Analytic Exercises 1

Pattern Recognition
NTUA 9th Semester

Efthymios Tzinis 03112007

November 17, 2017

Exercise 1.

a)

We divide the general proof in the forward and reverse steps:

Forward: We use the basic bayessian property: $P(X,Y) = P(X|Y)P(Y)$

$$P(C|AB) = P(C|B) \Rightarrow P(C|AB)P(AB) = P(C|B)P(AB) \Rightarrow$$

$$\Rightarrow P(ABC) = P(A|B)P(B)P(C|B) \Rightarrow P(AC|B)P(B) = P(A|B)P(B)P(C|B) \Rightarrow$$
$$Thus: \; P(AC|B) = P(A|B)P(C|B)$$

Reverse:

$$P(AC|B) = P(A|B)P(C|B) \Rightarrow P(AC|B)P(B) = P(A|B)P(B)P(C|B) \Rightarrow$$

$$\Rightarrow P(ABC) = P(AB)P(C|B) \Rightarrow P(C|AB)P(AB) = P(C|B)P(AB) \Rightarrow$$
$$Thus: \; P(C|AB) = P(C|B)$$

b)

We take as a fact: $a \perp b, c|d \Leftrightarrow P(abc|d) = P(a|d)P(bc|d)$
We integrate these probabilities with respect to the variable c:

$$\int_c P(abc|d) = \int_c P(a|d)P(bc|d)$$

$$Thus, P(ab|d) = \int_c P(a|d)P(b|d) \Leftrightarrow a \perp b|d$$

Exercise 2

We suppose that a-priori probabilities for the 2 classes are equal:

$$P(x \in 1) = P(x \in -1) = \frac{1}{2}$$

a)

The classifier which is based on the algorithm $\mathcal{A}_1$ has the following Confusion Matrix on a probabilistic perspective:

Actual vs Estimated	-1	1
-1	1	0
1	$\frac{1}{3}$	$\frac{2}{3}$

Algorithm $\mathcal{A}_1^{'}$ takes advantage of the algorithm $\mathcal{A}_1$ which classifies all instances of $x \in -1$ and thus, the only wrong estimated instances could be found only in class $\mathcal{C} = 1$. So the algorithm does the following steps iteratively:

- **Step 0:** $Dataset = InitialDataset$
- **Step 1:** Classify the data set according to the classifier of algorithm $\mathcal{A}_1$.
- **Step 2:** $Dataset = Instances\ Classified\ in\ class\ \mathcal{C} = -1$
- **Step 3:** If $(P(error) > \frac{1}{100})$ Go to **Step 1**.
- **Step 4:** Return

By this Procedure, our algorithm can reduce its error with every iteration. In particular, suppose the initial data set consists of **D** samples. In the first iteration, the error is $P(error)^{(1)} = \frac{1}{3}P(x \in 1) = \frac{1}{6}$. In other words, $\frac{D}{6}$ misclassified instances. In the second one, the data set consists of $\frac{4D}{3}$ samples, where the $\frac{D}{3}$ belong to class $\mathcal{C} = 1$. By the end of this cycle the error would be: $P(error)^{(2)} = \frac{1}{3}\frac{1}{3}P(x \in 1) = \frac{1}{12}$ In other words, $\frac{D}{12}$ misclassified instances. by induction we can derive the general form of the error according to the iterations of our algorithm. In general, the error is given by the following equation:

$$P(error)^{(n)} = \frac{\frac{D}{2*(3)^n}}{D} = \frac{1}{2}\frac{1}{3^n}$$

In order to achieve an error smaller than 1% we would need:

$$P(error)^{(n)} < \frac{1}{100} \Rightarrow \frac{1}{2}\frac{1}{3^n} < \frac{1}{100} \xRightarrow{n \in \mathbb{N}} n = 4\ Iterations$$

b)

Following the previous contemplation, the respective confusion matrix from probabilistic scope now becomes:

Actual vs Estimated	-1	1
-1	$\frac{2}{3}$	$\frac{1}{3}$
1	$\frac{1}{3}$	$\frac{2}{3}$

The algorithm $\mathcal{A}_2^{'}$ is called recursively by dividing in every iteration the data in respect with its estimations described by the matrix above. Consequently, after some iterations our algorithm will construct a tree where every level conducts a smaller total error than the previous level. In the level of the leaves of the tree, $\mathcal{A}_1$ is used in order to count the total misclassified samples. In every level we open **only** the left most and right most sub trees with the recursive form of our algorithm when the columns derived by the 2 new leafs are feeded in the initial algorithm $\mathcal{A}_1$. So the total depth of our tree is the only important parameter of our algorithm. The steps of the algorithm are described below (We suppose that $\mathcal{A}_1$ returns the number of data samples which were misclassified):

- **Step 0:** $Dataset = InitialDataset$ and $Depth = n$
- **Step 1:** Classify the data set according to the classifier of algorithm $\mathcal{A}_1$.
- **Step 2:** $DA = Instances\ Classified\ in\ class\ \mathcal{C} = -1$ and $DB = Instances\ Classified\ in\ class\ \mathcal{C} = 1$

- **Step 3:** $Missclassified\ samples = \mathcal{A}_2(DA, n-1) + \mathcal{A}_2(DB, n-1) + \mathcal{A}_1(inside\ columns)$
- **Step 3:** If $(P(error) = \frac{Missclassified\ samples}{InitialDataset} > \frac{1}{100})$ Set the $Depth = n{+}1$ and Go to **Step 1**.
- **Step 4:** Return

Same as the first inquiry, the induce will give us the general form of the misclassified instances and thus the total P(error) for every level. For example, for Depth $=1$ we will get $P(error) = P(x \in 1)\frac{1}{3} + P(x \in -1)\frac{1}{3} = \frac{1}{3}$. For Depth $=2$ we expand the tree 1 level lower by splitting the data. So in the next level we will get for the left sub-tree $P(x\ classified\ as 1, x \in -1) = \frac{1}{2}\frac{1}{3}\frac{1}{3} = \frac{1}{2}\frac{1}{9}$, for the right sub-tree $P(x\ classified\ as - 1, x \in 1) = \frac{1}{2}\frac{1}{3}\frac{1}{3} = \frac{1}{2}\frac{1}{9}$ and for the middle derived columns forwarded through the first algorithm we have $P(x\ classified\ as{-1}, x \in 1) = \frac{1}{2}\frac{1}{3}\frac{2}{3}\frac{1}{3} = \frac{1}{2}\frac{2}{27}$ and $P(x\ classified\ as 1, x \in -1) = \frac{1}{2}\frac{1}{3}\frac{2}{3}\frac{1}{3} = \frac{1}{2}\frac{2}{27}$. As a result the total error would be: $P(error) = \frac{1}{2}(2 * \frac{1}{9} + 2 * \frac{2}{27}) = \frac{5}{27} < \frac{1}{3}$. By induction we can conclude to the following formula for total error:

$$P(error)^{(n)} = \frac{1}{2} * \frac{\sum_{i=1}^{n} 2^i}{3^n}$$

For depth = n = 12 we get the demanded error.

Exercise 3.

We suppose that: $P(\omega_1) = P(\omega_2) = \frac{1}{2}$. Let x^* to be the decision boundary which divides the whole space in regions $\mathcal{R}_1$ and $\mathcal{R}_2$ where $P(\mathbf{x}|\omega_1)P(\omega_1) > P(\mathbf{x}|\omega_2)P(\omega_2)$ and $P(\mathbf{x}|\omega_1)P(\omega_1) < P(\mathbf{x}|\omega_2)P(\omega_2)$ respectively.

a)

Let $P(x|\omega_1) \sim N(1,1)$ $P(x|\omega_2) \sim N(1,2)$ then $P(x|\omega_1) = \frac{1}{\sqrt{2\pi}} \exp^{-\frac{(x-1)^2}{2}}$, $P(x|\omega_2) = \frac{1}{\sqrt{8\pi}} \exp^{-\frac{(x-1)^2}{8}}$. For calculating the decision point x^* we have to solve the following equation:

$$P(x^*|\omega_1)P(\omega_1) = P(x^*|\omega_2)P(\omega_2) \Rightarrow \frac{1}{\sqrt{2\pi}} \exp^{-\frac{(x^*-1)^2}{2}} = \frac{1}{\sqrt{8\pi}} \exp^{-\frac{(x^*-1)^2}{8}} \Rightarrow$$

$$\Rightarrow \exp^{-\frac{(x^*-1)^2}{2}} = \frac{1}{2} \exp^{-\frac{(x^*-1)^2}{8}} \Rightarrow -\frac{(x^*-1)^2}{2} = ln(\frac{1}{2}) - \frac{(x^*-1)^2}{8} \Rightarrow$$

$$\Rightarrow (x^*-1)^2 = \frac{8ln2}{3} \Rightarrow x^* = 1 \pm \sqrt{\frac{8ln2}{3}}$$

From that we derive the aforementioned regions:

$$\mathcal{R}_1 = (1 - \sqrt{\frac{8ln2}{3}}, 1 + \sqrt{\frac{8ln2}{3}})\ and\ \mathcal{R}_2 = (-\infty, 1 - \sqrt{\frac{8ln2}{3}}) \cup (1 + \sqrt{\frac{8ln2}{3}}, \infty)$$

Consequently, the P(error) can be computed:

$$P(error) = \int_{\mathcal{R}_1} P(x|\omega_2)P(\omega_2)dx + \int_{\mathcal{R}_2} P(x|\omega_1)P(\omega_1)dx =$$

$$= \frac{1}{2}\int_{\mathcal{R}_1} \frac{1}{\sqrt{8\pi}} \exp^{-\frac{(x-1)^2}{8}} + \frac{1}{2}\int_{\mathcal{R}_2} \frac{1}{\sqrt{2\pi}} \exp^{-\frac{(x-1)^2}{2}}$$

We will use the function $Z = \frac{X-\mu}{\sigma}$ according to the Normal Distribution Tables to compute the integrals. We also know that:

$$\int_{\mathcal{R}_1} \frac{1}{\sqrt{8\pi}} \exp^{-\frac{(x-1)^2}{8}} = 1 - \int_{\mathcal{R}_2} \frac{1}{\sqrt{8\pi}} \exp^{-\frac{(x-1)^2}{8}}$$

And because in both distributions the mean values are equal to 1.

$$\int_{\mathcal{R}_2} \frac{1}{\sqrt{8\pi}} \exp^{-\frac{(x-1)^2}{8}} = 2\int_{-\infty}^{1-\sqrt{\frac{8ln2}{3}}} \frac{1}{\sqrt{8\pi}} \exp^{-\frac{(x-1)^2}{8}}$$

$$\int_{\mathcal{R}_2} \frac{1}{\sqrt{2\pi}} \exp^{-\frac{(x-1)^2}{2}} = 2\int_{-\infty}^{1-\sqrt{\frac{8ln2}{3}}} \frac{1}{\sqrt{2\pi}} \exp^{-\frac{(x-1)^2}{2}}$$

Thus, we acquire from the tables the values and the total error would be:

$$P(error) \approx \frac{1}{2}(1 - 2 * 0.36) + \frac{1}{2}(2 * 0.23) = 0.37$$

b)

The distributions now are in 2 dimensional space but because of their form the decisions boundaries are easily calculated. It is clear that the only Region where the two distributions overlap is: $\mathcal{R} = 1 < x_1 < 2,\ 3 < x_2 < 5$. In this region it is obvious that the second distribution is dominant. With the same reasoning as before, we derive that the total error would be:

$$P(error) = \int_{\mathcal{R}} P(x1, x2|\omega_1)P(\omega_1)dx_1dx_2 = 0.1 * 0.5 * 1 * 2 = 0.1$$

Exercise 4.

a)

According to the Bayes decision theory error probability is derived from the correct probability, assuming that the decision boundary (point in this case) is called θ:

$$P(error) = 1 - P(correct) = 1 - \int_{\theta}^{\infty} P(x|\omega_1)P(\omega_1)dx - \int_{-\infty}^{\theta} P(x|\omega_2)P(\omega_2)dx \Rightarrow$$

$$\Rightarrow P(error) = P(\omega_1) + P(\omega_2) - P(\omega_1)\int_{\theta}^{\infty} P(x|\omega_1)dx - P(\omega_2)\int_{-\infty}^{\theta} P(x|\omega_2)dx \Rightarrow$$

$$\Rightarrow P(error) = P(\omega_1)\int_{-\infty}^{\theta} P(x|\omega_1)dx + P(\omega_2)\int_{\infty}^{\theta} P(x|\omega_2)dx$$

because a-priori probabilities are constants and $\int_{-\infty}^{\infty} P(x|\omega_2)dx = \int_{-\infty}^{\infty} P(x|\omega_1)dx1$.

b)

For calculating the decision boundary which is the optimal decision for minimizing the probability error we find the roots of the derivative.

$$\frac{dP(error)}{d\theta} = 0 \Rightarrow P(\omega_1)P(\theta|\omega_1) - P(\omega_2)P(\theta|\omega_2) = 0 \Rightarrow$$

$$P(\omega_1) * P(\theta|\omega_1) = P(\omega_2) * P(\theta|\omega_2)$$

c)

c) In general, the above equation could have many additional roots and in N-dimensional spaces where N¿2 the decision boundaries could be hyperplanes. But even in the 1 dimensional space there could be more than one solutions to this equations and consequently it could not imply individualistic values for the parameter θ. If we could somehow obtain that the pattern of the 2 distributions had only one common point then and only then θ would be unique.

d)

d) There is a possibility that if we force the decision to change when it passes through different regions divided by decision points even if the probability of the a-posteriori of the first distribution is greater than the other.

A more tangential example would be to choose as a distribution a dirac distribution where $P(x|\omega_1) = \delta(x_0) = 1$ then if we have another arbitrary distribution $P(x|\omega_2)$ and choose apt a-priori probabilities, the only point that the above equation could be true is $\theta = x_0$. If we choose this point then every time we could sum to the probability error the value of $P(x|\omega_1)P(\omega_1) = P(\omega_1)$ but if we choose another point then we could acquire the minimum error. If the argument $P(\omega_1) > \int_{\mathcal{R}_1} P(x|\omega_2)P(\omega_2)dx$ is valid.

Exercise 5.

Assuming that $P(x|\omega_i) \sim \mathcal{N}(\mu_i, \Sigma),\ i = 1, 2$ where Σ is common for the two distributions and is symmetric. Then the decision boundary equation would be:

$$\frac{P(\omega_1)}{\sqrt{2\pi|\Sigma|}} e^{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)} = \frac{P(\omega_2)}{\sqrt{2\pi|\Sigma|}} e^{-\frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)} \Rightarrow$$

$$ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) = \frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - \frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2) \Rightarrow$$

$$ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) = \frac{1}{2}\left(\mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2 + x^T\Sigma^{-1}\mu_1 - \mu_1^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_2 + \mu_2^T\Sigma^{-1}x\right) \Rightarrow$$

$$\xrightarrow[for\ arbitrary\ a,b\ a^T\Sigma^{-1}b=b^T\Sigma^{-1}a]{\Sigma\ Symmetric\ then\ \Sigma^{-1}\ Symmetric} ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) = \frac{1}{2}\left(\mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2 - 2\mu_1^T\Sigma^{-1}x + 2\mu_2^T\Sigma^{-1}x\right)$$

$$Finally,\ ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) = \frac{1}{2}\left(\mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2 - 2\left((\mu_1-\mu_2)^T\Sigma^{-1}x\right)\right)$$

Considering that vectors μ_1, μ_2 and $x \in \mathbb{R}^n$ then the right part of the above equation represents a hyperplane in $\mathbb{R}^n$. This hyperplane is also the decision boundary of our classifier. In order to attain that this boundary does not cross in the middle of the mean values vectors we have to check if both vectors μ_1, μ_2 lie on the same side of the hyperplane. Because the general form of a hyperplane equation is: $w^T x + w_0 = 0$ in our case we simply derive that:

$$w_0 = -\ ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) + \frac{1}{2}\left(\mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2\right),\ w = -(\mu_1-\mu_2)^T\Sigma^{-1}x$$

In other words, we demand either

$$w^T\mu_1 + w_0 < 0\ and\ w^T\mu_2 + w_0 < 0 \tag{1}$$

$$w^T\mu_1 + w_0 > 0\ and\ w^T\mu_2 + w_0 > 0 \tag{2}$$

These two conditions are translated into a more proper form, considering the hyperplane:

$$1 \Rightarrow ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) > -\frac{1}{2}\left((\mu_2-\mu_1)^T\Sigma^{-1}(\mu_2-\mu_1)\right) \tag{3}$$

$$1 \Rightarrow and\ ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) > \frac{1}{2}\left((\mu_2-\mu_1)^T\Sigma^{-1}(\mu_2-\mu_1)\right) \tag{4}$$

$$2 \Rightarrow ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) < -\frac{1}{2}\left((\mu_2-\mu_1)^T\Sigma^{-1}(\mu_2-\mu_1)\right) \tag{5}$$

$$2 \Rightarrow and\ ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) < \frac{1}{2}\left((\mu_2-\mu_1)^T\Sigma^{-1}(\mu_2-\mu_1)\right) \tag{6}$$

From equations 3, 4, 5, 6 we conduct that the inequality that derives is the following:

$$\left|ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right)\right| > \frac{1}{2}\left|(\mu_2-\mu_1)^T\Sigma^{-1}(\mu_2-\mu_1)\right| = \frac{1}{2}Mahalanobis\_Distance(\mu_1,\mu_2)$$

Exercise 6.

a)

In figure 1 we demonstrate the Exponential Distribution for $\theta = 2$.

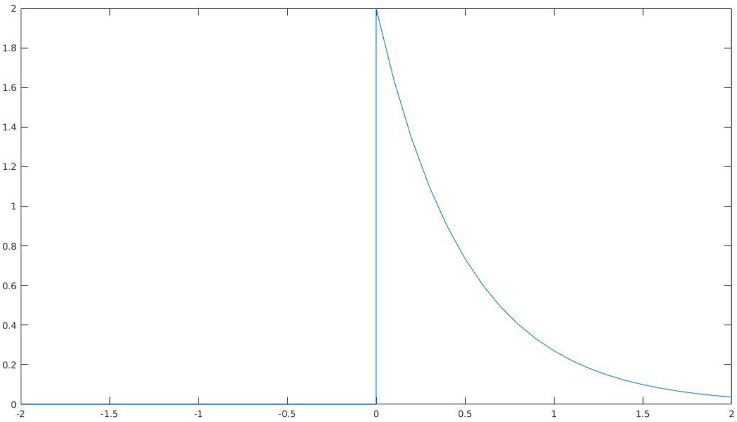

Figure 1: Exponential Distribution for $\theta = 2$

b)

According to the Maximum Likelihood Theory, the ideal estimation for parameter θ is given from the maximization of the a-posteriori probability. Instead of maximizing the a-posteriori itself we can try to maximize the logarithm of this value. Thus:

$$\hat{\theta} = argmax\left[\sum_{i=1}^{N} lnP(x_i|\theta)\right] = argmax\left(N * ln\theta - \theta\sum x_i\right)$$

From finding the root of the derivative we take:

$$\frac{d\left(N * ln\theta - \theta\sum x_i\right)}{d\theta} = 0 \Rightarrow \frac{N}{\hat{\theta}} - \sum_{i=1}^{N} x_i = 0 \Rightarrow \hat{\theta} = \frac{1}{\frac{1}{N}\sum_{i=1}^{N} x_i}$$

c)

From the previous analysis: $\frac{1}{\hat{\theta}} = \frac{\sum_{i=1}^{N} x_i}{N}$ However, every x_i follows the exponential distribution but exponential distribution is a Gamma distribution with respective parameters. In our case: $x_i \sim \Gamma(1, \frac{1}{\theta}),\ i = 1, 2, ..., N$ The probabilistic distribution function for Gamma distribution is given by: $\Gamma(a, b) = \frac{x^{a-1}e^{-\frac{x}{b}}}{b^a\Gamma(a)}$ It can be easily induced that: $\Gamma(a_1, b) + \Gamma(a_2, b) = \Gamma(a_1 + a_2, b)$ In this way we can use it for all our variables. Thus:

$$\frac{1}{\hat{\theta}} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}\Gamma\left(N, \frac{1}{\theta}\right)$$

d)

$$E(\hat{\theta}) = E\left(\frac{1}{\frac{1}{\hat{\theta}}}\right) \overset{ByDefinition}{=} N\int_{-\infty}^{\infty}\frac{\Gamma(N, \frac{1}{\theta})}{x}dx = N\int_{-\infty}^{\infty}\frac{\theta^N x^{N-1}e^{x\theta}}{x\Gamma(N)}dx =$$

$$= N \int_{-\infty}^{\infty} \frac{\theta^{N-1} x^{N-2} e^{x\theta}}{\Gamma(N-1)} \frac{\Gamma(N-1)}{\Gamma(N)} \theta dx = \frac{N\theta}{N-1} \int_{-\infty}^{\infty} \frac{\theta^{N-1} x^{N-2} e^{x\theta}}{\Gamma(N-1)} dx =$$

$$= \frac{N\theta}{N-1} \int_{-\infty}^{\infty} \Gamma\left(N-1, \frac{1}{\theta}\right) dx \xRightarrow{DistributionIntegral} E(\hat{\theta}) = \frac{N\theta}{N-1}$$

Obviously, our estimator is not indifferent but when the number of the consisting variables approach to infinity then our estimator becomes totally unbiased:

$$\lim_{N\to\infty} E(\hat{\theta}) = \lim_{N\to\infty} \frac{N\theta}{N-1} \overset{DLH}{=} \theta$$

We know that:

$$E(\hat{\theta}^2) = E\left(\frac{1}{\frac{1}{\hat{\theta}^2}}\right) = N^2 \int_{-\infty}^{\infty} \frac{\Gamma(N, \frac{1}{\theta})}{x^2} dx = N^2 \int_{-\infty}^{\infty} \frac{\theta^N x^{N-1} e^{x\theta}}{x^2 \Gamma(N)} dx = \frac{N^2\theta^2}{(N-1)(N-2)}$$

The variance of our estimator is:

$$Var(\hat{\theta}) = E(\hat{\theta}^2) - E(\hat{\theta})^2 = \frac{N^2\theta^2}{(N-1)(N-2)} - \frac{N^2\theta^2}{(N-1)^2} = \frac{N^2\theta^2}{(N-1)^2(N-2)}$$

Exercise 7.

a)

According to maximum likelihood theory:

$$\hat{\theta} = argmax \left[\sum_{i=1}^{N} lnP(x_i|\theta)\right] = argmax\,(-Nln\theta) \Rightarrow \hat{\theta} = max(x_i)$$

Because the function $-Nln\theta$ is descending and thus the maximum value of it is the same as the minimum value of θ. In this way the only way to achieve this, is: $\theta > max(x_i),\ i = 1, 2, ..., N$.

b)

According to the previous statement for $\hat{\theta}$:

$$P(Y \le x) = P(max(x_i) \le x) = P(x_1 \le x, ..., x_N \le x) \xRightarrow{x_i iid}$$

$$\Rightarrow P(Y \le x) = \prod_{i=1}^{N} P(x_i \le x) = \prod_{i=1}^{N} \frac{x}{\theta} = \left(\frac{x}{\theta}\right)^n$$

The probabilistic distribution function would be:

$$\frac{dP(Y \le x)}{dx} = n\left(\frac{x}{\theta}\right)^{n-1}$$

Of course this is a Betta Distribution with parameters (N,1) where:

$$B(x,a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)})\ andB(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

In this way the expected value of our estimator is:

$$E(\hat{\theta}) = \int_{-\infty}^{\infty} N\left(\frac{x}{\theta}\right)^{N-1} xdx = \int_{-\infty}^{\infty} (N+1)\left(\frac{x}{\theta}\right)^{N} \frac{N\theta}{N+1}dx$$

$$E(\hat{\theta}) = \frac{N\theta}{N+1}\int_{-\infty}^{\infty} Beta(N+1,1)dx = \frac{N\theta}{N+1}$$

Of course, our estimator is not unbiased because $E(\hat{\theta}) \neq \theta$. But the estimator is asymptotically unbiased and the proof is the same as the previous exercise:

$$\lim_{N\to\infty} E(\hat{\theta}) = \lim_{N\to\infty} \frac{N\theta}{N+1} \overset{DLH}{=} \theta$$

Exercise 8

According to the same contemplation of the maximum likelihood theory we can find the best estimation for q by the maximization of a-posteriori probability.

$$\hat{q} = argmax\left(\prod_{i=1}^{N} P(x_i|q)\right) = argmax\left(q^{\sum_{i=1}^{N} x_i}(1-q)^{N-\sum_{i=1}^{N} x_i}\right)$$

$$Let\ f(q) = q^{\sum_{i=1}^{N} x_i}(1-q)^{N-\sum_{i=1}^{N} x_i}\ Therefore:$$

$$\frac{df(q)}{dq} = \sum_{i=1}^{N} x_i q^{\sum_{i=1}^{N} x_i - 1}(1-q)^{N-\sum_{i=1}^{N} x_i} - (N - \sum_{i=1}^{N} x_i) q^{\sum_{i=1}^{N} x_i}(1-q)^{N-\sum_{i=1}^{N} x_i - 1}$$

$$\frac{df(q)}{dq} = 0 \Rightarrow q^{\sum_{i=1}^{N} x_i}(1-q)^{N-\sum_{i=1}^{N} x_i}\left(\frac{\sum_{i=1}^{N} x_i}{q} - \frac{N-\sum_{i=1}^{N} x_i}{1-q}\right) = 0$$

Which is the equation that it was needed by definition.

Exercise 9.

a)

We will try to show that the $P'(error)$ in the (d-1) space will be always greater or equal to the respective error of the initial dimensional space and thus by induction we can show easily that generally, the error in a reduced feature space will be greater or equal from the derived one.

Assuming that we have the two distributions in a d-dimensional space have the following distributions functions: $P(x|\omega_1)\ and\ P(x|\omega_2)$ then the decisions boundaries can be found by solving the equation:

$$P(x|\omega_1)P(\omega_1) = P(x|\omega_2)P(\omega_2)$$

This equation will divide the d-dimensional into Regions $\mathcal{R}_1\ and\ \mathcal{R}_2$ where the first distribution is dominant and inverse. Moreover, the projection of a vector

x to a reduced dimensional space (without loss of generality we will assume in a d-1 space) is achieved by a projection through the k-dimension.

$$y = Ax, where\ A\ is\ the\ projection\ matrix$$

y has dimensions $(d-1)*1$, x has dimensions $d*1$ and the matrix A has dimensions $(d-1)*d$. The initial probabilistic error in d-dimensional space is:

$$P(error) = \int_{\mathcal{R}_1} P(x|\omega_2)P(\omega_2)dx + \int_{\mathcal{R}_1} P(x|\omega_1)P(\omega_1)dx$$

However, from the assumption we can acquire the new probability distributions in (d-1) space from the following equations:

$$P'(x|\omega_1) = AP(x|\omega_1)\ and\ P'(x|\omega_2) = AP(x|\omega_2)$$

But we can also see them as integrals through the dimension that is being reduced. Let k be this dimension. Then, we can express them as:

$$P'(x|\omega_1) = \int_k P(x|\omega_1)dx\ and\ P'(x|\omega_2) = \int_k P(x|\omega_2)dx$$

But the thing is that if we try to solve the new equation, in (d-1) projected space, for the decision boundaries we will see that the previous regions will probably change. Lets expand our intuition:

$$P'(x|\omega_1)P(\omega_1) = P'(x|\omega_2)P(\omega_2) \Rightarrow AP(x|\omega_1)P(\omega_1) = AP(x|\omega_2)P(\omega_2)$$

$$A\left(P(x|\omega_1)P(\omega_1) - P(x|\omega_2)P(\omega_2)\right) = 0$$

We can see that this is the system of a typical form $Ax = 0$. So if we exclude the cases where initially decided for the boundaries because the equation $P(x|\omega_1)P(\omega_1) = P(x|\omega_2)P(\omega_2)$ was true in d-dimensional space, we can see that maybe there are some other vectors that are solutions to the last equation. All vectors of the form $x_n = (P(x|\omega_1)P(\omega_1) - P(x|\omega_2)P(\omega_2))$ which $x_r \in Null\ Space(A)$ are solutions to the above decision problem. Of course these vectors could exist because the matrix of the projection is not full rank. In this way the decision regions will now convert to $\mathcal{R'}_1\ and\ \mathcal{R'}_2$. If the regions are exactly the same as before the error will not change but lets speculate for the purpose of our proof that they will change in a region $\mathcal{R}$.

Obviously, the new probabilistic error would be:

$$P'(error) = \int_{\mathcal{R'}_1} P'(x|\omega_2)P(\omega_2)dx + \int_{\mathcal{R'}_2} P'(x|\omega_1)P(\omega_1)dx$$

As a final step of our proof we can easily see that by integrating according to the reduced dimension we can obtain the initial P(error) but computed into different regions from the initial ones. In particular:

$$P'(error) = \int_{\mathcal{R'}_1}\int_k P(x|\omega_2)P(\omega_2)dx + \int_{\mathcal{R'}_2}\int_k P(x|\omega_1)P(\omega_1)dx \Rightarrow$$

$$P'(error) = \int_{\mathcal{R'}_1} P(x|\omega_2)P(\omega_2)dx + \int_{\mathcal{R'}_1} P(x|\omega_1)P(\omega_1)dx \geq$$

$$\int_{\mathcal{R}_1} P(x|\omega_2)P(\omega_2)dx + \int_{\mathcal{R}_1} P(x|\omega_1)P(\omega_1)dx = P(error)$$

By dividing the lower dimensional space into different boundaries because of the Null Space Vectors of the projection matrix then we obligate the error to be integrated through different decision regions and conclusively to be greater or equal than the initial one.

b)

Although, as it is proven above the probabilistic error of the reduced dimensional space is always greater or equal to the respective error of the initial dimensional space, sometimes this reduction is very useful and other times is a necessity. For example, in many cases the data we have in our training system are ample and thus the computations are very complicated and time consuming. By projecting the data and representing them in a lower dimensional feature space we simplify the complexity of our system and also keep only the valuable characteristics (those with high variance). Moreover, it is widely known that in cases of little data, features that are redundant will conclude to a lower classification score than the classifier in the reduced space.