CHAPTER 4

Discriminative Learning Algorithm for Exponential-Family Distributions

In this chapter, we describe an efficient, growth transformation (GT)-based approach to the discriminative parameter estimation problem in classifier design where each class is characterized by an exponential-family distribution discussed in Chapter 1. The next chapter extends the results here into the more difficult but practically more useful case of hidden Markov models (HMMs).

4.1 EXPONENTIAL-FAMILY MODELS FOR CLASSIFICATION

In this section, we derive the GT formulas for estimating parameters of exponential family distributions. This class of densities covers a large number of contemporary statistical models and is of important theoretical and practical interests. The derived formulas "grow" the unified rational-form discriminative training criterion $O(\Lambda)$ defined in Chapter 2. In the next chapter, we will present the derivation for the Gaussian mixture density HMM, which is widely used in modern speech recognition.

Let us start from the problem of C-class classification. Let the data of each class be i.i.d. (independent and identically distributed) that are modeled by an exponential family distribution. Although parameter estimation for this problem has a nice closed-form solution under ML training, it is more complicated under discriminative training. In the latter case, the objective function $O(\Lambda)$ is difficult to optimize directly but because it is a rational function as expressed in (3.2), we can construct the auxiliary functions of F and then V based on F (see Section 1.5.5). Optimizing $V(\Lambda;\Lambda')$ becomes a relatively easy problem and it leads to the GT formula for all types of discriminative criteria unified by (3.2).

Assume that there are R observation samples $x_r (r = 1, \ldots, R)$ in the training set and that each sample x_r is a vector with dimension D. Each sample x_r is associated with a reference label (e.g., a class index) $S_r \in \{c_i \mid i = 1, \ldots, C\}$, where C denotes the total number of classes in the task. Using the above notations, the task is considered a C-class classification problem, where each observation sample x_r is to be classified into one of the C classes.

Note that, denoting by θ_i the natural parameter vector of the exponential family distribution of the ith class, and denoting by $\Lambda = \{\theta_i\}$ the whole model parameter set, $p(x|c_i)$ takes the following form (as (1.6) in Chapter 1):

$$p(x|c_i;\Lambda) = p(x|\theta_i) = b(x) \cdot \exp\left\{\theta_i^{\mathrm{T}} T(x) - A(\theta_i)\right\} \tag{4.1}$$

4.2 CONSTRUCTION OF AUXILIARY FUNCTIONS

According to Section 3.2.2, the objective function of discriminative training is a rational function. Following the background material presented in Section 1.5.5, we can construct the objective function of

$$O(\Lambda) = \frac{G(\Lambda)}{H(\Lambda)} \tag{4.2}$$

in the same form of (1.26), where $G(\Lambda)$ and $H(\Lambda)$ are the nominator and denominator of (3.2). Then, the GT-based optimization algorithm constructs the auxiliary function of

$$F(\Lambda;\Lambda') = G(\Lambda) - O(\Lambda')H(\Lambda) + D \tag{4.3}$$

where D is a quantity independent of the parameter set Λ, and Λ' denotes the parameter set obtained from the immediately previous iteration of the algorithm. The purpose of constructing (4.3) is that it is easier to optimize than (4.2). However, (4.2) is still difficult to optimize, and we desire to introduce another auxiliary function from $V(\Lambda;\Lambda')$ in (4.3). This new function is constructed by

$$V(\Lambda;\Lambda') = \int_{\chi} f(\chi,\Lambda')\log f(\chi,\Lambda)\mathrm{d}\chi \tag{4.4}$$

where the positive, real-valued function $f(x, \Lambda) > 0$ is constructed to satisfy

$$F(\Lambda;\Lambda') = \int_{\chi} f(\chi,\Lambda)\mathrm{d}\chi \tag{4.5}$$

Then, we have

$$\begin{aligned}\log F(\Lambda;\Lambda') - \log F(\Lambda';\Lambda') &= \log \frac{F(\Lambda;\Lambda')}{F(\Lambda';\Lambda')} \\ &= \log \int_{\chi} \frac{f(\chi,\Lambda')}{F(\Lambda';\Lambda')} \frac{f(\chi,\Lambda)}{f(\chi,\Lambda')} \mathrm{d}\chi \geq \int_{\chi} \frac{f(\chi,\Lambda')}{F(\Lambda';\Lambda')} \log \frac{f(\chi,\Lambda)}{f(\chi,\Lambda')} \mathrm{d}\chi \\ &= \frac{1}{F(\Lambda';\Lambda')} \left[\int_{\chi} f(\chi,\Lambda') \log f(\chi,\Lambda) \mathrm{d}\chi - \int_{\chi} f(\chi,\Lambda') \log f(\chi,\Lambda') \mathrm{d}\chi \right] \\ &= \frac{1}{F(\Lambda';\Lambda')} \left[V(\Lambda;\Lambda') - V(\Lambda';\Lambda') \right] \end{aligned} \tag{4.6}$$

The inequality above is attributable to Jensen's inequality being applied to the concave log function. The result of (4.6) states that an increase in the auxiliary function $V(\Lambda; \Lambda')$ guarantees an increase in $\log F(\Lambda; \Lambda')$. Because logarithm is a monotonically increasing function, this also guarantees an increase of $F(\Lambda; \Lambda')$ and hence the original objective function $O(\Lambda)$. The technique that "transforms" the parameters from Λ' to Λ so as to increase or "grow" the values of the auxiliary functions and hence the value of the original objective function is called the growth-transformation (GT) technique. We now apply this GT technique to the exponential-family distribution with the unified discriminative optimization criterion formulated in (3.2).

4.3 GT LEARNING FOR EXPONENTIAL-FAMILY DISTRIBUTIONS

In this section, we derive the GT formula for general exponential-family distributions. The formula "grows" the unified discriminative training criterion $O(\Lambda)$. As discussed above, $O(\Lambda)$ is difficult to optimize directly but because it is a rational function as expressed in (3.2), we can construct the auxiliary functions of (1) F and then (2) V based on F. Optimizing $V(\Lambda; \Lambda')$ becomes a relatively easier problem and it leads to the final GT formula for all types of discriminative criteria unified by (3.2). In the next section, we will present the derivation for two specific exponential-family distributions — multinomial distribution and Gaussian distribution.

In the rational function of

$$O(\Lambda) = \frac{G(\Lambda)}{H(\Lambda)} \tag{4.7}$$

as the unified form of the discriminative objective function (3.2) for maximum mutual information, minimum classification error, and minimum phone error/minimum word error, we have

$$G(\Lambda) = \sum_s p(X, s|\Lambda) C(s) \text{ and } H(\Lambda) = \sum_s p(X, s|\Lambda) \tag{4.8}$$

where we use $s = s_1, \ldots, s_R$ to denote the class label (including correct or incorrect labels) for each of the R training tokens, respectively, and use $X = x_1, \ldots, x_R$ to denote the observation samples for these R training tokens. Note that each observation sample is a feature vector.

For the auxiliary function of

$$F(\Lambda; \Lambda') = G(\Lambda) - O(\Lambda')H(\Lambda) + D \tag{4.9}$$

we substitute (4.8) into (4.9) to obtain the new auxiliary function

$$\begin{aligned} F(\Lambda; \Lambda') &= \sum_s p(X, s|\Lambda) C(s) - O(\Lambda') \sum_s p(X, s|\Lambda) + D \\ &= \sum_s p(X, s|\Lambda) \left[C(s) - O(\Lambda') \right] + D \end{aligned} \tag{4.10}$$

The main terms in the auxiliary function $F(\Lambda;\Lambda')$ above can be interpreted as the average deviation of the accuracy count.

Because $p(s)$ is the prior probability of s, it is irrelevant for optimizing Λ. Using $p(X,s|\Lambda) = p(s) \cdot p(X|s, \Lambda)$, we obtain

$$\begin{aligned} F(\Lambda; \Lambda') &= \sum_s [C(s) - O(\Lambda')] p(s) p(X|s, \Lambda) + D \\ &= \sum_s \sum_\chi \left[\Gamma(\Lambda') + \mathrm{d}(s) \right] p(\chi|s, \Lambda) \end{aligned} \tag{4.11}$$

where

$$\Gamma(\Lambda') = \delta(\chi, X) p(s) \left[C(s) - O(\Lambda') \right] \tag{4.12}$$

In (4.10), $D = \sum_s \mathrm{d}(s)$ is a quantity independent of the parameter set Λ. In (4.12), $\delta(\chi, X)$ is the Kronecker delta function for discrete valued observations, and χ represents the entire data space where X belongs. The summation over this data space is introduced here for accommodating the parameter-independent constant D; that is, $\sum_s \sum_\chi \mathrm{d}(s) p(\chi|s, \Lambda) = \sum_s \mathrm{d}(s) = D$ is a Λ-independent constant.

In the case where the observation vector is continuous valued, the summation operation above will be replaced with integration, and $\delta(\chi, X)$ in (4.12) needs to be a Dirac delta function.

We now proceed to construct the new auxiliary function of (4.4). To achieve this, we first identify from (4.5) and (4.11) that

$$f(\chi, s, \Lambda) = \left[\Gamma(\Lambda') + \mathrm{d}(s) \right] p(\chi|s, \Lambda)$$

To ensure that $f(\chi, s, \Lambda)$ above is positive, $\mathrm{d}(s)$ should be selected to be sufficiently large so that $\Gamma(\Lambda') + \mathrm{d}(s) > 0$ (note that $p(\chi|s, \Lambda)$ is nonnegative). This issue will be discussed in greater details in Section 5.4.

Then, using (4.4), we have

$$\begin{aligned} V(\Lambda;\Lambda') &= \sum_s \sum_\chi \left[\Gamma(\Lambda') + \mathrm{d}(s)\right] p(\chi \,|s,\Lambda') \log \left\{ \underbrace{\left[\Gamma(\Lambda') + \mathrm{d}(s)\right]}_{\text{optimization - indept}} p(\chi|s,\Lambda) \right\} \\ &= \sum_s \sum_\chi \left[\Gamma(\Lambda') + \mathrm{d}(s)\right] p(\chi \,|s,\Lambda') \log p(\chi \,|s,\Lambda') + \text{Const.} \\ &= \sum_s p(X,s|\Lambda') \left(C(s) - O(\Lambda')\right) \log p(X|s,\Lambda') \\ &\quad + \sum_s \sum_\chi \mathrm{d}(s) p(\chi|s,\Lambda') \log p(\chi \,|s,\Lambda') + \text{Const.} \end{aligned} \tag{4.13}$$

The key reason which makes this new auxiliary function (4.13) easier to optimize than that in (4.11) is the new logarithm in $\log p(X|\, s, \Lambda)$, which was absent in (4.11). As for the case of ML learning case for exponential-family distributions, this enables drastic simplification of the new auxiliary function of $V(\Lambda; \Lambda')$, which we outline below.

We first ignore the optimization-independent constant in (4.13), and divide $V(\Lambda; \Lambda')$ by another optimization-independent quantity, $p(X|\Lambda')$, in order to convert the joint probability $p(X, s\,|\Lambda')$ to the posterior probability $p(s, X|\Lambda')$. We then obtain an equivalent auxiliary function of

$$\begin{aligned} U(\Lambda;\Lambda') = &\overbrace{\sum_s p(s|X,\Lambda') \left(C(s) - O(\Lambda')\right) \log p(X|s,\Lambda')}^{\text{term - I}} \\ &+ \underbrace{\sum_s \sum_\chi \mathrm{d}'(s) p(\chi \,|s,\Lambda') \log p(\chi \,|s,\Lambda)}_{\text{term - II}} \end{aligned} \tag{4.14}$$

where

$$\mathrm{d}'(s) = \mathrm{d}(s)/p(X|\Lambda') \tag{4.15}$$

Note that $X = X_1, \ldots, X_R$ is a large aggregate of all training data with R independent tokens. For each token X_r, it is independent of each other and it depends only on the rth label. Hence, $\log p(X|\, s, \Lambda)$ can be decomposed, enabling simplification of both term-I and term-II in (4.14). We now elaborate on the simplification of these two terms.

$$\begin{aligned}\text{term - I} &= \sum_s p(s|X,\Lambda')\left(C(s) - O(\Lambda')\right)\sum_{r=1}^{R}\log p(x_r|s_r,\Lambda) \\ &= \sum_s p(s|X,\Lambda')\left(C(s) - O(\Lambda')\right)\sum_{\substack{r=1 \\ s_r=c_i}}^{R}\sum_{i=1}^{C}\log p(x_r|c_i,\Lambda)\end{aligned} \tag{4.16}$$

The simplification process for the second term in (4.14) is below (using the notations $\tilde{\mathbf{s}} = s_1, \ldots, s_{r-1}, s_{r+1}, \ldots, s_R$, $\tilde{\chi} = \chi_1, \ldots, \chi_{r-1}, \chi_{r+1}, \ldots, \chi_R$):

$$\begin{aligned}\text{term - II} &= \sum_s \mathrm{d}(s)\sum_{\chi_1,\ldots,\chi_R} p(\chi_1,\ldots,\chi_R|s,\Lambda')\sum_{r=1}^{R}\log p(\chi_r|s_r,\Lambda) \\ &= \sum_s \mathrm{d}'(s)\sum_{r=1}^{R}\sum_{\chi_r} p(\chi_r|s_r,\Lambda')\underbrace{\sum_{\tilde{\chi}} p(\tilde{\chi}|\tilde{s},\Lambda')}_{=1}\log p(\chi_r|s_r,\Lambda) \\ &= \sum_{r=1}^{R}\sum_{i=1}^{I}\mathrm{d}(r,i)\sum_{\chi_{r,t}} p(\chi_r|c_i,\Lambda')\log p(\chi_r|c_i,\Lambda)\end{aligned} \tag{4.17}$$

where

$$\mathrm{d}(r,i) = \sum_{s,\, s_r=c_i}\mathrm{d}'(s) \tag{4.18}$$

Substituting (4.16) and (4.17) into (4.14), and using (4.1), we have:

$$\begin{aligned}U(\Lambda;\Lambda') = &\sum_s p(s|X,\Lambda')\left(C(s) - O(\Lambda')\right)\sum_{\substack{r=1 \\ s_r=c_i}}^{R}\sum_{i=1}^{C}\log p(x_r|\theta_i) \\ &+ \sum_{r=1}^{R}\sum_{i=1}^{C}\mathrm{d}(r,i)\sum_{\chi_{r,t}} p(\chi_r|c_i,\Lambda')\log p(\chi_r|\theta_i)\end{aligned} \tag{4.19}$$

Because $p(\cdot|\theta_i)$ is an exponential density and therefore its logarithm is a linear function of the data, $U(\Lambda;\Lambda')$ becomes ready to be maximized, which we proceed below.

Setting, $\dfrac{\partial U(\Lambda;\Lambda')}{\partial\theta_i} = 0$, $i = 1, \ldots, C$, we obtain

$$0 = \sum_s p\left(s\middle|X,\Lambda'\right)\left(C(s) - O(\Lambda')\right)\sum_{\substack{r=1\\ s_r=c_i}}^{R}\left(T(x_r) - \frac{\partial A(\theta_i)}{\partial \theta_i}\right)$$

$$+ \sum_{r=1}^{R} \mathrm{d}(r,i)\sum_{\chi_{r,t}} p\left(\chi_r\middle|c_i,\Lambda'\right)\left(T(\chi_r) - \frac{\partial A(\theta_i)}{\partial \theta_i}\right)$$

with the constraints:

$$\sum_{\chi_{r,t}} p\left(\chi_r\middle|c_i,\Lambda'\right) = 1$$

$$\sum_{\chi_{r,t}} p\left(\chi_r\middle|c_i,\Lambda'\right) T(\chi_r) = \mathbb{E}_{p(\chi|\theta_i')}[T(\chi)]$$

Using

$$D_i = \sum_{r=1}^{R} \mathrm{d}(r,i)$$

we obtain the solution that satisfies

$$\frac{\partial A(\theta_i)}{\partial \theta_i} = \frac{\sum_s p\left(s\middle|X,\Lambda'\right)\left(C(s) - O(\Lambda')\right)\sum_{\substack{r=1\\ s_r=c_i}}^{R} T(x_r) + D_i \cdot \mathbb{E}_{p(\chi|\theta_i')}[T(\chi)]}{\sum_s p(s|X,\Lambda')\left(C(s) - O(\Lambda')\right)\sum_{\substack{r=1\\ s_r=c_i}}^{R} 1 + D_i} \tag{4.20}$$

If we define

$$\Delta\gamma(i,r) = \sum_s p\left(s\middle|X,\Lambda'\right)\left(C(s) - O(\Lambda')\right)\delta(s_r = c_i) \tag{4.21}$$

Then we can rewrite (4.20) as

$$\frac{\partial A(\theta_i)}{\partial \theta_i} = \frac{\sum_{r=1}^{R} \Delta\gamma(i,r)T(x_r) + D_i \cdot \mathbb{E}_{p(\chi|\theta_i')}[T(\chi)]}{\sum_{r=1}^{R} \Delta\gamma(i,r) + D_i} \tag{4.22}$$

Based on (4.22), we will present details in deriving the model estimation formulas for two important exponential-family distributions in the next section. These are multinomial distribution

and Gaussian distributions. The former is commonly used to model discrete distribution and the latter is widely used to model continuous random variables. Equation (4.22) is also applicable to all other members of the exponential family.

4.4 ESTIMATION FORMULAS FOR TWO EXPONENTIAL-FAMILY DISTRIBUTIONS

4.4.1 Multinomial Distribution

In this section we discuss the discriminative training formulas when $p(x|\theta_i)$ is a single-observation multinomial distribution. Readers are referred to Section 1.4.3 for a general introduction of multinomial distribution and its properties.

For the single observation multinomial distribution of the ith class, its standard form is

$$p(x|v_i) = \prod_{k=1}^{K} v_{i,k}^{x(k)}$$

where $x = [x(1), \ldots, x(K)]^{\mathrm{T}}$ is a K-dimensional observation vector and $v_i = [\, v_{i,1}, \ldots, v_{i,K}]^{\mathrm{T}}$ is the K-dimensional parameter vector.

The exponential-family form of the is single observation multinomial distribution

$$p(x|\theta_i) = b(x) \cdot \exp\left\{\theta_i^{\mathrm{T}} T(x) - A(\theta_i)\right\}$$

where

$$T(x) = \tilde{x} = [x(1),\ldots,x(K-1)]^{\mathrm{T}} \tag{4.23}$$

with $\tilde{x}$ being the observation vector that contains the first $K-1$ elements of x. Given the above definition, according to properties of the multinomial distribution, we have the following

$$\mathbb{E}_{p(\chi|\theta_i')}[T(\chi)] = \mathbb{E}_{p(\chi|v_i')}[\tilde{\chi}] = \tilde{v}_i' \tag{4.24}$$

where $\tilde{v}_i' = [v'_{i,1}, \ldots, v'_{i,K-1}]^{\mathrm{T}}$ is an parameter vector that contains only the first $K-1$ parameters.

Substituting (1.16), (4.23), and (4.24) into (4.22), and denote by $\tilde{v}_i' = [v'_{i,1}, \ldots, v'_{i,K-1}]^{\mathrm{T}}$, we have

$$\tilde{v}_i = \frac{\sum_{r=1}^{R} \Delta\gamma(i,r)\tilde{x}_r + D_i \cdot \tilde{v}_i'}{\sum_{r=1}^{R} \Delta\gamma(i,r) + D_i} \tag{4.25}$$

By summing both sides of (4.25) over $k = 1, \ldots, K-1$, we have

$$\sum_{k=1}^{K-1} v_{i,k} = \frac{\sum_{r=1}^{R} \Delta\gamma(i,r) \sum_{k=1}^{K-1} x_{r,k} + D_i \cdot \sum_{k=1}^{K-1} v'_{i,k}}{\sum_{r=1}^{R} \Delta\gamma(i,r) + D_i}$$

from which, we obtain

$$\begin{aligned} v_{i,K} &= 1 - \sum_{k=1}^{K-1} v_{i,k} \\ &= 1 - \frac{\sum_{r=1}^{R} \Delta\gamma(i,r) \sum_{k=1}^{K-1} x_{r,k} + D_i \cdot \sum_{k=1}^{K-1} v'_{i,k}}{\sum_{r=1}^{R} \Delta\gamma(i,r) + D_i} \\ &= \frac{\sum_{r=1}^{R} \Delta\gamma(i,r) \left(1 - \sum_{k=1}^{K-1} x_{r,k}\right) + D_i \cdot \left(1 - \sum_{k=1}^{K-1} v'_{i,k}\right)}{\sum_{r=1}^{R} \Delta\gamma(i,r) + D_i} \\ &= \frac{\sum_{r=1}^{R} \Delta\gamma(i,r) x_{r,K} + D_i \cdot v'_{i,K}}{\sum_{r=1}^{R} \Delta\gamma(i,r) + D_i} \end{aligned} \tag{4.26}$$

Combining (4.25) and (4.26), we have the GT estimation formula for multinomial distribution:

$$v_i = \frac{\sum_{r=1}^{R} \Delta\gamma(i,r) x_r + D_i \cdot v'_i}{\sum_{r=1}^{R} \Delta\gamma(i,r) + D_i} \tag{4.27}$$

4.4.2 Multivariate Gaussian Distribution

In this section, we will derive and present discriminative training formulas when $p(\mathbf{x}|\theta_i)$ is a multivariate Gaussian distribution. Readers are referred to Section 1.4.3 for a general introduction of multivariate Gaussian distribution and its properties.

For a multivariate Gaussian distribution of the ith class, its standard form is

$$p(x|\lambda_i) = N(x|\mu_i, \sum_i) = \frac{1}{(2\pi)^{\frac{D}{2}} \left|\sum_i\right|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x-\mu_i)^{\mathrm{T}} \sum_i^{-1} (x-\mu_i)\right\} \tag{4.28}$$

and its exponential family form is

$$p(x|\theta_i) = h(x) \cdot \exp\left\{\theta_i^{\mathrm{T}} T(x) - A(\theta_i)\right\}$$

where

$$T(x) = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x \\ \mathrm{Vec}\left(xx^{\mathrm{T}}\right) \end{bmatrix} \tag{4.29}$$

Given the above definition, according to the property of the multivariate Gaussian distribution, we have

$$\mathbb{E}_{p(\chi|\theta_i')}[T(\chi)] = \mathbb{E}_{p(\chi|\mu_i', \Sigma_i')} \begin{bmatrix} x \\ \mathrm{Vec}\left(xx^{\mathrm{T}}\right) \end{bmatrix} = \begin{bmatrix} \mu_i' \\ \mathrm{Vec}\left(\Sigma_i' + \mu_i'\mu_i'^{\mathrm{T}}\right) \end{bmatrix} \tag{4.30}$$

Substituting (1.22), (1.23), (4.29), and (4.30) into (4.22), we finally obtain

$$\begin{bmatrix} \mu_i \\ \mathrm{Vec}\left(\mu_i\mu_i^{\mathrm{T}} + \Sigma_i\right) \end{bmatrix} = \begin{bmatrix} \dfrac{\sum_{r=1}^{R} \Delta\gamma(i,r)x_r + D_i \cdot \mu_i'}{\sum_{r=1}^{R} \Delta\gamma(i,r) + D_i} \\ \dfrac{\sum_{r=1}^{R} \Delta\gamma(i,r)\mathrm{Vec}\left(x_r x_r^{\mathrm{T}}\right) + D_i \cdot \mathrm{Vec}\left(\Sigma_i' + \mu_i'\mu_i'^{\mathrm{T}}\right)}{\sum_{r=1}^{R} \Delta\gamma(i,r) + D_i} \end{bmatrix}$$

After rearrangement and canceling out the Vec() function at both sides, we can obtain the parameter updating formulas as

$$\mu_i = \frac{\sum_{r=1}^{R} \Delta\gamma(i,r)x_r + D_i \cdot \mu_i'}{\sum_{r=1}^{R} \Delta\gamma(i,r) + D_i} \tag{4.31}$$

$$\Sigma_i = \frac{\sum_{r=1}^{R} \Delta\gamma(i,r) x_r x_r^{\mathrm{T}} + D_i \cdot \left[\Sigma_i' + \mu_i' \mu_i'^{\mathrm{T}}\right]}{\sum_{r=1}^{R} \Delta\gamma(i,r) + D_i} - \mu_i \mu_i^{\mathrm{T}} \tag{4.32}$$

Equations (4.27), (4.31), and (4.32) give the discriminative training formula for the multinomial distribution and Gaussian distribution. The computation of $\Delta\gamma(i, r)$ will be presented in greater details in Chapter 6, and the issues of setting the constant D_i is discussed in Section 5.4.