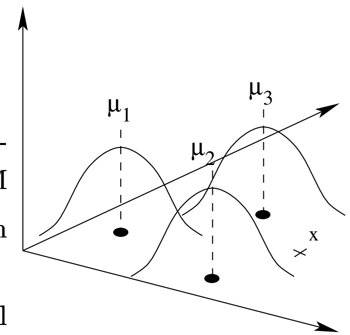
- Reduce the number of training samples – find representative prototypes (Reduced Kernel Discriminant Analysis, SOM + LVQ)
- Different smoothing along different dimensions: adaptive PNN
- Generalized Fisher training PNN, Gaussian clustering network, probabilistic mapping network

1.2 Radial Basis Function Networks

- Very much like Gaussian Mixture Models
- Instead of all training samples uses prototypes. These can be initialize for ex. by SOM
- Discriminative training: training criterion MSE between actual and desired outputs
- Localized response can be justified from real neural systems (Cochlear cells).

\+ Fast training (few orders of magnitude faster than MLP), good generalization

\- Needs lot of training data

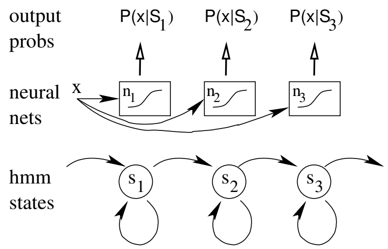


$$\begin{bmatrix} P(x|C_1) \\ P(x|C_2) \\ \vdots \\ P(x|C_N) \end{bmatrix} = f(\mathbf{A} \begin{bmatrix} K(\mu_1 - x) \\ K(\mu_2 - x) \\ \vdots \\ K(\mu_M - x) \end{bmatrix})$$

2. Applications

- Networks static - only very simple speech recognition applications. Used more in speaker recognition tasks.
- Hybrid systems used in speech recognition

2.1. Hybrids - Neural Networks as preprocessors for Hidden Markov Models



Methods:

- Use NN's as vector quantizer: Map observations to phonemes or other set of discrete values. Calculate HMM emission probabilities for these.
- Estimate state posteriori probabilities for each frame by NNs (MLP, time delay MLP, recurrent MLP, RBF, PNN). Use HMMs to tie these to a coherent whole.

A closer look taken at the end of this presentation: Case study 2.

2.2. Hybrids - Neural Networks as postprocessors for Hidden Markov Models

Methods:

- Rescore results from HMM (provides segmentation) with NN
- Use NN to distinguish between highly confusable units (/p/,/b/) (also requires segmentation from HMM).
- Word spotting: Use a “loose” HMM spotter (accepts a match easily) and a neural network to reject false hits from HMM.

A closer look later at this presentation: Case study 1.

2.3. Hybrids - Unified models

Methods

- Implement algorithms (Viterbi, Baum-Welch) as neural networks (recurrent MLP)
- Alphanets: Extended recurrent MLPs performing exactly same operations as HMMs

3.1. Case study 1: Segmental Neural Nets in BBN BYBLOS system

(see references [89,90] in ch. 8)

- Hybrid MLP / HMM system (or hybrid generalized RBF / HMM system)
- Addresses problems associated with HMMs
  - conditional independence assumption
  - poor representation for duration and other segmental features

3.1.1 Structure

- One net for each phoneme
- Let's assume that the segmentation is known
- Warp each segment to span 5 frames or use Discrete Cosine transform and trunctuate the results
- duration info in a separate duration model, which is a smoothed histogram of durations

- segment score = net score + duration score
- 80 features (16 features/frame $\times$ 5 frames)
- No efficient search algorithm for all word sequences and segmentations. Can not be used as such.

3.1.2 N-best rescoring

- HMMs can give N-best results, which are usually given to the language model, which rescores the hypotheses.
- Here, 20 best hypotheses are given to Segmental Neural Net.
- Sentence is rescored. score = HMM score + segment score (net+duration) + grammar score + hypothesized number of words and phonemes
- Can be trained with MSE or even discriminatively (best vs negative training for the rest in n-best list)

3.2. Case study 1: Neural nets as HMM state emission probability estimators in Decipher

(Renals, Morgan, Bourlard, Cohen, Franco: "Connectionist Probability Estimators in HMM Speech Recognition", IEEE transactions on speech and audio processing, 2.1.2, Jan 1994)

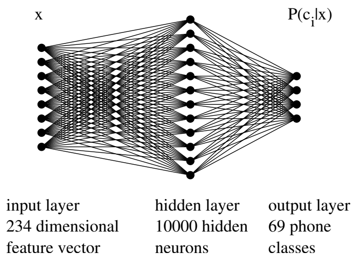
Networks either MLPs or RBFs

3.2.1 MLPs:

- Regular, recurrent or predictive
- Predictive network's degree of match is seen from the prediction error. The net is not a discriminatively taught.

3.2.2 RBFs:

- unsupervised RBF-layer (k-means)
- sigmoid, softmax or linear output layer (linear not necessarily very probabilistic, but fast to train)



MLP-networks output is turned into probability by softmax function

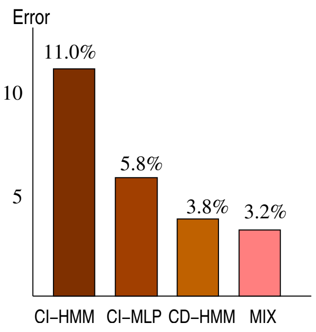
$$f(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

3.2.3 MLP vs RBF

- RBF can be very fast to train
  - RBF layer by k-means
  - linear output
  - backprop possible but slow – discriminative training for local functions (RBFs as opposed to sigmoids) not very useful
- With sufficient training time, MLP better

3.2.4 Hybrid MLP/HMM vs pure HMM

- Context Independent MLP easily wins Context Independent HMM
- CI MLP loses to context dependent HMMs.
- CD HMM has 50 times the number of models and 35 times number of parameters as compared to CI MLP.
- CI MLP is considerably faster in recognition, but harder to train
- The combination of CI MLP and CD HMM does very well



4. Conclusions

- Probabilistic neural networks are used in speaker recognition
- In speech recognition, they are mostly used in hybrid systems

5. Home exercises

1. Explain shortly the relations between following concepts: PNN (Probabilistic Neural Net), RBF (Radial Basis Function network), GMM (Gaussian Mixture models) and Parzen window estimation.
2. Both case studies presented here use neural nets and hidden Markov models. What is the fundamental difference between these two approaches ?