ECE 8110 Machine Learning Midterm 1

March 7, 2014

Salvatore Giorgi
salvatore.giorgi@temple.edu

Exercise 1. *Consider two probability distributions defined by*

$$p(x|\omega_1) = \begin{cases} 1, & \alpha - \frac{1}{2} \leq x \leq \alpha + \frac{1}{2} \\ 0, & \textit{elsewhere} \end{cases}$$

$$p(x|\omega_2) = \begin{cases} 1, & -\frac{1}{2} \leq x \leq \frac{1}{2} \\ 0, & \textit{elsewhere} \end{cases}.$$

(a) Sketch the probability of error $P(E)$ for a maximum likelihood classifier as a function of α and $P(\omega_1)$. Label all critcal points.

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

Solution for (a):

For a maximum likelihood classifier, we use the following decision rule: for an observation x, decide ω_1 if

$$P(\omega_1|x) > P(\omega_2|x);$$

and choose ω_2 otherwise. From this, we get the probability of error

$$P(\text{error}|x) = \begin{cases} P(\omega_2|x), & \text{if } x \in \omega_1 \\ P(\omega_1|x), & \text{if } x \in \omega_2 \end{cases},$$

which in turn gives us

$$P(\text{error}) = \int P(\text{error}|x)p(x)dx.$$

Using the fact that $P(\omega_1) + P(\omega_2) = 1$ and Baye's formula

$$P(\omega_i|x) = \frac{P(x|\omega_i)P(\omega_i)}{p(x)}$$

the total probability of error can be written as

$$P(\text{error}) = \int_{R_1} P(\text{error}|x)p(x)dx + \int_{R_2} P(\text{error}|x)p(x)dx$$

$$= \int_{R_1} P(x|\omega_2)P(\omega_2)dx + \int_{R_2} P(x|\omega_1)P(\omega_1)dx = \int_{R_1} P(x|\omega_2)\big(1 - P(\omega_1)\big)dx + \int_{R_2} P(x|\omega_1)P(\omega_1)dx.$$

$$= P(\omega_1)\Big[\int_{R_2} P(x|\omega_1)dx - \int_{R_1} P(x|\omega_2)dx\Big] + \int_{R_1} P(x|\omega_2)dx. \quad (1)$$

Next, we note that if $|\alpha| > 1$, the two distributions will not overlap. Thus, the integrals above will evaluate to zero and our probability of error will equal zero. Thus, we only need consider $-1 \leq \alpha \leq 1$. For this case, we also not that

$$\int_{R_2} P(x|\omega_1)dx = \int_{R_1} P(x|\omega_2)dx$$

and therefore the first part of (1) becomes zero and therefore, the probability of error does not depend on the prior probabilities. The final part of (1) gives us

$$P(\text{error}) = \int_{R_1} P(x|\omega_2)dx = 1 - |\alpha|, \quad (2)$$

which is plotted in Fig. 1.

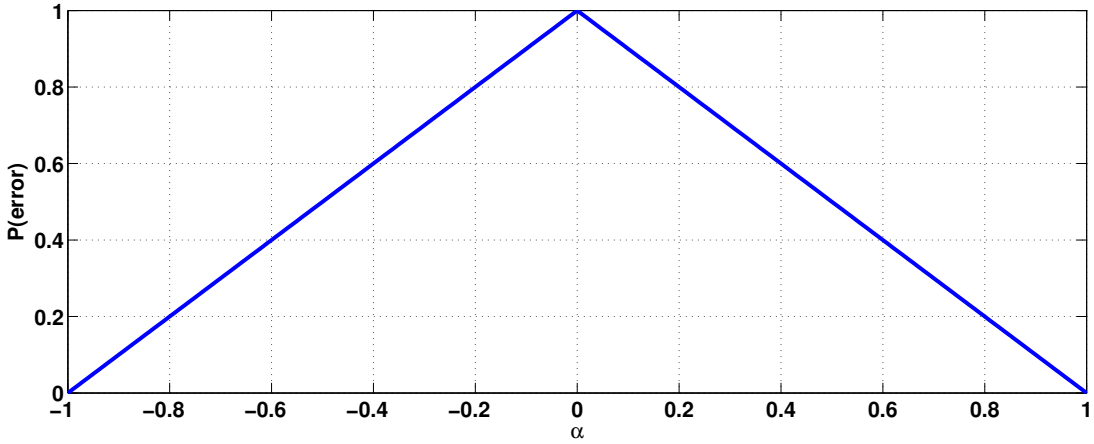


Figure 1: Probability of Error

For this decision rule and when $\alpha = 0$, the two distributions will completely overlap and the decision rule above will gives a $P(\text{error}) = 1$. To overcome this, we can use a different decision rule, in the case when $\alpha = 0$: random guessing based on the prior probabilities. This decision rule is: for the observations decide ω_1 $\big(100 \times P(\omega_1)\big)\%$ of the time and decide ω_2 for the remaining observations.

Solution for (b):

We have the following observations:

1) The above integral for the $P(error)$ is just the convolution of the two pdfs.
2) Given two independent random variables X, Y, the pdf of their sum $Z = X + Y$ is equal to the convolution of their respective pdfs.
3) The sum Z of two independent Gaussian random variables $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ is also Gaussian with $Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

The second two observations can be found in any introductory probability text book, such as [1]. With these observations, if we change the uniform distributions to Gaussian distributions,

we have that the probability of error, as a function of α, will have a closed form expression equal to the pdf of $Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$. (Note, we are not saying that $P(\text{error})$ is a random variable. We are just making some observations based on the convolution integral which simplify our calculation. What we are saying is that the analytical form of the $P(\text{error})$ will have the same form as a Gaussian RV.)

For this case, again, we have that $P(\text{error})$ is maximum at $\alpha = 0$. The maximum will be $\frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}}$. Again, if $\alpha = 0$ we can change the decision strategy to a similar one discussed above, which is based on the prior probabilities.

Exercise 2. *Let x have a uniform density:*

$$p(x|\theta) = \begin{cases} \frac{1}{\theta}, & 0 \leq x \leq \theta \\ 0, & \textit{otherwise} \end{cases}.$$

Suppose that n samples $D = \{x_1, \cdots, x_n\}$ are drawn independently from $p(x|\theta)$. Derive an expression for the maximum likelihood estimate of θ. Hint: compute the likelihood of the data given θ and differentiate. Discuss what happens to this estimate as $n \to \infty$.

Solution: In general, $p(D_n|\theta)$ is called the likelihood of θ with respect to the data D_n and, under the assumption that the n samples were drawn independently from the same distribution, we can write

$$p(D_n|\theta) = \prod_{k=1}^{n} p(x_k|\theta). \tag{3}$$

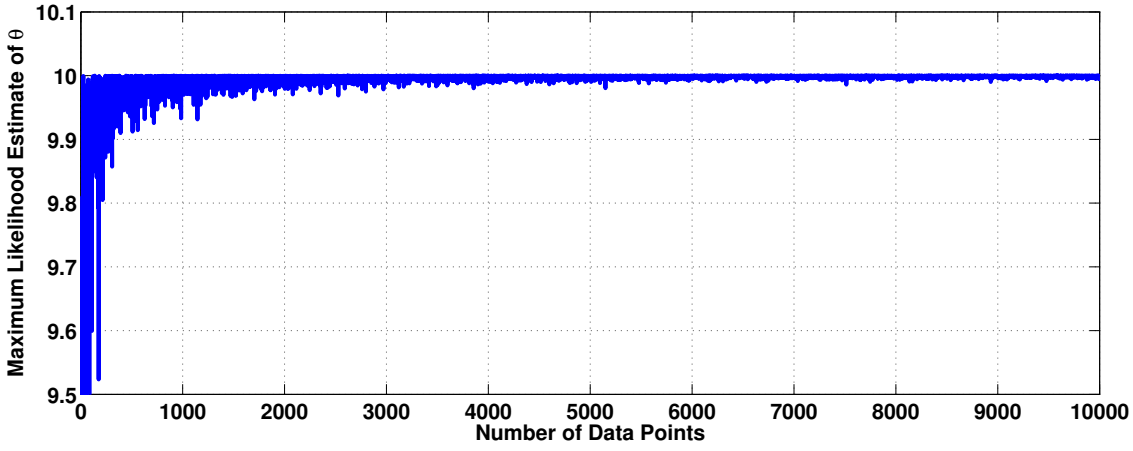
With each $p(x_k|\theta) \sim U(0, \theta)$ we can rewrite the above as

$$p(D_n|\theta) = \prod_{k=1}^{n} \frac{1}{\theta} = \theta^{-n} \tag{4}$$

using the pdf of a uniform random variable. Note, that we are assuming a known parametric form, which is completely determined by its parameters. This assumption, along with the I.I.D. data sets, are the two assumptions needed for a maximum likelihood estimate.

The value of θ that maximizes the above likelihood function, which we will denote $\hat{\theta}$, is the maximum likelihood esitmate of θ. Since each sample x_i is assumed to be taken from $U(0, \theta)$, we then have that $\theta \geq x_i$ for all i. Also, we note that the likelihood function is a decreasing function of θ. Thus, to maximize it, we must choose the smallest value of θ possible. So with the constraint that $\theta \geq x_i$ we have that $\hat{\theta} = \max\{x_1, \cdots, x_n\}$.

As $n \to \infty$ we will get $\max_i\{x_i\}$ closer and closer to θ, or in other words $|\theta - \max\{x_i\}| < \epsilon$ with $\epsilon \to 0$ as $n \to \infty$. We show this behavoir with a simulation. We take, $\theta = 10$, n data points from $U(0, \theta)$, let $\hat{\theta} = \max_i\{x_i\}$, and increase n.



Exercise 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 would 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.

Solution:

First, we discuss how the minimum mean square error and the maximum likelihood estimates are equal in this case. We follow the exposition given in [2]. Given a random signal y and an estimate of this signal $\hat{y}$, we wish to minimize the squared error between the two. Since our signal y is a random variable, then we must minimize the mean of the squared error (MMSE). One method of solving this minimization is to assume some prior knowledge of the estimate in the form of a pdf, which is a Bayesian approach. (We already know that maximum likelihood is a special case of Bayesian, so we have some idea that the two should be connected, given certain assumptions.) We denote the Bayesian mean square error as BMSE and we consider the problem of minimizing BMSE, i.e.,

$$\hat{y}_{\text{BMSE}} = \min E\{(y - \hat{y})^2\}.$$

In this case the expectation is take with respect to the joint pdf $p(x, y)$ where x is our data. Notice also we have that

$$E\{(y - \hat{y})^2\} = E\{\hat{y}^2\} + E\{y^2\} - 2yE\{\hat{y}\}$$
$$= \text{Var}(\hat{y}) + \big(E\{\hat{y}\}\big)^2 + y^2 - 2yE\{\hat{y}\} = \text{Var}(\hat{y}) + \big(E\{\hat{y}\} - y\big)^2.$$

Hence, our estimator will depend on its bias and its own variance. We see that this will be minimum when we have an unbiased estimator (when $y = E\{\hat{y}\}$), since we are assuming a fixed distribution.

Another Bayesian approach is a maximum a posteriori (MAP) estimate. For this problem, we choose $\hat{y}$ to maximize the posterior pdf, i.e.,

$$\hat{y}_{\text{MAP}} = \arg\max_y \big(p(y|x)\big).$$

Using Bayes formula, we have the equivalent problem of finding

$$\hat{y}_{\text{MAP}} = \arg\max_y \big(p(x|y)p(y)\big),$$

which is see is similar to a maximum likelihood (ML) estimate, except for the prior pdf $p(y)$

$$\hat{y}_{\text{ML}} = \arg\max_y \big(p(x|y)\big).$$

When this prior is uniform, we see that MAP and ML are equivalent, since $p(y)$ is a constant. Also, we note that for a Gaussian distribution, the maximum of its pdf is at its mean, which is its expected value. Thus, if $p(y|x)$ is Gaussian, then $E\{y|x\} = \hat{y}_{\text{ML}}$, which is the solution to the MMSE problem above.

From the above we have that under the assmption that the prior $p(y)$ is Uniform, then ML is equivalent to MAP. Additionally, under the assumption that the posterior $p(y|x)$ is Gaussian, we get that MMSE is equivalent to MAP. Thus, for ML to be equal to MMSE we need the prior $p(y)$ to be uniformly distributed and the posterior $p(y|x)$ to be Gaussian.

From the transfer function of the filter we have the following:

$$H(z) = \frac{1}{1-\alpha z^{-1}} = \frac{Y(z)}{X(z)} \Rightarrow X(z) = (1-\alpha z^{-1})Y(z)$$

$$\Rightarrow y[n] = x[n] + \alpha y[n-1] \Rightarrow x[n] = y[n] - \alpha y[n-1].$$

Since $x[n]$ is a Gaussian white noise disturbance, we see that $y[n] - \alpha y[n-1]$ also follows the same distribution. We assume we can only measure the output y and that we know past values of $y[i]$ for all $0 \leq i < n$. We use a maximum likelihood estimate of $y[n]$:

$$L(\alpha) = f(y[n]|\alpha) = \prod_{k=1}^{n-1} f_w(y[k] - \alpha y[k-1]) \Rightarrow$$

$$\log L(\alpha) = -\log(2\pi\sigma) - \frac{1}{2\sigma^2} \sum_{k=1}^{n-1} \big(y[k] - \alpha y[k-1]\big)^2.$$

We take the partial derivative with respect to α and set equal to zero:

$$\frac{\partial \log L(\alpha)}{\partial \alpha} = 0 \Rightarrow$$

$$\hat{\alpha} = \frac{1}{n-1} \sum_{k=1}^{n-1} \frac{y[k]}{y[k-1]} \tag{5}$$

Equation (5) gives us our maximum likelihood equation, which is the sample mean. Since we are working with Gaussian distributions, we know that this is an unbiased estimator, and therefore the solution to the MMSE problem.

Next, we check that this is equivalent to the MAP problem:

$$\hat{\alpha}_{\text{MAP}} = \arg\max_{\alpha} \log f(\alpha|y[n]) = \arg\max_{\alpha} \log f(y[n]|\alpha) + \log f(\alpha).$$

We note that in the above problem statement, we are assuming all values of α are equally likely. Hence, the problem can be rewritten as:

$$\hat{\alpha}_{\text{MAP}} = \arg\max_{\alpha} \log f(y[n]|\alpha) + c,$$

where c is a constant. When we differentiate, this constant will vanish and the maximization will be the same as in the ML case.

References

[1] S. Ross, *Introduction to Probability Models, Ninth Edition*, Academic Press, Inc, Orlando, FL, 2006.

[2] S. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*, Prentice Hall, Hoboken, NJ, 1993.