Automatic Speech Recognition (CS753)
Lecture 17: Discriminative Training for HMMs

Instructor: Preethi Jyothi
Sep 28, 2017
Discriminative Training
Recall: MLE for HMMs

Maximum likelihood estimation (MLE) sets HMM parameters so as to maximise the objective function:

\[ \mathcal{L} = \sum_{i=1}^{N} \log P_{\lambda}(X_i|M_i) \]

where
- \( X_1, ..., X_i, ... X_N \) are training utterances
- \( M_i \) is the HMM corresponding to the word sequence of \( X_i \)
- \( \lambda \) corresponds to the HMM parameters

What are some conceptual problems with this approach?
Discriminative Learning

- *Discriminative models* directly model the class posterior probability or learn the parameters of a joint probability model discriminatively so that classification errors are minimised.

- As opposed to *generative models* that attempt to learn a probability model of the data distribution.

- [Vapnik] “*one should solve the (classification/recognition) problem directly and never solve a more general problem as an intermediate step*”

[Vapnik]: V. Vapnik, Statistical Learning Theory, 1998
Discriminative Learning

- Two central issues in developing discriminative learning methods:
  
  1. Constructing suitable objective functions for optimisation
  
  2. Developing optimization techniques for these objective functions
Maximum mutual information (MMI) estimation: Discriminative Training

- MMI aims to directly maximise the posterior probability (criterion also referred to as conditional maximum likelihood)

\[
\mathcal{F}_{\text{MMI}} = \sum_{i=1}^{N} \log P_{\lambda}(M_{i}|X_{i})
\]

\[
= \sum_{i=1}^{N} \log \frac{P_{\lambda}(X_{i}|M_{i})P(W_{i})}{\sum_{W'} P_{\lambda}(X_{i}|M_{W'})P(W')}
\]

- \(P(W)\) is the language model probability
Why is it called MMI?

- Mutual information $I(X, W)$ between acoustic data $X$ and word labels $W$ is defined as:

$$I(X, W) = \sum_{x,w} \Pr(X, W) \log \frac{\Pr(X, W)}{\Pr(X) \Pr(W)}$$

$$= \sum_{x,w} \Pr(X, W) \log \frac{\Pr(W | X)}{\Pr(W)}$$

$$= H(W) - H(W | X)$$

where $H(W)$ is the entropy of $W$ and $H(W | X)$ is the conditional entropy.
Why is it called MMI?

• Assume $H(W)$ is given via the language model. Then, maximizing mutual information becomes equivalent to minimizing conditional entropy

$$H(W|X) = -\frac{1}{N} \sum_{i=1}^{N} \log \Pr(W_i|X_i)$$

$$= -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\Pr(X_i|W_i) \Pr(W_i)}{\sum_{W'} \Pr(X_i|W') \Pr(W')}$$

• Thus, MMI is equivalent to maximizing:

$$F_{\text{MMI}} = \sum_{i=1}^{N} \log \frac{P_\lambda(X_i|M_i)P(W_i)}{\sum_{W'} P_\lambda(X_i|M_{W'})P(W')}$$
MMI estimation

\[ F_{\text{MMI}} = \sum_{i=1}^{N} \log \frac{P_\lambda(X_i|M_i)P(W_i)}{\sum_{W'} P_\lambda(X_i|M_{W'})P(W')} \]

- Numerator: Likelihood of data given correct word sequence
- Denominator: Total likelihood of the data given all possible word sequences

How do we compute this?
Recall: Word Lattices

- A word lattice is a pruned version of the decoding graph for an utterance.
- Acyclic directed graph with arc costs computed from acoustic model and language model scores.
- Lattice nodes implicitly capture information about time within the utterance.

![Diagram of a word lattice](image-url)

Image from [GY08]: Gales & Young, Application of HMMs in speech recognition, NOW book, 2008
**MMI estimation**

\[
\mathcal{F}_{\text{MMI}} = \sum_{i=1}^{N} \log \frac{P_\lambda(X_i|M_i) P(W_i)}{\sum_{W'} P_\lambda(X_i|M_{W'}) P(W')} 
\]

- **Numerator**: Likelihood of data given correct word sequence
- **Denominator**: Total likelihood of the data given all possible word sequences
  - Estimate by generating lattices, and summing over all the word sequences in the lattice
MMI Training and Lattices

- Computing the denominator: Estimate by generating lattices, and summing over all the words in the lattice.

- Numerator lattices: Restrict G to a linear chain acceptor representing the words in the correct word sequence. Lattices are usually only computed once for MMI training.

- HMM parameter estimation for MMI uses the extended Baum-Welch algorithm \([V96,WP00]\).

- Like HMMs, can DNNs also be trained with an MMI-type objective function? Yes!

---

\([V96]\): Valtchev et al., Lattice-based discriminative training for large vocabulary speech recognition, 1996

\([WP00]\): Woodland and Povey, Large scale discriminative training for speech recognition, 2000
MMI results on Switchboard

- Switchboard results on two eval sets (SWB, CHE). Trained on 300 hours of speech. Comparing maximum likelihood (ML) against discriminatively trained GMM systems and MMI-trained DNNs.

<table>
<thead>
<tr>
<th></th>
<th>SWB</th>
<th>CHE</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM ML</td>
<td>21.2</td>
<td>36.4</td>
<td>28.8</td>
</tr>
<tr>
<td>GMM MMI</td>
<td>18.6</td>
<td>33.0</td>
<td>25.8</td>
</tr>
<tr>
<td>DNN CE</td>
<td>14.2</td>
<td>25.7</td>
<td>20.0</td>
</tr>
<tr>
<td>DNN MMI</td>
<td>12.9</td>
<td>24.6</td>
<td>18.8</td>
</tr>
</tbody>
</table>

[V et al.]: Vesely et al., Sequence discriminative training of DNNs, Interspeech 2013
Another Discriminative Training Objective: Minimum Phone/Word Error (MPE/MWE)

• MMI is an optimisation criterion at the sentence-level. Change the criterion so that it is directly related to sub-sentence (i.e. word or phone) error rate.

• MPE/MWE objective function is defined as:

\[
F_{\text{MPE/MWE}} = \sum_{i=1}^{N} \log \frac{\sum_{W} P_{\lambda}(X_{i}|M_{W})P(W)A(W, W_{i})}{\sum_{W'} P_{\lambda}(X_{i}|M_{W'})P(W')}
\]

where \(A(W, W_{i})\) is phone/word accuracy of the sentence \(W\) given the reference sentence \(W_{i}\) i.e. the total phone count in \(W_{i}\) minus the sum of insertion/deletion/substitution errors of \(W\).
The MPE/MWE criterion is a weighted average of the phone/word accuracy over all the training instances.

\[ F_{\text{MPE/MWE}} = \sum_{i=1}^{N} \log \frac{\sum_{W} P_{\lambda}(X_{i}|M_{W}) P(W) A(W, W_{i})}{\sum_{W'} P_{\lambda}(X_{i}|M_{W'}) P(W')} \]

- The MPE/MWE criterion can be computed either at the phone or word level for the MPE or MWE criterion, respectively.

- The weighting given by MPE/MWE depends on the number of incorrect phones/words in the string while MMI looks at whether the entire sentence is correct or not.
MPE results on Switchboard

- Switchboard results on eval set SWB. Trained on 68 hours of speech. Comparing maximum likelihood (MLE) against discriminatively trained (MMI/MPE/MWE) GMM systems

<table>
<thead>
<tr>
<th></th>
<th>SWB</th>
<th>%WER redn</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM MLE</td>
<td>46.6</td>
<td>-</td>
</tr>
<tr>
<td>GMM MMI</td>
<td>44.3</td>
<td>2.3</td>
</tr>
<tr>
<td>GMM MPE</td>
<td>43.1</td>
<td>3.5</td>
</tr>
<tr>
<td>GMM MWE</td>
<td>43.3</td>
<td>3.3</td>
</tr>
</tbody>
</table>

[V et al.]: Vesely et al., Sequence discriminative training of DNNs, Interspeech 2013
How does this fit within an ASR system?
Estimating acoustic model parameters

- If $A$: speech utterance and $O_A$: acoustic features corresponding to the utterance $A$,

$$W^* = \arg \max_W P_\lambda(O_A|W)P_\beta(W)$$

- ASR decoding: Return the word sequence that jointly assigns the highest probability to $O_A$

- How do we estimate $\lambda$ in $P_\lambda(O_A|W)$?
  - MLE estimation
  - MMI estimation
  - MPE/MWE estimation

Covered in this class