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.
MLE and MMIE

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

$$F_{MLE} = \log P(X|\lambda) = \sum_{u=1}^{U} \log P(X_u | W_u)$$

- Maximum mutual information estimation (MMIE) (also referred to as conditional maximum likelihood, CML) aims to directly maximize the posterior probability:

$$F_{MMIE} = \sum_{u=1}^{U} \log P(X_u | W_u) \frac{P(W_u)}{P(X_u)} = \sum_{u=1}^{U} \log \frac{P(X_u | W_u)P(W_u)}{\sum_{W} P(X_u | W)P(W)}$$

$X_u$ is the acoustic observation sequence for the $u$th utterance

Optimizing the MMIE objective function

- No straightforward efficient optimization approach for $F_{MMIE}$
- Gradient-based approaches are straightforward but slow
- Extended Baum-Welch (EBW) algorithm provides update formulae similar to forward-backward recursions used in MLE
- Extended by Povey (PhD thesis, 2003) using notions of strong-sense and weak-sense auxiliary functions
- For large vocabulary tasks, estimating the denominator is expensive (an unpruned decoding) — in practice it is estimated using word lattices to restrict the set of words sequences that are summed over

MMIE

$$F_{MMIE} = \sum_{u=1}^{U} \log \frac{P(X_u | W_u)P(W_u)}{\sum_{W} P(X_u | W)P(W)}$$

- The denominator sums over all possible word sequences estimated by the full acoustic and language models in recognition
- The first term of the numerator is identical to the MLE objective function
- MMIE training corresponds to maximizing the likelihood, while simultaneously minimizing the denominator term
- Discriminative criterion: maximize the probability of the correct sequence (as in MLE) while simultaneously minimizing the probability of all possible word sequences

MPE: Minimum phone error

- Basic idea: adjust the optimization criterion so it is directly related to word error rate (cf. Minimum Bayes Risk (MBR))
- Minimum phone error (MPE) criterion

$$F_{MPE} = \sum_{u=1}^{U} \sum_{W} P(X_u | W)A(W, W_u)$$

$$= \sum_{u=1}^{U} \sum_{W} P(X_u | W)P(W)A(W, W_u)$$

$A(W, W_u)$ is the phone transcription accuracy of the sentence $W$ given the reference $W_u$

- $F_{MPE}$ 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>

Combining multiple feature streams

- **Basic idea** Different representations of the speech signal are possible: if they result in complementary errors than it may reduce error rates to combine them
- Combination at the feature level: linear discriminant analysis (and related methods) to combine feature streams
- Combination at the acoustic model level: combine frame-level probability estimates (multi-stream methods)
- Combination at the system level: combine the word sequence outputs of different recognizers (ROVER)

Feature combination

- **Basic idea** Compute different feature vectors for each frame and train acoustic models on all of them
- Simplest approach: concatenate feature vectors at each frame
  - Increases the dimensionality
  - May be strong correlations between the feature streams (can cause problems for diagonal covariance Gaussians)
- Transform concatenated feature vectors (linear discriminant analysis (LDA), principal component analysis (PCA))
  - dimension reduction
  - decorrelation
- PCA estimates a global transform; LDA estimates a transform per-class / per-state / per-component

LDA: Linear discriminant analysis

- LDA aims to find a linear transformation (from $d$ dimensions to $p$ dimensions, $p \leq d$) given by a matrix $\theta^T$:
  \[
  z = \theta^T x
  \]
- $\theta^T$ projects $x$ to a vector $z$ in a lower dimension space
- The LDA transform $\theta^T$ is chosen to simultaneously
  - maximise the between class covariance $\Sigma_b$
  - minimise the within class covariance $\Sigma_w$
  using the eigenvectors corresponding to the $p$ largest eigenvalues of $\Sigma_b^\Sigma_w^{-1}$
- HLDA: Heteroscedastic Linear Discriminant Analysis
  - In LDA classes share the same within-class covariance matrix
  - In HLDA a different covariance matrix is estimated for each class
- Both HLDA and LDA assume a Gaussian distribution
- NB: “class” may be a phone, a state or a Gaussian component, depending on the amount of data
Example: STRAIGHT features

- Conventional PLP and MFCC computation use a fixed size analysis window
- STRAIGHT spectral representation (Kawahara et al., 1999): smoothed spectral representation computed using a pitch adaptive window
- Requires a use of a pitch tracker to obtain $F_0$
- Resolution of STRAIGHT spectrogram follows the values of the fundamental frequency
- Can use STRAIGHT spectral analysis to obtain STRAIGHT MFCCs (and STRAIGHT PLPs)
- For recognition, combine STRAIGHT and conventional MFCCs using HLDA, reducing from 78 dimensions ($39+39$) to 39

Results on CTS

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Female</th>
<th>Male</th>
<th>SW1</th>
<th>S23</th>
<th>Cell</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC (no CMN/CVN)</td>
<td>42.7</td>
<td>41.8</td>
<td>43.6</td>
<td>36.5</td>
<td>43.3</td>
<td>47.9</td>
</tr>
<tr>
<td>STRAIGHT (no CMN/CVN)</td>
<td>45.7</td>
<td>44.5</td>
<td>46.9</td>
<td>40.0</td>
<td>46.6</td>
<td>50.3</td>
</tr>
<tr>
<td>MFCC+CMN/CVN+VTLN</td>
<td>37.6</td>
<td>37.0</td>
<td>38.3</td>
<td>31.8</td>
<td>37.1</td>
<td>43.5</td>
</tr>
<tr>
<td>STRAIGHT +CMN/CVN+VTLN</td>
<td>39.2</td>
<td>38.2</td>
<td>40.1</td>
<td>33.6</td>
<td>39.0</td>
<td>44.5</td>
</tr>
<tr>
<td>MFCC + STRAIGHT +CMN/CVN+VTLN+HLDA</td>
<td>34.7</td>
<td>33.8</td>
<td>35.6</td>
<td>28.6</td>
<td>34.7</td>
<td>40.5</td>
</tr>
</tbody>
</table>

Results on Meetings

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Female</th>
<th>Male</th>
<th>CMU</th>
<th>ICSI</th>
<th>LDC</th>
<th>NIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC+VTLN</td>
<td>38.4</td>
<td>38.5</td>
<td>38.5</td>
<td>42.7</td>
<td>23.9</td>
<td>52.1</td>
<td>30.9</td>
</tr>
<tr>
<td>STRAIGHT+VTLN</td>
<td>39.3</td>
<td>38.3</td>
<td>39.7</td>
<td>44.7</td>
<td>24.8</td>
<td>53.1</td>
<td>31.2</td>
</tr>
<tr>
<td>MFCC+STR AIGHT +VT LN</td>
<td>42.1</td>
<td>44.4</td>
<td>41.0</td>
<td>45.6</td>
<td>28.5</td>
<td>55.4</td>
<td>37.0</td>
</tr>
<tr>
<td>MFCC+STR AIGHT VT LN+HLDA</td>
<td>36.6</td>
<td>36.3</td>
<td>36.7</td>
<td>41.0</td>
<td>22.5</td>
<td>51.2</td>
<td>28.5</td>
</tr>
</tbody>
</table>
Example: Discriminative features

- Can also use the outputs of other statistical models as a feature stream
- Neural networks (e.g., multi-layer perceptrons – MLPs) when trained as a phone classifier output a posterior probability \( P(\text{phone|data}) \)
- This is a locally discriminative model
- MLP probability estimates can be used as an additional feature stream, modelled by the HMM/GMM system (Tandem)
- Advantages of discriminative features
  - can be estimated from a large amount of temporal context (e.g., ±25 frames)
  - encode phone discrimination information
  - only weakly correlated with PLP or MFCC features

Example: meeting speech recognition

- Tandem (LCRC – left context, right context) features (Karafiat, 2007)
- Derived from multiple stages of MLPs that try to estimate phoneme state posterior probabilities
- Wide context input to these is not only the feature vector at the current time, but 25 surrounding frames as well
- Separate MLPs for left and right context

<table>
<thead>
<tr>
<th>System</th>
<th>Training criterion</th>
<th>PLP</th>
<th>LCRC+PLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>ML</td>
<td>28.7</td>
<td>25.2</td>
</tr>
<tr>
<td>SAT</td>
<td>ML</td>
<td>27.6</td>
<td>23.9</td>
</tr>
<tr>
<td>SAT</td>
<td>MPE</td>
<td>24.5</td>
<td>21.7</td>
</tr>
</tbody>
</table>

Tandem features

- Tandem features
  - Speech
  - Time buffer
  - Mel-bank

- LCRC features
  - PLP, ∆, ∆∆
  - Log
  - DCTs
  - Neural network

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
- Combining features can take advantage of approaches which are complementary, but still make different errors
- Increasing emphasis on approaches which view the features as another model to be optimized