### Discriminative Training

Steve Renals

Automatic Speech Recognition — ASR Lecture 13
11 March 2013

#### Discriminative training

- **Basic idea** Estimate the parameters of a speech recognizer so as to make the fewest classification errors (optimize the word error rate)
- **Generative model:** estimate the parameters so that the model reproduces the training data with the greatest probability (maximum likelihood)
- Generative modelling only results in minimum classification error if certain conditions are met, including:
  - the model is correct (i.e. the true data source is an HMM)
  - infinite training data
- This never happens in practice
- **Discriminative training:**
  - Focus on learning *boundaries* between classes
  - Consider incorrect word sequences as well as correct word sequences
  - This is related to direct optimisation of the posterior probability of the words given the acoustics $P(W | X)$

#### Discriminative training criteria

- **Minimum classification error (MCE)**
  - Correct parameters if misrecognition occurs
  - Discriminant function is the difference between the log likelihood of the correct sentence and the average likelihood of incorrect competitors
  - Used mainly for small vocabularies
  - Uses training data inefficiently (only considers misrecognised examples)
- **Maximum mutual information estimation (MMIE)**
  - Maximise the mutual information between the acoustics and word sequence
  - Variant of conditional maximum likelihood
- **Minimum Bayes Risk (MBR)**
  - Optimise the word error rate rather than a likelihood ratio
  - Use the string edit distance between competing and reference utterances
  - Minimum phone error (MPE) training

#### Maximum likelihood estimation (MLE)

- Maximum likelihood estimation (MLE) sets the parameters so as to maximize an objective function $F_{\text{MLE}}$:

$$F_{\text{MLE}} = \sum_{u=1}^{U} \log P_{\lambda}(X_u | M(W_u))$$

for training utterances $X_1 \ldots X_U$ where $W_u$ is the word sequence given by the transcription of the $u$th utterance, $M(W_u)$ is the corresponding HMM, and $\lambda$ is the set of HMM parameters
- This objective function can be maximised by the EM algorithm (also known as Baum-Welch algorithm or Forward-Backward algorithm when applied to HMMs)
MLE — Updating the mean

- Update equation for the mean vector $\hat{\mu}_{jm}$ for Gaussian component $m$ of GMM associated with state $s_j$ is:

$$\hat{\mu}_{jm} = \frac{\sum_{u=1}^U \sum_{t=1}^T \gamma_u^t(s_j, m) x_t^u}{\sum_{u=1}^U \sum_{t=1}^T \gamma_u^t(s_j, m)}$$

where $\gamma_u^t(s_j, m)$ is the probability of the model occupying mixture component $m$ of state $j$ at time $t$ given training sentence $X_u$.

- Some extra notation:

$$\Theta_{jm}^u(M) = \sum_{t=1}^T \gamma_u^t(s_j, m) x_t^u \quad \Gamma_{jm}^u(M) = \sum_{t=1}^T \gamma_u^t(s_j, m)$$

$$\hat{\mu}_{jm} = \frac{\sum_{u=1}^U \Theta_{jm}^u(M(W_u))}{\sum_{u=1}^U \Gamma_{jm}^u(M(W_u))}$$

Maximum mutual information estimation

- Maximum mutual information estimation (MMIE) aims to directly maximise the posterior probability (sometimes called conditional maximum likelihood). Using the same notation as before, with $P(w)$ representing the language model probability of word sequence $w$:

$$F_{MMIE} = \sum_{u=1}^U \log P_\lambda(M(W_u) | X_u)$$

$$= \sum_{u=1}^U \log \frac{P_\lambda(X_u | M(W_u)) P(W_u)}{\sum_{w' \neq w} P_\lambda(X_u | M(w'_u)) P(w'_u)}$$

Maximum mutual information estimation

- Numerator: $P_\lambda(X_u | M(W_u)) P(W_u)$
  - the likelihood of the data given the correct word sequence — similar to the MLE objective function. $M_{num}$ is combined acoustic & language models used in the numerator

- Denominator: $\sum_{w' \neq w} P_\lambda(X_u | M(w'_u)) P(w'_u)$
  - the total likelihood of the data given all possible word sequences — obtained by summing over all possible word sequences estimated by the full acoustic and language models in recognition ($M_{den}$):

$$P(X | M_{den}) = \sum_{w'} P_\lambda(X_u | M(w'_u)) P(w'_u)$$

- The objective function $F_{MMIE}$ is optimised by making the correct word sequence likely, and all other word sequences unlikely

Extended Baum-Welch

- No EM-based optimization approach for $F_{MMIE}$
- Gradient-based approaches are straightforward but slow
- Approximation: Extended Baum-Welch (EBW) algorithm provides update formulae similar to forward-backward recursions used in MLE.

- Extended Baum-Welch — Updating the mean:

$$\mu_{jm}^u = \frac{\sum_{u=1}^U \left[ \Theta_{jm}^u(M_{num}) - \Theta_{jm}^u(M_{den}) \right] + D \mu_{jm}^u}{\sum_{u=1}^U \left[ \Gamma_{jm}^u(M_{num}) - \Gamma_{jm}^u(M_{den}) \right] + D}$$

- Can interpret $D$ as a weight between old and new estimates; in practice $D$ estimated for each Gaussian to ensure variance updates are positive
EBW and Lattices

- Computing $\Theta_{jm}^u(M_{\text{den}})$ involves summing over all possible word sequences — estimate by generating lattices, and summing over all words in the lattice.
- In practice also compute numerator statistics using lattices (useful for summing multiple pronunciations).
- Generate numerator and denominator lattices for every training utterance.
- Denominator lattice uses recognition setup (with a weaker language model).
- Each word in the lattice is decoded to give a phone segmentation, and forward-backward is then used to compute the state occupation probabilities.
- Lattices not usually re-computed during training.

MPE: Minimum phone error

- Basic idea: adjust the optimization criterion so it is directly related to word error rate.
- Minimum phone error (MPE) criterion:
  \[
  F_{\text{MPE}} = \frac{\sum_{u=1}^{U} \log \left( \frac{\sum_{W} P_s(X_u | M(W)) P(W) A(W, W_u)}{\sum_{W'} P_s(X_u | M(W'_u)) P(W'_u)} \right)}{
  \sum_{W'} P_s(X_u | M(W'_u)) P(W'_u)}
  \]
- $A(W, W_u)$ is the phone transcription accuracy of the sentence $W$ given the reference $W_u$.
- $F_{\text{MPE}}$ is a weighted average over all possible sentences $w$ of the raw phone accuracy.
- Although MPE optimizes a phone accuracy level, it does so in the context of a word-level system: it is optimized by finding probable sentences with low phone error rates.

Example: meeting speech recognition

<table>
<thead>
<tr>
<th>System</th>
<th>Training criterion</th>
<th>PLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>ML</td>
<td>28.7</td>
</tr>
<tr>
<td>SAT</td>
<td>ML</td>
<td>27.6</td>
</tr>
<tr>
<td>SAT</td>
<td>MPE</td>
<td>24.5</td>
</tr>
</tbody>
</table>

Summary

- Discriminative methods optimize a criterion other than maximum likelihood (e.g., more directly related to the error rate).
- But, we still want to optimize all parameters according to a consistent criterion.
  - MMI — directly optimize the posterior probability of the word sequence given the data.
  - MPE — scale the posterior word sequence probability by an estimate of the phone error rate.
- Discriminative training has a number of technical issues relating to smoothing the parameter updates.
- Reading: sec 27.3.1 of Young (2008).