

Assignment 4

Teaching assistant: Wilson Tam (yct@cs.cmu.edu)

Due on: Nov 24th before class

Instructions: Problem 1 and Problem 2 are compulsory, while you can choose either Problem 3 or Problem 4.

Problem 1: Expectation Maximization

The expectation-maximization (EM) algorithm is a technique for maximum likelihood estimation involving hidden variables. The aim of this problem is to help you understand how the optimization function is established in the context of a simple Gaussian mixture model (GMM) with K mixtures. Assume the GMM generates the observation vectors $X=x_1, x_2, \dots, x_T$ from $t=1$ to $t=T$ independently. However, we do not know which Gaussian mixture was picked to generate the observation at time t . In other words, the hidden variables here are the mixture indices ($1..K$) to generate the observation at each time t . Denote the mixture index at time t as q_t . Therefore, the hidden mixture sequence is $Q=q_1, q_2, \dots, q_T$.

For your information, a multivariate Gaussian distribution has the following form:

$$p(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n \cdot |\Sigma|}} \cdot e^{-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)} \quad (1)$$

where μ and Σ denote the mean and the covariance parameters respectively, and n denotes the feature dimension.

1. Derive the log likelihood of X when Q is observed i.e. $\log p(X, Q)$?
2. Hence, write down $\log p(X)$ in terms of $p(X, Q)$?
3. Using the Jensen's inequality $\log \sum_y p(y) f(y) \geq \sum_y p(y) \log f(y)$ where $p(y)$ is a probability distribution, and the posterior distribution $P^{(j-1)}(q_t|X)$ computed using an old model from the $(j-1)$ th training iteration, show that the lower-bound of $\log p(X)$ denoted as $Q(X; \{\lambda_k, \mu_k, \Sigma_k\})$, known as the EM auxiliary function, is:

$$Q(X; \{\lambda_k, \mu_k, \Sigma_k\}) = \sum_{t=1}^T \sum_{k=1}^K P^{(j-1)}(q_t = k|X) \left(-\frac{1}{2}(x_t - \mu_k)^t \Sigma_k^{-1}(x_t - \mu_k) - \frac{1}{2} \log |\Sigma_k| + \log \lambda_k \right) + \text{Constant}$$

4. Hence, compute the derivative of $Q(X; \{\mu_k\})$ with respect to the mean vector of the k -th Gaussian. Show that the reestimation formula for μ_k is:

$$\hat{\mu}_k^{(j)} = \frac{\sum_{t=1}^T P^{(j-1)}(q_t = k|X) \cdot x_t}{T_k}$$

where $T_k = \sum_{t=1}^T P^{(j-1)}(q_t = k|X)$ and $\sum_{k=1}^K T_k = T$

How does the reestimation formula compare with the MLE formula with a single Gaussian (i.e. without hidden variables)?

Problem 2: Acoustic Adaptation

A multivariate Gaussian distribution is given with known mean and covariance parameters. Given the adaptation data $X = \{x_1, x_2, \dots, x_T\}$, we want to find a transformation matrix $A_{n \times n}$ such that the likelihood of the *transformed* adaptation data is maximized. An input vector x can be transformed linearly as $y = Ax + b$ where b is a bias vector.

1. Derive the log likelihood of the *transformed* adaptation data.
2. Derive the maximum likelihood estimation of A via matrix differentiation.
3. Is transformation of the feature vector identical to transformation of the mean vector μ in a single Gaussian distribution? Explain briefly.
4. A GMM is applied with K mixtures. Using the results from Problem 1, write down the EM auxiliary function involving A , i.e. $Q(X;A)$.
5. By differentiation, write down the MLE condition for A .

Bonus: If a diagonal covariance of each Gaussian mixture is assumed, derive a reestimation formula for A in closed form.

Problem 3: Discriminative training

1. In acoustic modeling using HMM, explain briefly why discriminative training is usually desirable?
2. Does discriminative training always help? Give a scenario when discriminative training may hurt.
3. The maximum mutual information (MMI) criterion can be expressed as follows:

$$I(X, W) = \log \frac{p(X|W) \cdot P(W)}{\sum_{W' \neq W} p(X|W') \cdot P(W')} \quad (2)$$

Will you employ a (strong) trigram LM or a (weak) unigram LM for $P(W)$? Explain briefly your choice.

4. In MMIE for continuous speech recognition, explain why the competing classes are usually represented either by a word lattice or a N-best list, but not a 1-best competitor.
5. Assume a two-class $\{\omega_1, \omega_2\}$ classification problem with each class modeled by a Gaussian distribution. As a continuation of problem 2, write down the MMI criterion involving the feature transformation matrix A assuming the observation vectors X come from class ω_1 . You can ignore the term $P(W)$ without the loss of generality.
6. Compute the derivative of the above MMI criterion with respect to A . Can you obtain a closed-form solution for A ?

Problem 4: Search

1. What is the difference between synchronous (e.g. Viterbi) and asynchronous search (e.g. A^*).
2. What is inadmissible? How does it happen in A^* search?
3. Programming exercise: Implement A^* search to extract 3-best hypotheses from a word lattice. Download the lattice file from the course website. Report each extracted word hypothesis with a total score. Explain your choice of the heuristic function $h(\cdot)$. The word lattice has the following format:

```
StartNode EndNode Info(a=acoustic score, l=lmscore, wp=word penalty) Word CombinedScore
.. .. .. .. ..
FinalNode
..
node t=timestamp
.. ..
```

Each word and the score (in \log_{10}) is associated to an edge in the word lattice.