

Special Issues in Speech Recognition: Discriminative Training

Roger Hsiao

InterACT, Language Technologies Institute, Carnegie Mellon University
Pittsburgh, PA

Outline

- 1 Maximum Likelihood Estimation and Bayes' Rule
- 2 Maximum Mutual Information Estimation
- 3 Minimum Classification Error
- 4 Summary

Synopsis

- 1 Objective
- 2 The focus of today's class is about **discriminative training on acoustic model**.
- 3 Highlights
 - 1 Revisit maximum likelihood training and Bayes's decision theory.
 - 2 Cover two discriminative training algorithms based on very different theories.
 - 3 Model training is interesting and challenging.

Maximum Likelihood Estimate

- In class, we learn Baum-Welch algorithm which provides **maximum likelihood estimate (MLE)** for hidden Markov model (HMM).
- In MLE training, we try to find parameters such that the model gives high likelihood to the train data.

$$\theta^* = \underset{\theta}{\operatorname{argmax}} P(X|W; \theta)$$

- Does MLE training always give the best estimate?

MLE and Bayes' Decision Theory

- What does it mean by **the best**?
 - we want best classification accuracy on the train set.
- Bayes Decision Theory tells us best classification can be achieved by selecting w^* s.t.

$$\begin{aligned}w^* &= \operatorname{argmax}_w P(w|X) \\ &= \operatorname{argmax}_w \frac{P(X|w)P(w)}{P(X)} \\ &= \operatorname{argmax}_w P(X|w)P(w)\end{aligned}$$

- So, this is the reason why MLE only considers $P(X|w)$.

MLE and Bayes' Decision Theory

- What does it mean by **the best**?
 - we want best classification accuracy on the train set.
- Bayes Decision Theory tells us best classification can be achieved by selecting w^* s.t.

$$\begin{aligned}w^* &= \operatorname{argmax}_w P(w|X) \\ &= \operatorname{argmax}_w \frac{P(X|w)P(w)}{P(X)} \\ &= \operatorname{argmax}_w P(X|w)P(w)\end{aligned}$$

- So, this is the reason why MLE only considers $P(X|W)$.
- Is there something wrong...?

MLE and Bayes' Decision Theory

- However, we are working on **training, not classification!**

$$\begin{aligned}P(w|X) &= \frac{P(X|w)P(w)}{P(X)} \\ &= \frac{P(X|w)P(w)}{\sum_w P(X|w)P(w)}\end{aligned}$$

We care about θ ,

$$\frac{P(X|w; \theta)P(w)}{\sum_w P(X|w; \theta)P(w)}$$

- MLE considers $P(X)$ as a constant. It is true for classification, but during training, $P(X)$ depends on all the parameters of different models!
- Can we do better than MLE?

Conditional Maximum Likelihood Estimation

- How about we optimize for the posterior probability that Bayes decision rule cares about?

$$\begin{aligned}\theta^* &= \operatorname{argmax}_{\theta} P(W|X; \theta) \\ &= \operatorname{argmax}_{\theta} \frac{P(X|W; \theta)P(W)}{P(X; \theta)} \\ &= \operatorname{argmax}_{\theta} \frac{P(X|W; \theta)P(W)}{\sum_w P(X|w; \theta)P(w)}\end{aligned}$$

- This is maximizing the posterior probability or conditional likelihood (not to be confused with MAP in Bayesian learning).
- What does it mean intuitively?

Maximum Mutual Information Estimation I

- Before we continue, let's take a quick look about mutual information (empirical version),

$$\begin{aligned} I(X, W_i) &= \log \frac{P(X, W_i)}{P(X)P(W_i)} \\ &= \log \frac{P(X|W_i; \theta)P(W_i)}{P(W_i) \sum_j P(X|W_j; \theta)P(W_j)} \end{aligned}$$

- If we have equal prior, maximizing conditional likelihood is the same as maximizing mutual information.
- This procedure is known as **maximum mutual information estimation (MMIE)**.

Maximum Mutual Information Estimation II

- We can further simplify the formula:

$$\begin{aligned}
 P(W_i|X; \theta) &= \frac{P(X|W_i; \theta)P(W_i)}{P(X; \theta)} \\
 &= \frac{P(X|W_i; \theta)P(W_i)}{\sum_j P(X|W_j; \theta)P(W_j)} \\
 &= \frac{P(X|W_i; \theta)P(W_i)}{P(X|W_i; \theta)P(W_i) + \sum_{j \neq i} P(X|W_j; \theta)P(W_j)} \\
 &= \frac{1}{1 + \frac{\sum_{j \neq i} P(X|W_j; \theta)P(W_j)}{P(X|W_i; \theta)P(W_i)}}
 \end{aligned}$$

- Then, maximizing conditional likelihood is equivalent to maximizing,

$$\frac{P(X|W_i; \theta)P(W_i)}{\sum_{j \neq i} P(X|W_j; \theta)P(W_j)}$$

Maximum Mutual Information Estimation III

- In log domain, we have

$$\begin{aligned} & \log \frac{P(X|W_i; \theta)P(W_i)}{\sum_{j \neq i} P(X|W_j; \theta)P(W_j)} \\ &= \log P(X|W_i; \theta)P(W_i) - \log \left(\sum_{j \neq i} P(X|W_j; \theta)P(W_j) \right) \end{aligned}$$

- Therefore, MMIE can be considered as maximizing the log likelihood difference between the true class and the competing classes.
- What is different between MLE and MMIE?

MLE v.s. MMIE

- Theoretically speaking, if we use a correct family of prior and likelihood distribution, MLE is in fact better than MMIE because the smaller variance in its estimator [A. Nadas, 1983].
- In practice, **none** of the **prior (language model)** and the **likelihood (acoustic model) distribution** use a correct model - MMIE often performs better than MLE in practice.
- While MLE training only involves parameters from the correct model, MMIE also operates on the parameters from competing models.
- MMIE is more computationally intensive than MLE.
- **How do we apply MMIE in speech recognition?**

Optimization for MMIE

- To actually perform MMIE, we need an optimization algorithm and one of the simplest algorithms is the **gradient ascent**.

$$F(\theta) = \log P(X|W_i; \theta)P(W_i) - \log\left(\sum_{j \neq i} P(X|W_j; \theta)P(W_j)\right)$$

- Then, we can optimize θ iteratively:

$$\theta^{t+1} = \theta^t + \lambda_t \nabla F(\theta^t)$$

- It converges (almost surely) as long as λ is small enough and it follows:

$$\sum_{t=0}^{\infty} \lambda_t = \infty \quad \sum_{t=0}^{\infty} \lambda_t^2 < \infty \quad \lambda_t > 0$$

- This is how we perform MMIE in 80's.

Extended Baum-Welch Algorithm for MMIE

- Gradient ascent is theoretically sound, but slow convergence and tuning difficulty are critical drawbacks.
- Baum-Welch (BW)

$$\mu_j = \frac{\sum_t \gamma_t^{\text{ref}}(j) x_t}{\sum_t \gamma_t^{\text{ref}}(j)}$$

- Extended Baum-Welch (EBW)

$$\mu_j = \frac{\sum_t \gamma_t^{\text{ref}}(j) x_t - \sum_t \gamma_t^{\text{com}}(j) x_t^i + D_j \mu_j^0}{\sum_t \gamma_t^{\text{ref}}(j) - \sum_t \gamma_t^{\text{com}}(j) + D_j}$$

- Compared to BW, EBW also considers competing classes.
- The D constant controls the learning rate.
- EBW is not guaranteed to converge but it works very well in practice.

MMIE Training Procedure

- 1 Use MLE to train a baseline system.
- 2 Decode the train set using an unigram LM.
 - So we can obtain lattices for MMIE.
 - Why do we use unigram LM?
- 3 Perform EBW algorithm for several iterations.

MMIE in Practice

Table: WER of different ASR systems using MLE and MMIE.

System	Specs	MLE	MMIE	Rel. imprv.
Mandarin	1300hr, 108K vocab	15.1%	13.8%	8.6%
Iraqi	320hr, 62K vocab	39.2%	35.9%	8.4%
Farsi	110hr, 33K vocab	50.7%	46.7%	7.9%

Beyond Bayes' Rule: Minimum Error Rate

- Both MLE and MMIE are based on the framework of Bayes' rule.
- Later on, researchers are interested in whether we can **optimize the recognition error directly**.
- In other words, we are trying to define an objective function that well represents the recognition error, and we try to optimize the model parameters for such objective.
- This approach is known as minimum error rate training, and the candidate that we are going to discuss today is the **minimum classification error (MCE)** training.

Minimum Classification Error I

- In speech recognition, we use edit distance between reference and hypothesis to compute **word error rate (WER)**.
- However, the error function involves discrete entities, which are difficult to relate them to HMM parameters.
- MCE suggests the following simplified, smoothed error measure:

$$e_i(X) = -d_i(X, \theta) + \left[\frac{1}{N-1} \sum_{j \neq i} d_j(X, \theta)^\eta \right]^{\frac{1}{\eta}}$$

Minimum Classification Error II

$$e_i(X) = -d_i(X, \theta) + \left[\frac{1}{N-1} \sum_{j \neq i} d_j(X, \theta)^\eta \right]^{\frac{1}{\eta}}$$

- e_i represents error of the reference class i : $e_i > 0$ implies recognition error and $e_i \leq 0$ means correct recognition.
- d_i is known as the discriminant functions which are non-negative. It can be considered as a scoring function that how much the recognizer thinks the observation X should be classified as class i .
 - d_i has to be differentiable.
- N is the number of classes.
- η controls how to weigh the competing classes.

Minimum Classification Error III

- The MCE objective is defined as:

$$l_i(X) = \text{sigmoid}(e_i(X))$$

where the sigmoid function is defined as,

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}}$$

- The sigmoid function serves as a smoothed step function so MCE concentrates on errors that are relative easy to fix.
- The objective function, l , can be considered as a **soft recognition error function**.
- During MCE training, the utterances which are correctly recognized are skipped.
- The optimization is based on gradient descent.

MCE v.s. MMIE

- MMIE optimizes **posterior probability**, while MCE optimizes **soft recognition error**.
- Both MMIE and MCE requires a reasonable initial model (say, from MLE).
- MCE can use N-best lists or lattices as competitors.
- MCE does not have efficient optimization algorithm.
- Can we use MCE to simulate MMIE?

Summary

- In today's class, we revisited Bayes' rule and we learned why **MLE may not be the best choice for model training.**
 - But MLE is often a good choice.
- MLE optimizes likelihood, MMIE optimizes conditional likelihood (posterior probability) and MCE optimizes soft recognition error.
 - Both MLE and MMIE use Bayes' rule, MCE defines its own objective function.
- Discriminative training is **computationally expensive.**
 - It considers competing classes while MLE only considers the reference.
 - Optimization is more difficult.
- However, discriminative training often **out performs** MLE training.