Discriminative Training

Notes on Speech and Audio Processing

Chia-Ping Chen

Department of Computer Science and Engineering

National Sun Yat-Sen University

Kaohsiung, Taiwan ROC

Introduction

- With the maximum-likelihood criterion, we train the model parameters so that the data likelihood for the correct model is non-decreasing.
- This method, however, does not consider the likelihood of incorrect models. It may well happen that the increase of likelihood in a wrong model is more than the correct model.
- A model training method that incorporates the likelihood of competing models is called discriminative training.

Issues with MLE

- The posterior probability for model M is

$$P(M|X) = \frac{P(X|M)P(M)}{\sum_{M'} P(X|M')P(M')} = \frac{1}{1 + \sum_{M' \neq M} \frac{P(X|M')P(M')}{P(X|M)P(M)}}$$

- To minimize the error probability, we need to maximize the posterior probability of the correct class, say M.
- Changing parameters to increase $P(X|M)$ does not necessarily increase $P(M|X)$.

Maximum Mutual Information

- The mutual information between M and X is

$$I(M,X|\lambda) = \log \frac{P(M,X|\lambda)}{P(M|\lambda)P(X|\lambda)}$$
$$= \log \frac{P(X|M,\lambda)}{\sum_{M'} P(X|M',\lambda)P(M'|\lambda)}$$

- This is quite similar to the posterior probability (except for $\log$ and $P(M|\lambda)$). Therefore, increasing MI also increases the posterior probability.
- The denominator can consist an infinite number of terms. There are feasible approximations to the denominator such as lattice or N-best.

Corrective Training

- The parameters are modified for any data where the correct model have a lower likelihood than the best model by a margin Δ.
- In other words,

$$P(X|M_r,\lambda) \geq P(X|M_c,\lambda) + \Delta \ \Rightarrow \ \lambda \rightarrow \lambda'$$

such that

$$\begin{cases} P(X|M_c,\lambda') \geq P(X|M_c,\lambda) \\ P(X|M_r,\lambda') \leq P(X|M_r,\lambda) \end{cases}$$

Discriminant Functions

- A framework for classification using discriminant functions is as follows. We define a discriminant function for each class,

$$g_j(X;\lambda), \quad j = 1, \ldots, K.$$

- The classification rule is simply

$$j^* = \arg\max_j \; g_j(X;\lambda).$$

- A sample X of class j will be classified correctly if

$$g_j(X;\lambda) > g_i(X;\lambda) \;\; \forall i \neq j.$$

Misclassification Measure

- We can define a misclassification measure based on the values of discriminant functions. Specifically, for data X of class j, we can define

$$d_j(X;\lambda) = \log\left\{\frac{1}{K-1}\sum_{k\neq j} e^{\eta g_k(X;\lambda)}\right\}^{\frac{1}{\eta}} - g_j(X;\lambda).$$

- If X is classified correctly, then $d_j(X) < 0$. Put differently, if $d_j(X) > 0$, then a classification error occurs.

Minimum Classification Error

- The number of errors in the training data is lower bounded by the number of data where $d_j(X) > 0$. Therefore, we can use $d_j(X)$ as the argument of a step function to count the number of errors.
- For the entire data set, this amounts to

$$E(\lambda) = \sum_j \sum_{X \in M_j} F(d_j(X;\lambda)),$$

where $F(x)$ is approximating a step function. $E(\lambda)$ is minimized to get the optimal model parameters λ^*.