ECE 8527: Exam 1

Andrew Powell

March 25, 2014

1 Problem 1

Consider two probability distributions defined by:

$$p(x|\omega_1) = \begin{cases} 1 & \alpha - 1/2 \leq x \leq \alpha + 1/2 \\ 0 & elsewhere \end{cases} \text{ and } p(x|\omega_2) = \begin{cases} 1 & -1/2 \leq x \leq 1/2 \\ 0 & elsewhere \end{cases}$$

1.1 A

Sketch the probability of error, p(E), for a maximum likelihood classifier as a function of α and P(ω_1). Label all critical points.

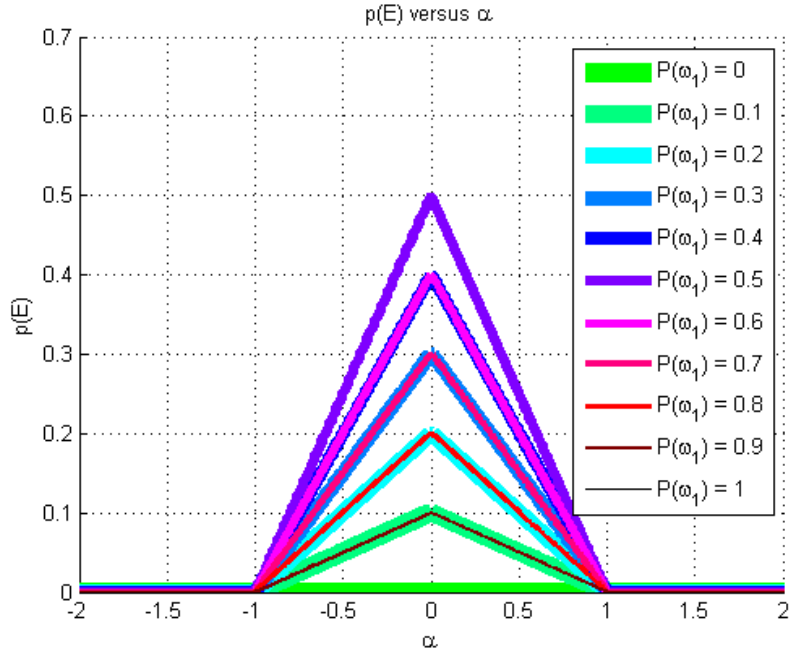
$$\begin{aligned} P(error) &= \begin{cases} p(\omega_1|x) & p(\omega_1|x) > p(\omega_2|x) \\ 0 & elsewhere \end{cases} + \begin{cases} p(\omega_2|x) & p(\omega_1|x) \leq p(\omega_2|x) \\ 0 & elsewhere \end{cases} \\ &= \begin{cases} p(x|\omega_1)P(\omega_1) & \frac{p(\omega_1|x)}{p(\omega_2|x)} > \frac{1-P(\omega_1)}{P(\omega_1)} \\ 0 & elsewhere \end{cases} + \begin{cases} p(x|\omega_2)(1-P(\omega_1)) & \frac{p(\omega_1|x)}{p(\omega_2|x)} \leq \frac{1-P(\omega_1)}{P(\omega_1)} \\ 0 & elsewhere \end{cases} \end{aligned} \tag{1}$$

The probability of error $P(error)$ is calculated based on Equation 1. If the notation is not clear, the probability of error is the sum of the posteriors $p(\omega_1|x)$, when $p(\omega_1|x)$ is greater than $p(\omega_2|x)$, and $p(\omega_2|x)$, when $p(\omega_2|x)$ is greater than or equal to $p(\omega_1|x)$. Conceptually, Equation 1 is indicative to deciding to which class a data point x belongs, based on which posterior is the largest. However, there is still a probability the data point x belongs to the class not chosen. This probability, which is also the probability of error $P(error)$, is one minus the posterior of the chosen class, or simply the posterior of the class not chosen.

Figure 1 is the result of calculating $P(error)$ with respect to the prior $P(\omega_1)$ and the parameter α. A few inferences matching Figure 1 can be made based on early observations. These observations include both likelihoods $p(x|\omega_1)$ and $p(x|\omega_2)$ have probability densities that are square pulses of the same width, and the parameter α from the likelihood $p(x|\omega_1)$ causes the likelihood's probability density to shift along its x axis.

Based on the observations, it is inferred the probability of error $P(error)$ will peak when the parameter α is 0, which signifies both the likelihoods $p(x|\omega_1)$ and $p(x|\omega_2)$ overlap and thus cause the most uncertainty in a class decision for any value of the prior $P(\omega_1)$. The probability of error $P(error)$ will result in 0 when the parameter α is greater than the magnitude of the parameter α, which signifies the the likelihoods are separated and thus the selection of the correct class ω for any data point x is clear.

Figure 1



Another important inference has to do with the prior $P(\omega_1)$. From a conceptual standpoint (common sense, really), the uncertainty in a class decision will peak when the prior $P(\omega_1)$ is equal to $1 - P(\omega_1) = .5$. In any classifier relying on Bayesian decision theory the priors are effectively trivial if all the priors are equal to each other. Equal priors indicates no prior information is known and thus a class (i.e. ω) can only be chosen based on the likelihood of a particular class (i.e. $p(x|\omega)$), given one or more features (i.e. the data points x).

Though obvious to even point out, the probability of error $P(error)$ will result in 0 for the prior $P(\omega_1)$ is either 0 or 1, signifying all the data points fall under a single class.

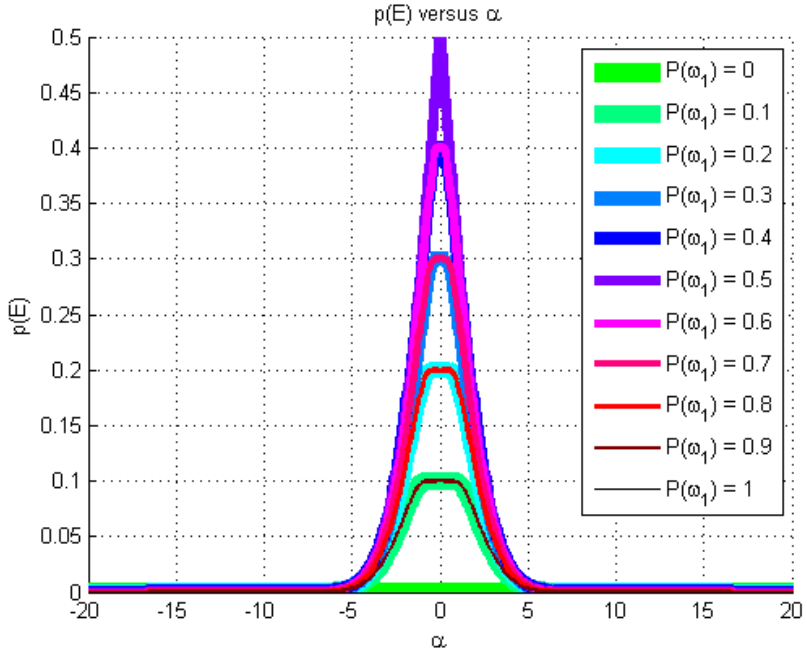
1.2 B

Suppose you estimated these distributions to be Gaussian distributions rather than uniform by analyzing a large amount of training data drawn from each distribution. How would your result in (a) change?

Figure 2 displays exactly how the results change for normally distributed likelihoods. The parameter α for the likelihood $p(x|\omega_1)$ is assumed to equal the likelihood's expected value, whereas the expected value for the second likelihood $p(x|\omega_2)$ is assumed to equal 0. variances for both likelihoods are assumed to equal 1, for simplicity.

Most of the observations and inferences made in Part A are applicable to Part B. For instance, the uncertainty peaks when the parameter α is equal to 0, which indicates both likelihoods overlap and thus the class decision rests entirely on the priors. The greater the magnitude of the parameter α is, the lower the probability of error $P(error)$. The probability of error also peaks when all priors are equal to each other; that is, when $P(\omega_1) = 1 - P(\omega_1) = .5$. Finally, the probability of error $P(error)$ is 0 if either $P(\omega_1) = 1$ or $P(\omega_1) = 0$ is true.

Figure 2



A significant difference is the "shape" the probability of error $P(error)$ takes for normally distributed likelihoods, as demonstrated in Figure 2.

2 Problem 2

Let x have a uniform density: $p(x|\theta) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \theta \\ 0 & elsewhere \end{cases}$. Suppose that n samples, $D = \{x_1, x_2, ..., x_n\}$ are drawn independently from $p(x|\theta)$. Derive an expression for the maximum estimate of θ. Hint: compute the likelihood of the data given θ and differentiate. Discuss what happens to this estimate as $n \to \infty$.

$$\begin{aligned} p(D|\theta) &= \prod_{k=1}^{n} p(x_k|\theta) \\ &= \prod_{k=1}^{n} \begin{cases} \frac{1}{\theta} & 0 \leq x_k \leq \theta \\ 0 & elsewhere \end{cases} \\ &= \begin{cases} \theta^{-n} & 0 \leq x_k \leq \theta \\ 0 & elsewhere \end{cases} \end{aligned} \tag{2}$$

Equation 2 is the result of independently drawing n amount of samples by which the data set D is defined. In other words, since the probability of drawing a value given the parameter θ (i.e. $p(x_k|\theta)$) does not affect the probability of another value being drawn for another sample, the *likelihood* of getting a particular data set D given the parameter θ (i.e. $p(D|\theta)$) is simply the product of all the samples' $p(x_k|\theta)$.

Since a parameterized model of the data's likelihood function $p(D|\theta)$ is known, the *Maximum Likelihood Estimation (MLE)* of the parameter $\hat{\theta}$ is the θ that causes the largest likelihood to result from the likelihood function $p(D|\theta)$. At first glance, it seems apparent the next step is to find the derivative of the likelihood function $p(D|\theta)$ with respective to the parameter θ.

In cases for which a particular function in question has a maximum, setting the derivative to zero can in fact result in the parameter that maximizes the original function. However, simply setting the derivative of Equation 2 to zero results in the parameter θ being undefined. The possible range over which each sample's value x_k is defined needs consideration in order to find the MLE of the parameter $\hat{\theta}$.

It follows from Equation 2, specifically the equality $0 \leq x_k \leq \theta$, the likelihood function $p(D|\theta)$ would result in zero if at least one of the conditions $\theta \geq \max\{x_k\}$ and $0 \leq \min\{x_k\}$ are false. It also follows from Equation 2, specifically the expression θ^{-n}, the likelihood goes to infinity as the parameter θ goes to zero. However, seeing as the condition $\theta \geq \max\{x_k\}$ causes the likelihood function $p(D|\theta)$ to fall zero once the parameter θ falls below the largest x_k in the data set D, the smallest parameter θ that can possibly maximize the likelihood function $p(D|\theta)$ can only be $\theta = \max\{x_k\}$.

Therefore, the MLE of the parameter $\hat{\theta}$ is $\max\{x_k\}$.

3 Problem 3

A zero-mean unit variance discrete-time Gaussian white noise signal, $x[n]$, is applied to a digital filter: $H(z) = \frac{1}{1-\alpha z^{-1}}$. Assume you only have access to the output of this filter, but you do know the form of the filter (you just don't know the specific value of α), and you can assume the input is zero-mean Gaussian white noise. Derive or explain how you construct a maximum

likelihood estimate of the filter coefficient. Hint: think about the pdf for the difference of two random variables. Second hint: think about the role correlation can play in this estimate.

$$y[n] = x[n] + \alpha y[n-1] \tag{3}$$

In order to derive an expression to estimate the parameter α with MLE, the likelihood function $l(\alpha) = p(D|\alpha)$ is needed. The first step in determining the likelihood function $l(\alpha)$ is to first perform the inverse z-transform on the digital filter in order to obtain a difference equation that calculates the output $y[n]$ for its Gaussian input $x[n]$. The difference equation is shown in Equation 3.

$$\begin{aligned}
p(y[n]|\alpha) &= p(x[n] + \alpha y[n-1]) \\
&\approx N(\mu, \sigma^2) \\
&\approx N(E[y[n]], \sigma^2) \\
&\approx N(E[x[n] + \alpha y[n-1]], \sigma^2) \\
&\approx N(E[x[n]] + E[\alpha y[n-1]], \sigma^2) \\
&\approx N(0 + E[\alpha y[n-1]], \sigma^2) \\
&\approx N(\alpha y[n-1], \sigma^2)
\end{aligned} \tag{4}$$

Because the input $x[n]$ of the digital filter is normally distributed, the output of the digital filter $y[n]$ is too normally distributed with a mean equal to the expected value of the output $y[n]$. The derivation is shown in Equation 4.

$$\begin{aligned}
p(D_N|\alpha) &= \prod_{n=0}^{N-1} N(\alpha y[n-1], \sigma^2) \\
p(D_N|\alpha) &= \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y[n]-\alpha y[n-1])}{\sigma}\right)^2} \\
\ln(p(D_N|\alpha)) &= \ln\left(\prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y[n]-\alpha y[n-1])}{\sigma}\right)^2}\right) \\
\ln(p(D_N|\alpha)) &= \sum_{n=0}^{N-1}\left[\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \ln e^{-\frac{1}{2}\left(\frac{y[n]-\alpha y[n-1]}{\sigma}\right)^2}\right] \\
\ln(p(D_N|\alpha)) &= \sum_{n=0}^{N-1}\left[k - \frac{1}{2}\left(\frac{y[n]-\alpha y[n-1]}{\sigma}\right)^2\right] \\
\ln(p(D_N|\alpha)) &= Nk - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(y[n]-\alpha y[n-1])^2 \\
\ln(p(D_N|\alpha)) &= Nk - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(y^2[n] - 2\alpha y[n]y[n-1] + \alpha^2 y^2[n-1]) \\
\ln(p(D_N|\alpha)) &= Nk - \frac{N}{2\sigma^2}(R_{yy}[0] - 2\alpha R_{yy}[1] + \alpha^2 R_{yy}[0])
\end{aligned} \tag{5}$$

The likelihood function $p(D|\alpha)$ is assumed to equal $p(D|\alpha, N)$ or $p(D_N|\alpha)$, where N is the size of each data set D. As shown in Equation 5, the expression of the likelihood function $p(D_N|\alpha)$

is determined by the product of N independently drawn samples, which are simply the outputs of the digital filter. The expression is simplified through application of the natural-log operation.

$$\begin{aligned}\frac{d\left[\ln(p(D_N|\alpha))\right]}{d\alpha} &= \frac{d\left[Nk - \frac{N}{2\sigma^2}(R_{yy}[0] - 2\alpha R_{yy}[1] + \alpha^2 R_{yy}[0])\right]}{d\alpha} \\ \frac{d\left[\ln(p(D_N|\alpha))\right]}{d\alpha} &= \frac{N}{\sigma_{yy}^2}[R_{yy}[1] - \alpha R_{yy}[0]]\end{aligned} \tag{6}$$

Since the simplified likelihood function, $\ln(p(D_N|\alpha))$, is known, calculating the parameter α that maximizes the simplified likelihood function is done by first differentiating the modified likelihood function with respect to the parameter α. The derivation is shown in Equation 6.

$$\begin{aligned}0 &= \frac{N}{\sigma_{yy}^2}[R_{yy}[1] - \alpha R_{yy}[0]] \\ 0 &= R_{yy}[1] - \alpha R_{yy}[0] \\ \alpha &= \frac{R_{yy}[1]}{R_{yy}[0]}\end{aligned} \tag{7}$$

Finally, the Maximum Likelihood Estimate of the parameter α is acquired by setting the derivative to 0 and then solving for the parameter α, as shown in Equation 7.