**Automatic Speech Recognition**
**Lecture Note 5: Statistical Pattern Classification**

1. **Speech as a Stochastic Process**

   Speech signal is a random process. That is, there is uncertainty about the production, the transmission, and the receiving of speech signal. Such uncertainty stems from speaker, channel, and environmental variations. To describe a random process, we need some basic knowledge of the probability theory.

2. **Basic Probability Theory**

   The probability theory is a natural framework in which one describes processes with uncertainty. The following brief account is limited to the description of fundamental and related concepts.

   - **Random experiments**

     A random experiment has an outcome which is not predictable beforehand. The set of all possible outcomes is called the sample space. A subset in the sample space may have an associated non-negative probability. A mapping from all such subsets to the range $[0, 1]$ is called a probability measure. For a given probability measure, the probability of the union of disjoint subsets is the sum of the probabilities of each subset. The probability of the entire sample space is 1.

   - **Random variables**

     A random variable $X$ is a function that maps the sample space of a random experiment to the set of real numbers such that

     $$F(x) = Pr(X \leq x),$$

     called the *distribution function*, is defined for all $x$.

   - **Probability functions**

     For a *discrete* random variable, a *probability mass function*,

     $$P(x) = Pr(X = x),$$

     is often used to describe the probability. Note that $P(x)$ consists of the magnitudes of the step functions in $F(x)$.

For a *continuous* random variable, a *probability density function* $f(x)$, which is related to $F(x)$ by

$$F(x) = \int_{-\infty}^{x} f(u)du,$$

is often used. (Here an upper-case letter denotes a probability mass and a lower-case letters denotes a probability density.)

Note that

$$\sum_{x} P(x) = 1, \quad \text{if } X \text{ is discrete}$$
$$\int p(x)dx = 1, \quad \text{if } X \text{ is continuous}$$

- **Joint probability**

  For two random variables, $X$ and $Y$, the joint probability distribution function is defined by

  $$F(x, y) = Pr(X \leq x, Y \leq y).$$

  The *marginal* distribution function of $X$ and $Y$ are respectively,

  $$Pr(X \leq x) = F(x) = F(x, \infty)$$
  $$Pr(Y \leq y) = F(y) = F(\infty, y).$$

  If both $X$ and $Y$ are discrete, a corresponding joint probability mass function can be defined which satisfies

  $$P(x, y) = Pr(X = x, Y = y).$$

  Similarly, if both $X$ and $Y$ are continuous, a density function $f(x, y)$ can be defined which satisfies

  $$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v)dudv.$$

- **conditional probability**

  For two sets $A$ and $B$ with defined probability, the conditional probability that $A$ occurs given $B$ occurs is defined as

  $$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}.$$

For two discrete random variables, $X$ and $Y$, the conditional probability of $\{X = x\}$ given $\{Y = y\}$ is

$$Pr(X = x|Y = y) = \frac{P(x, y)}{P(x)},$$

where $P(x, y), P(x)$ are the probability mass functions. For two continuous random variables, the conditional probability density function $f(y|x)$ is defined by

$$\begin{aligned} f(y|x)dy &= \frac{Pr(y < Y \leq y + dy,\ x < X \leq x + dx)}{Pr(x < X \leq x + dx)} \\ &= \frac{f(x, y)dxdy}{f(x)dx} \\ \rightarrow f(y|x) &= \frac{f(x, y)}{f(x)}, \end{aligned}$$

where $f(x, y),\ f(x)$ are the probability density functions.

Given a discrete random variable, say $Y$, and a continuous random variable, say $X$, we may be interested in the joint or conditional probability functions. In this case, the probability is distributed discretely among a countable number of values of $Y$, and for the probability distributed to $y$, it is further distributed continuously among values of $x$. Although the probability can be described with the joint distribution function, it can be equivalently described with the probability mass function $P(y)$ and the conditional probability density function of $f(x|y)$. Specifically, the joint probability density function is

$$f(x, y) = P(y)f(x|y),$$

and the conditional probability of $Y = y$ given $X = x$ is

$$\begin{aligned} P(y|x) &= \frac{Pr(Y = y,\ x < X \leq x + dx)}{Pr(x < X \leq x + dx)} \\ &= \frac{f(x, y)dx}{\sum_v f(x, v)dx} \\ \rightarrow P(y|x) &= \frac{f(x, y)}{\sum_v f(x, v)} \\ &= \frac{P(y)f(x|y)}{\sum_v P(v)f(x|v)}. \end{aligned}$$

The point here is that the conditional probability of a discrete random variable given the value of a continuous random variable can be expressed by conditional density functions. Note that had $X$ been discrete, the conditional density functions can be replaced by the conditional mass functions,

$$P(y|x) = \frac{P(x,y)}{P(x)} = \frac{P(x,y)}{\sum_v P(x,v)} = \frac{P(y)P(x|y)}{\sum_v P(v)P(x|v)}.$$

3. **Class-Related Probability Functions**

   Assume that the observed data is $X$ and it can be from any class $\omega$. We want to make a decision of $\omega$, based on $X$. The following probabilities are crucial in solving this problem.

   - **a priori (or prior) probability** $P(\omega)$

     This represents our knowledge about the frequencies of different classes, prior to observing anything.

   - **conditional probability function** $f(x|\omega)$

     This is the probability function when $X$ is sampled from class $\omega$. This is also known as the likelihood function.

   - **a posteriori probability (or posterior)** $P(\omega|x)$

     This is the conditional probability of class $\omega$ given that $X = x$ has been observed. This quantity is directly related to classification error, as we will see next.

4. **Minimum Error Classification**

   The maximum a posteriori (MAP) decision rule is

   $$\omega^* = \arg\max_{\omega} P(\omega|x).$$

   This rule guarantees the minimum classification error probability. To see this, note that the probability of error is

   $$Pr(error) = \int f(x, error)dx = \int P(error|x)f(x)dx.$$

   The probability of error given $x$ is

   $$P(error|x) = 1 - P(\omega|x),$$

if the decision rule yields $\omega$ when $x$ is observed. The MAP criterion apparently minimizes $P(error|x)$ for all $x$, therefore it minimizes $Pr(error)$.

In principle, if we have $P(\omega|x)$ as a function of $x$, then the classification problem is solved with optimal solution. In practice, however, it may be difficult to obtain or even approximate this function.

5. **Likelihood-Based MAP Classification**

   By the Bayes' rule,

$$P(\omega|x) = \frac{f(x|\omega)P(\omega)}{f(x)}.$$

   (This is also shown earlier in the note with a slightly different form in the denominator.) Since $f(x)$ is constant for fixed $x$, the MAP decision rule is equivalent to

$$\omega^* = \arg\max_{\omega} \left[\log f(x|\omega) + \log P(\omega)\right].$$

   In this form, the objective function is decomposed to two terms. Here the terms on the right-hand side are the likelihood function and the prior probability respectively. These terms can be learned with labeled samples, independently for each class.

6. **An Example: Single Gaussian Models**

   Single Gaussian models assume that the likelihood function for a class, say $\omega$, is a Gaussian distribution, i.e.,

$$f(x|\omega) = \frac{1}{(\sqrt{2\pi})^d |\Sigma_\omega|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_\omega)^T \Sigma_\omega^{-1}(x-\mu_\omega)\right],$$

   where $d$ is the dimension of $x$. The MAP classifier simply chooses the class that maximizes the discriminant functions, $g_\omega(x)$, with

$$g_\omega(x) = -\frac{1}{2}(x-\mu_\omega)^T \Sigma_\omega^{-1}(x-\mu_\omega) - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_\omega| + \log P(\omega).$$

   Here the term $\frac{d}{2}\log(2\pi)$ is irrelevant and can be dropped. Further simplication is possible under additional assumptions. See the textbook for a few examples.

7. **Parameter Learning from Samples**

We have said that, in practice the true probability functions are often unknown and we have to estimate these from samples. Specifically, in the aforementioned likelihood-based MAP classification, we need to know the likelihood function for each class. Here we will describe the maximum-likelihood (ML) parameter estimation approach for learning a parametric distribution from samples.

We assume that each likelihood function has a parameteric form. The parameters in the function are unknown and need to be learned from data set generated from the likelihood function. Let $\Theta$ denote the set of parameters in the likelihood function of class $\omega$ and $D$ be the set of data labeled $\omega$. The ML approach seeks the values of the parameters that maximizes the data likelihood, i.e.,

$$\Theta^* = \arg\max_{\Theta} f(D|\omega, \Theta).$$

Note that the estimation for $\omega$ does not use any data from other classes. So the parameters can be learned independently for each class.

In some cases, the maximization problem has a closed-form solution for the optimal parameters as functions of sample data values. In cases where such a closed-form solution is unavailable, a common method is the *expectation-maximization* (EM) algorithm.

8. **The EM Algorithm**

In maximum likelihood parameter estimation, the objective is to find the model parameter set which maximizes data likelihood. The EM (Expectation-Maximization) algorithm is commonly used in such parameter estimation because the data likelihood is non-decreasing with iterations.

An auxiliary function is defined by

$$Q(\Theta, \Theta_0) = E[\log p(S, x|\Theta)] = \sum_{s=1}^{M} P(s|x, \Theta_0) \log p(s, x|\Theta), \quad (1)$$

where $\Theta_0$ is the current model parameter set, $x$ is the (observed) data, and $S$ is the hidden variable, which is not observed but is assumed to be involved in the data generation process. We assume that the value

of $S$ is $\{1, 2, \ldots, M\}$ This function is maximized with respect to the unknown $\Theta$. Suppose the solution in maximizing (1) is $\Theta^*$, then

$$Q(\Theta^*, \Theta_0) \geq Q(\Theta_0, \Theta_0). \tag{2}$$

Furthermore, the log data-likelihood and the $Q$ function can be related by

$$\begin{aligned}
&Q(\Theta^*, \Theta_0) - Q(\Theta_0, \Theta_0) \\
&= \sum_{s=1}^{M} [P(s|x, \Theta_0) \log f(s, x|\Theta^*) - P(s|x, \Theta_0) \log f(s, x|\Theta_0)] \\
&= \sum_{s=1}^{M} P(s|x, \Theta_0) [\log f(x|\Theta^*) + \log P(s|x, \Theta^*)] \\
&\quad - \sum_{s=1}^{M} P(s|x, \Theta_0) [\log f(x|\Theta_0) + \log P(s|x, \Theta_0)] \\
&= \log f(x|\Theta^*) - \log f(x|\Theta_0) - \sum_{s=1}^{M} P(s|x, \Theta_0) \log \frac{P(s|x, \Theta_0)}{P(s|x, \Theta^*)} \\
&= \log f(x|\Theta^*) - \log f(x|\Theta_0) - D(P_0||P),
\end{aligned} \tag{3}$$

where $P_0$ and $P$ are the probability mass function of $S$ given $x$ and parameter sets $\Theta_0$ and $\Theta^*$ respectively. Therefore,

$$\begin{aligned}
&\log f(x|\Theta^*) - \log f(x|\Theta_0) \\
&= Q(\Theta^*, \Theta_0) - Q(\Theta_0, \Theta_0) + D(P_0||P) \\
&\geq Q(\Theta^*, \Theta_0) - Q(\Theta_0, \Theta_0) \\
&\geq 0,
\end{aligned} \tag{4}$$

since $D(p||q)$, the KL-distance between two probability functions $p$ and $q$, is always non-negative. It follows from (4) that the data likelihood does not decrease with each maximization. The expectation in (1) and the maximization in (2) is the reason for the name EM. The data likelihood converges to a local maximum with this algorithm.

In (1), $x$ represents a single data point. For a batch of data samples, $x_1, x_2, \ldots, x_N$, one may use the $Q$ function of

$$Q(\Theta, \Theta_0) = \sum_{i=1}^{N} E \log f(S_i, x_i|\Theta) = \sum_{i=1}^{N} \sum_{s_i=1}^{M} P(s_i|x_i, \Theta_0) \log f(s_i, x_i|\Theta). \tag{5}$$

To show that increasing the value of the $Q$ function increases the data likelihood, note that

$$\sum_{i=1}^{N} \log f(x_i|\Theta) - \sum_{i=1}^{N} \log f(x_i|\Theta_0) = Q(\Theta, \Theta_0) - Q(\Theta_0, \Theta_0) + \sum_{i=1}^{N} D(P_0^i || P^i), \tag{6}$$

where $P_0^i$ is the posterior probability of $S$ given $x_i$ with parameter $\Theta_0$.

9. **Example of EM: Gaussian Mixture Models**

In Gaussian mixture models (GMM), the conditional probability density function of the data from class $\omega$ is a weighted sum of Gaussian density functions. That is,

$$f_\omega(x|\Theta) = \sum_{k=1}^{K} P(k|\Theta) f(x|k, \Theta) = \sum_{k=1}^{K} c_k N(\mu_k, \sigma_k^2),$$

where $N(\mu, \sigma^2)$ is a Gaussian density function with mean $\mu$ and variance $\sigma^2$. Note that $\sum_k c_k = 1$. We will demonstrate how to use EM to estimate the parameters $c_k's, \mu_k's$, and $\sigma_k's$ from samples of class $\omega$. Since it is an interative algorithm, an initial set of parameters is needed to kick things off. This is often carried out by clustering. Given the current parameter set $\Theta_0$, the $Q$ function is

$$\begin{aligned} Q &= \sum_{i=1}^{N} \sum_{k=1}^{K} P(k|x_i, \Theta_0) \left[\log(P(k|\Theta) + \log f(x_i|k, \Theta)\right] \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} P(k|x_i, \Theta_0) \left[\log c_k + \log N(\mu_k, \sigma_k^2)\right]. \end{aligned}$$

In this form, the first term on the right-hand side is dependent only on $c_k$ while the second term is dependent only on $\mu_k, \sigma_k$. Therefore they can be maximized separately. Note that there is a constraint on $c_k$ that they must sum to 1, which may be incorporated into the optimization problem through a Lagrange multiplier. After a little algebra as shown in the textbook, the optimal (updated) values for the parameters are

$$c_k = \frac{1}{N} \sum_{i=1}^{N} P_k^i,$$

$$\mu_k = \frac{\sum_{i=1}^{N} P_k^i \, x_i}{\sum_{i=1}^{N} P_k^i},$$

and

$$\sigma_k^2 = \frac{\sum_{i=1}^{N} P_k^i (x_i - \mu_k)^2}{\sum_{i=1}^{N} P_k^i},$$

where

$$P_k^i = P(k|x_i, \Theta_0) = \frac{f(x_i, k|\Theta_0)}{f(x_i|\Theta_0)} = \frac{f(x_i|k, \Theta_0) P(k|\Theta_0)}{\sum_{k'} f(x_i|k', \Theta_0) P(k'|\Theta_0)}.$$