ECE 8110
Machine Learning
Exam 1 (Redo)
Vira Oleksyuk
03/16/2014

Problem#1

Consider two probability distributions defined by:

$$p(x|w_1) = \begin{cases} 1 & \alpha - \frac{1}{2} \le x \le \alpha + \frac{1}{2} \\ 0 & elsewere \end{cases}, \; p(x|w_2) = \begin{cases} 1 & -\frac{1}{2} \le x \le \frac{1}{2} \\ 0 & elsewere \end{cases}.$$

a) Sketch the probability of error, P(E), for a maximum likelihood classifier as a function of α and $p(w_1)$. Label all critical points.

Figure 1 shows graphical representation of distributions $p(x|w_1)$ and $p(x|w_2)$.

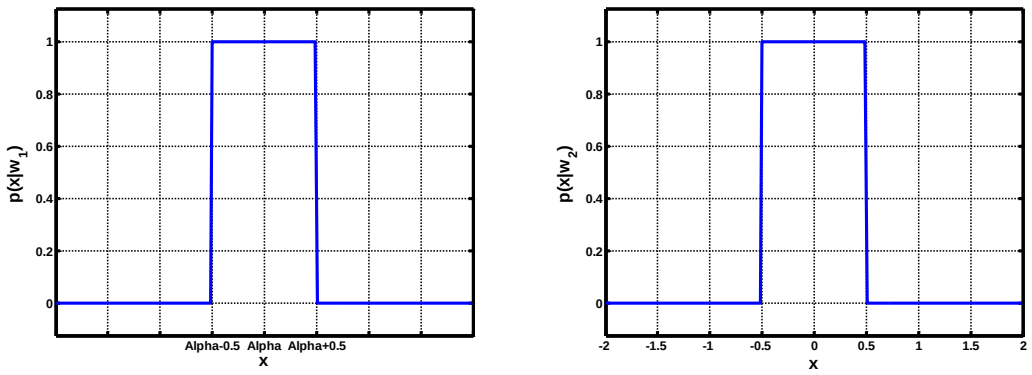


Figure 1. Given probability distributions of $p(x|w_1)$ and $p(x|w_2)$.

By performing convolution of those distributions we obtain following result (Figure 2):

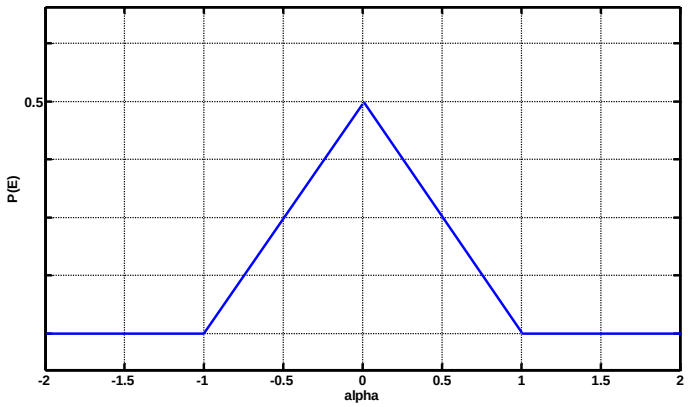


Figure 2. Probability of error for a maximum likelihood classifier.

Probability of error cannot exceed 0.5, which is the worst result from classifier. That is why the peak on Fig.2 is at 0.5.

We can support it by following equations.

Priors assumed to be equal (not specified) $P(w_1)=P(w_2)=0.5$.

$$P(error) = P(error)_{R1} + P(error)_{R2} = \int_{R1} P(error|x)p(x) + \int_{R2} P(error|x)p(x)$$

Bayes formula shows:

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

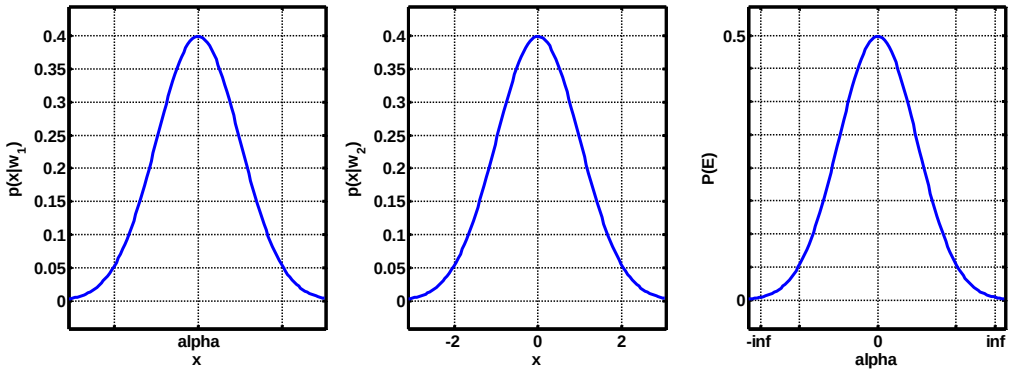
At the same time

$$P(error|x) = \begin{cases} P(w_1|x), & x \in w2 \\ P(w_2|x), & x \in w1 \end{cases}$$

So if we substitute everything in P(error), our result will graphically looks like Fig.2.

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

This problem may be seen as convolution of two Gaussians, which is also Gaussian. In part 1a) we were able to get total probability of error zero, where distributions did not overlap (-inf, -1] and [1, inf). We will not be able to have it in the case of Gaussian distributions. Values of Gaussian distributions go to zero at infinity.



(a) (b) (c)

Figure 3. Gaussian distributions $p(x|w_1)$ (a), and $p(x|w_2)$ (b), and total probability of error (c)

- The Gaussian distributions are considered as random variables. Mean and variance of the total error will have following characteristics $(\mu,\sigma^2) = (\mu_1+\mu_2,\sigma^2_{1+}\sigma^2_2)$ as supported by [3]

$$\sum_{i=1}^{n} Normal(\mu_i, \sigma_i^2) \sim Normal\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right), for\ (-\infty < \mu_i < \infty)\ and\ \ (\sigma_i^2 > 0\ )$$

Peak of P(E) also will not exceed 0.5 as well.

Problem#2

Let x have a uniform density:

$$P(x|\theta) = \begin{cases} 1/\theta, & 0 \le x \le \theta \\ 0, & otherwise \end{cases}$$

Suppose that n-samples D=$\{x_1, x_2, .., x_n\}$ are drawn independently from p(x| θ). Derive an expression for maximum likelihood estimate of θ with respect to data.

For total likelihood function we have to multiply every observed random sample likelihood function (when samples are drawn independently):

$$p(x1, x2, \dots, xn|\theta) = p(x1|\theta)p(x2|\theta) \dots p(xn|\theta) = \boldsymbol{p(D_n|\theta)}$$

$$\boldsymbol{p(D_n|\theta)} = \begin{cases} \frac{1}{\theta}\frac{1}{\theta}\dots\frac{1}{\theta}, & 0 \le x \le \theta \\ 0, & otherwise \end{cases} = \begin{cases} \frac{1}{\theta^n}, & 0 \le x \le \theta \\ 0, & otherwise \end{cases}$$

The resulting function is a decreasing function [2]. As a results, with increase of number of samples ($n \to \infty$), the likelihood function (MLE) will decrease. Which implies:

$$\theta = max(x1, x2, \dots, xn)$$

So the MLE of θ is [2]

$$\hat{\theta} = max(X1, X2, \dots, Xn)$$

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 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.

When we use maximum likelihood estimator, we consider parameters as unknown constants. The best estimate of their value is defined to be the one that minimizes the error and maximizes the probability of obtaining the samples actually observed [1].
We are looking for estimator, which is unbiased and has minimum variance [4]. In other words, we want its means converge to the true value, and variance converge to zero.
In Least Square Error estimator we are making no probabilistic assumptions [4] and approximating output from previous inputs and outputs.

In this problem, we assume input as a zero-mean Gaussian white noise, which is described as

$$s[n] = x[n] + e[n]$$

Write filter out

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

$$\frac{Y[z]}{X[z]} = \frac{1}{1 - \alpha z^{-1}}$$

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

$$y(n) = \alpha y(n-1) + x(n)$$

Input is a white noise with Gaussian distribution $N(0,1)$ with mean zero and unit variance. It will be denotes w(n). We can construct a square error solution

$$E = \sum_n [y(n) - \alpha y(n-1) + w(n)]^2$$

Rename signals as s(n) and s(n-1) and differentiate previous equation with respect to alpha.

$$\frac{dE}{d\alpha} = \frac{\sum_n [y(n) - \alpha y(n-1) + w(n)]^2}{d\alpha} = \frac{\sum_n [s^2(n) - 2\alpha * s(n)s(n-1) + \alpha^2 s^2(n-1)]}{d\alpha}$$

$$\frac{dE}{d\alpha} = \sum_n [0 - 2s(n)s(n-1) + 2\alpha s^2(n-1)] = 0$$

⇨

$$\sum_n s(n)s(n-1) = \alpha \sum_n s^2(n-1)$$

⇨

$$\alpha = \frac{\sum_n s(n)s(n-1)}{\sum_n s^2(n-1)} = \frac{c(1,0)}{c(1,1)} = \frac{r(1)}{r(0)}$$

Correlation of signals shows degree of interdependence of those signals [5]. Also correlation is able to extract important information about a system. If a signal is correlated with itself, it is autocorrelation. Autocorrelation uncovers energy of the signal and neglects its phase shift.

From the last equation we can see that coefficient of the filer can be found from autocorrelations of the outputs.

We are assuming that our samples were taken independently and were drawn from the normal distribution $N(0,1)$. Mat the same time, maximum likelihood estimates most probable future values from the sample data. Likelihood function will look like following [7]:

$$p(D_n|\theta) = p(x1, x2, \dots, xn|\theta) = p(x1|\theta)p(x2|\theta) \dots p(xn|\theta) = \prod_n p(X_k|\theta)$$

In our case. $p(X_k|\theta) \sim N(0,1)$ and the regression model is

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

Where ϵ_i,is an error, which is also normally distributed.

$$p(D_n|\theta) = \prod_n p(X_k|\theta) =$$

$$= \prod_n \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{1}{2\sigma^2}\sum(Y_i-\beta_0-\beta_1X_i)^2\right)} =$$

$$= \frac{1}{\sqrt{(2\pi\sigma^2)^n}} e^{\left(-\frac{1}{2\sigma^2}\sum(Y_i-\beta_0-\beta_1X_i)^2\right)} =$$

In order to maximize the likelihood function it is mathematically easier to work with log likelihood.

$$\ln\big(p(D_n|\theta)\big) = -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2}\sum(Y_i - \beta_0 - \beta_1 X_i)^2$$

Next, we differentiate with respect to β_0 and β_1.

$$\frac{\partial \ln\big(p(D_n|\theta)\big)}{\partial \beta_1} = -2\frac{1}{2\sigma^2}\sum(Y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0$$

$$\frac{\partial \ln\big(p(D_n|\theta)\big)}{\partial \beta_0} = -2\frac{1}{2\sigma^2}\sum(Y_i - \beta_0 - \beta_1 X_i)(-1) = 0$$

⇨ (Hats denote MLE)

$$\sum Y_i = n\bar{\beta_0} + \bar{\beta_1}\sum X_i$$

$$\sum X_i Y_i = \overline{\beta_0} \sum X_i + \bar{\beta}_1 \sum X_i^2$$

So those normal equation show, MLE is the same as LS estimator, as I specified before.

References

[1] R.O. Duda, P.E. Hart, and D. G. Stork, "Pattern Classification," 2nd ed., pp., New York : Wiley, 2001.

[2] Songfeng Zheng, Maximum Likelihood Estimation, http://people.missouristate.edu/songfengzheng/Teaching/MTH541/Lecture%20notes/MLE.pdf

[3] Hogg, Robert V.; McKean, Joseph W.; Craig, Allen T. (2004). *Introduction to mathematical statistics* (6th ed.). Upper Saddle River, New Jersey: Prentice Hall. http://en.wikipedia.org/wiki/List_of_convolutions_of_probability_distributions

[4] http://www.scribd.com/doc/99123544/Fundamentals-of-Statistical-Signal-Processing-Estimation-Theory

[5] http://www.ee.up.ac.za/main/_media/en/undergrad/subjects/esp411/esp411_lecture13.pdf

[6] Larry Rabiner, Linear Predictive Coding (LPC) Introduction Lecture 13 http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/digital%20speech%20processing%20course/lectures_new/Lecture%2013_winter_2012_6tp.pdf

[7] Chapter 1 – Linear Regression with 1 Predictor