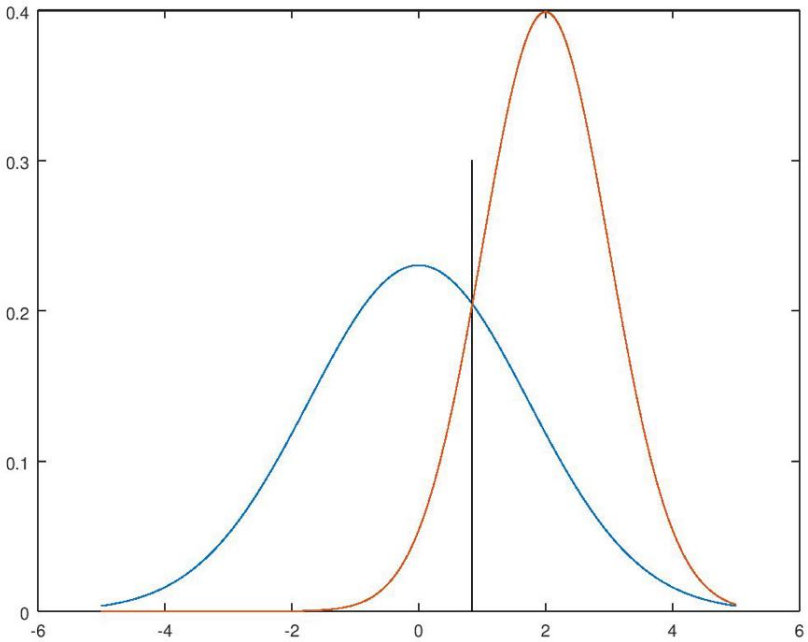
Questions

Answer 1



a) The likelihood ratio is $\frac{P(x|w_1)}{P(x|w_2)}$ which in this case becomes $\frac{1}{\sqrt{3}}e^{\frac{2x^2-12x+12}{6}}$

b) The Bayes Decision rule for this specific is quite trivial given that the lambdas basically say that the risk given for class 1 is just the likelihood of class 2 and vice versa. So the boundary becomes the point were the likelihood ratio is equal to 1, which is as plotted .84410 .

c) The error probability is basically $P(error \mid x) = P(w_2 \mid x)P(w_1) + P(w_1 \mid x)P(w_2)$ which simplifies into $P(error \mid x) = \frac{(P(x|w_1)+P(x|w_1))P(w_1)P(w_2)}{P(x)}$

Answer 2

a) Since $P(w_1) + P(w_2) = 1$, it is given that $P(w_1) = \frac{1}{3}$.

$$P(w_1|x) = \frac{P(x|w_1)P(w_1)}{P(x)} = \frac{(0.6(0.6)^x(0.4)^{1-x} + 0.4(0.4)^x(0.6)^{1-x}).\frac{1}{3}}{P(x|w_1)P(w_1) + P(x|w_2)P(w_2)}$$

Now, let's also find $P(x|w_2)P(w_2)$.

$$P(x|w_2)P(w_2) = (0.4)^x(0.6)^{1-x}\frac{2}{3} = (\tfrac{2}{3})^x 0.4$$

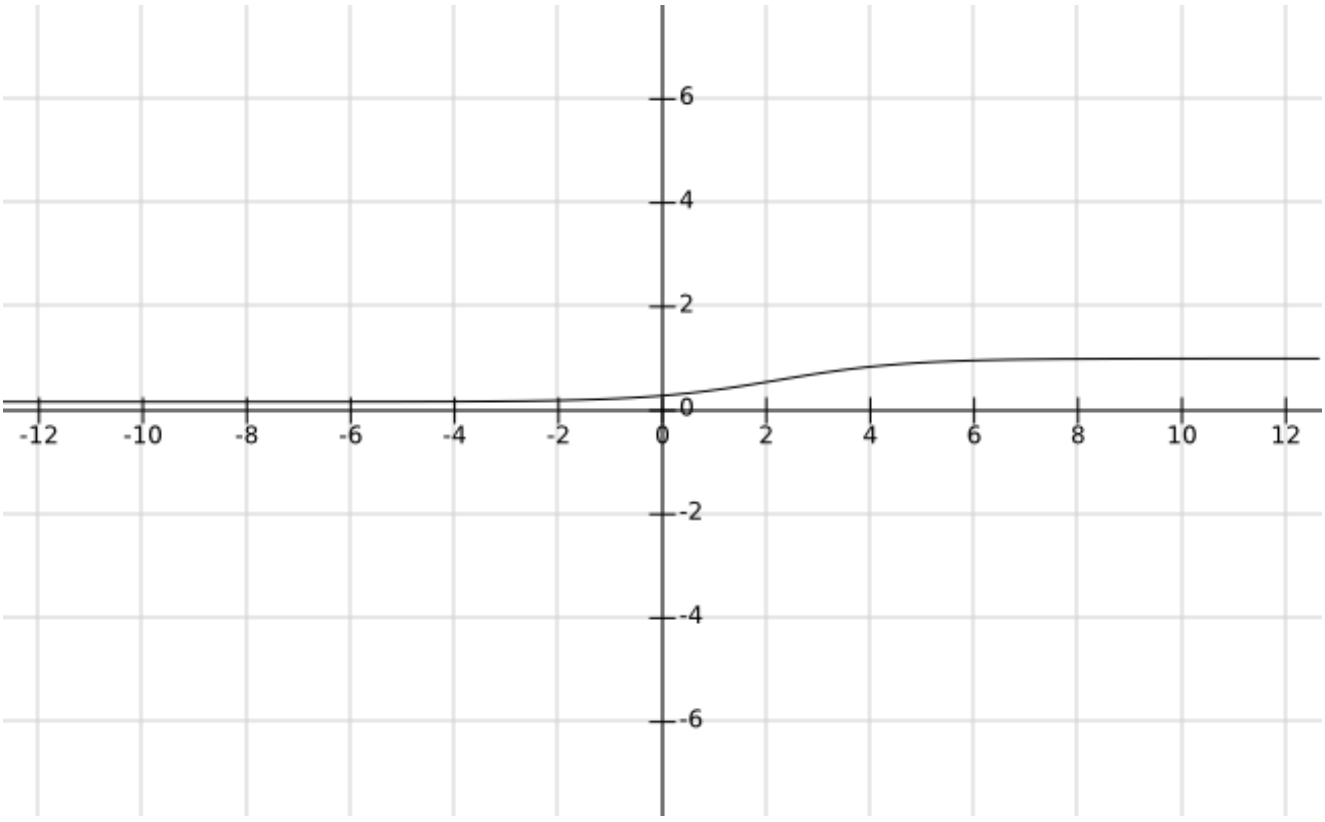
Continuing where we left off and doing some algebraic manipulation:

$$P(w_1|x) = \frac{0.08(\frac{3}{2})^x + 0.08(\frac{2}{3})^x}{0.08(\frac{3}{2})^x + 0.08(\frac{2}{3})^x + (\frac{2}{3})^x 0.4}$$

Then divide each term by 0.08 and multiply each term with $(\frac{3}{2})^x$.

$$P(w_1|x) = \frac{(\frac{9}{4})^x + 1}{(\frac{9}{4})^x + 6}$$

Does not look too bad. Now let's plot this:



While at it, let's also find:

$$P(w_2|x) = \frac{(\frac{2}{3})^x 0.4}{(0.6(0.6)^x(0.4)^{1-x} + 0.4(0.4)^x(0.6)^{1-x}).\frac{1}{3} + (\frac{2}{3})^x 0.4}$$

b) A basic decision strategy would be, choose w_1 is $P(w_1|x) > P(w_2|x)$, w_2 otherwise (We could also compare likelihood times prior, it would give the same result).

For $x = 0$:

$$P(w_1|x=0) = \frac{2}{7} = 0.29$$

$$P(w_2|x=0) = \frac{0.4}{0.2 \cdot 0.4 + 0.4 \cdot 0.2 + 0.4} = 0.71$$

So we choose w_2 (We didn't had to calculate the latter term, obviously, since the sum would have been 1 anyway, but no harm finding it explicitly).

c) $R(w_1|x=1) = \lambda_{11}P(w_1|x=1) + \lambda_{12}P(w_2|x=1) = P(w_2|x=1)$

$$= \frac{\frac{4}{5}}{0.6 \cdot 0.6 + 0.4 \cdot 0.4 + \frac{4}{5}} = 0.606$$

Answer 3

a)

$$P(w_1|\mathbf{x}) = \frac{P(\mathbf{x}|w_1)P(w_1)}{P(\mathbf{x})} = \frac{e^{\left(-\frac{1}{2}\left((0.3 \quad 0.3)-(0 \quad 0)\right)\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}^{-1}\left(\begin{pmatrix}0.3\\ 0.3\end{pmatrix}-\begin{pmatrix}0\\ 0\end{pmatrix}\right)\right)}}{\sqrt{(2\pi)^2\left|\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right|}} \cdot \frac{1}{3P(\mathbf{x})} =$$

$$0.12231141052 \cdot \frac{1}{3P(\mathbf{x})} = 0.04077047017 \cdot \frac{1}{P(\mathbf{x})}$$

$$P(w_2|\mathbf{x}) = \frac{P(\mathbf{x}|w_2)P(w_2)}{P(\mathbf{x})} = \frac{e^{\left(-\frac{1}{2}\left((0.3 \quad 0.3)-(1 \quad 1)\right)\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}^{-1}\left(\begin{pmatrix}0.3\\ 0.3\end{pmatrix}-\begin{pmatrix}1\\ 1\end{pmatrix}\right)\right)}}{\sqrt{(2\pi)^2\left|\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right|}} \cdot \frac{2}{3P(\mathbf{x})} =$$

$$0.08198779033 \cdot \frac{2}{3P(\mathbf{x})} = 0.05465852688 \cdot \frac{1}{P(\mathbf{x})}$$

$$P(w_3|\mathbf{x}) = \frac{P(\mathbf{x}|w_3)P(w_3)}{P(\mathbf{x})} = (\frac{e^{\left(-\frac{1}{2}\left(\begin{pmatrix}0.3 & 0.3\end{pmatrix}-\begin{pmatrix}0.5 & 0.5\end{pmatrix}\right)\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}^{-1}\left(\begin{pmatrix}0.3\\ 0.3\end{pmatrix}-\begin{pmatrix}0.5\\ 0.5\end{pmatrix}\right)\right)}}{\sqrt{(2\pi)^2\left|\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right|}}\cdot\frac{1}{2}+$$
$$\frac{e^{\left(-\frac{1}{2}\left(\begin{pmatrix}0.3 & 0.3\end{pmatrix}-\begin{pmatrix}-0.5 & 0.5\end{pmatrix}\right)\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}^{-1}\left(\begin{pmatrix}0.3\\ 0.3\end{pmatrix}-\begin{pmatrix}-0.5\\ 0.5\end{pmatrix}\right)\right)}}{\sqrt{(2\pi)^2\left|\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right|}}\cdot\frac{1}{2})\cdot\frac{3}{3P(\mathbf{x})} = (\frac{0.12858245064}{2}+$$
$$\frac{0.09525622229}{2})\cdot\frac{3}{3P(\mathbf{x})} = 0.11192\cdot\frac{1}{P(\mathbf{x})}$$

And $P(\mathbf{x}) = 0.04077047017 + 0.05465852688 + 0.11192 = 0.20734899705$

So we can see that, regardless of $P(\mathbf{x})$, the correct classification is w_3.

b)For the point (*,0.3) we just have to compare three likelihoods that are three normal distributions all with variance 1 and means 0 1 and 0.5 respectively. It is obvious that class 3 will yield the highest probability.

Answer 4

a) Let's first find the likelihood ratio, including the prior probabilities. Denote the ratio, which is a scalar, $R = R(x)$. Yet since prior probabilities are not given, we assume them to be 0.5 each.

$$R(x) = \frac{\frac{1}{\pi b}\cdot\frac{1}{1+(\frac{x-a_1}{b})^2}}{\frac{1}{\pi b}\cdot\frac{1}{1+(\frac{x-a_2}{b})^2}} = \frac{1+(\frac{x-a_2}{b})^2}{1+(\frac{x-a_1}{b})^2} = \frac{b^2+x^2+a_2^2-2xa_2}{b^2+x^2+a_1^2-2xa_1}$$

And since no loss funciton is given, we will assume unit-loss, which means we will decide w_1 when $R > 1$ and vice-versa. So our decision boundary is $R(x) = 1$.

$$R(x) = 1 = \frac{b^2+x^2+a_2^2-2xa_2}{b^2+x^2+a_1^2-2xa_1}$$

Or equivalently,

$b^2 + x^2 + a_2^2 - 2xa_2 = b^2 + x^2 + a_1^2 - 2xa_1$

$a_2^2 - 2xa_2 = a_1^2 - 2xa_1$

$x = \frac{a_2 + a_1}{2}$ (For $a_1 \neq a_2$)

is also the same decision boundary, independent of b, where x is a variable and a_1, a_2 are constants. And our rule is now decide w_1 when $\frac{a_1 + a_2}{2} > x$ and vice-versa (with the assumption that $a_2 > a_1$). Without loss of generality, we can simply reverse the cases if $a_1 > a_2$.

The reason for all these can be shown like this:

We decide w_1 when $R > 1$, or equivalently when:

$a_2^2 - 2xa_2 > a_1^2 - 2xa_1$

$a_2^2 - a_1^2 > 2xa_2 - 2xa_1$

$(a_2 + a_1)(a_2 - a_1) > 2x(a_2 - a_1)$

$a_2 + a_1 > 2x$ (For $a_2 > a_1$)

b)

$P(error) = \int_{-\infty}^{\infty} P(error|x)P(x)dx$

Where $P(error|x) = min[P(w_1|x), P(w_2|x)]$. So:

$P(error) = \int_{-\infty}^{\frac{a_1+a_2}{2}} P(w_2|x)P(x)dx + \int_{\frac{a_1+a_2}{2}}^{\infty} P(w_1|x)P(x)dx$

$= \int_{-\infty}^{\frac{a_1+a_2}{2}} P(x|w_2)P(w_2)dx + \int_{\frac{a_1+a_2}{2}}^{\infty} P(x|w_1)P(w_1)dx$

$= 0.5 \int_{-\infty}^{\frac{a_1+a_2}{2}} \frac{1}{\pi b} \cdot \frac{1}{1 + (\frac{x - a_2}{b})^2} dx + 0.5 \int_{\frac{a_1+a_2}{2}}^{\infty} \frac{1}{\pi b} \cdot \frac{1}{1 + (\frac{x - a_1}{b})^2} dx$

Which when put in to a calculator, it gets evaluated to the following:

$$\frac{0.5\arctan\left(\frac{a_1-a_2}{2b}\right)}{\pi} - \frac{0.5\arctan\left(\frac{-a_1+a_2}{2b}\right)}{\pi} + 0.5\text{sgn}\,(b)$$

Answer 5

Say we have a model such that $Y = aX + b$. Where b is the normally distributed error. When we take the log-likelihood of $Y \mid X$ we see that the value when we try to maximiz this expression w.r.t a it is the same as minimizing the error coefficient which turns out to be in the form of sum of squares.

Answer 6

Without loss of generality, assuming each $\boldsymbol{x}$ is consisted of 1s and 0s, it is easy to see that;

$P(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{d} \boldsymbol{\theta}_i^{\boldsymbol{x}_i} \cdot (1-\boldsymbol{\theta}_i)^{1-\boldsymbol{x}_i}$

Now, let $D = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$. The log-likelihood function L is:

$L(\boldsymbol{\theta}) = logP(D|\boldsymbol{\theta}) = log[P(\boldsymbol{x}_1|\boldsymbol{\theta}) \cdots P(\boldsymbol{x}_n|\boldsymbol{\theta})] = log \prod_{k=1}^{n} P(\boldsymbol{x}_k|\boldsymbol{\theta})$

$= \sum_{k=1}^{n} logP(\boldsymbol{x}_k|\boldsymbol{\theta}) = \sum_{k=1}^{n} log \prod_{i=1}^{d} \boldsymbol{\theta}_i^{\boldsymbol{x}_{\boldsymbol{k}i}} \cdot (1-\boldsymbol{\theta}_i)^{1-\boldsymbol{x}_{\boldsymbol{k}i}}$

$\sum_{k=1}^{n} \sum_{i=1}^{d} log[\boldsymbol{\theta}_i^{\boldsymbol{x}_{\boldsymbol{k}i}} \cdot (1-\boldsymbol{\theta}_i)^{1-\mathbf{x}_{\mathbf{k}i}}]$

$\sum_{k=1}^{n} \sum_{i=1}^{d} log[\boldsymbol{\theta}_i^{\boldsymbol{x}_{\boldsymbol{k}i}}] + log[(1-\boldsymbol{\theta}_i)^{1-\boldsymbol{x}_{\boldsymbol{k}i}}]$

$\sum_{k=1}^{n} \sum_{i=1}^{d} \boldsymbol{x}_{\boldsymbol{k}i} log[\boldsymbol{\theta}_i] + (1-\boldsymbol{x}_{\boldsymbol{k}i}) log[1-\boldsymbol{\theta}_i]$

Now, let's take derivative of $L(\boldsymbol{\theta})$ w.r.t $\boldsymbol{\theta}$. Since our L is not in vector notation, we can separately take partial derivative w.r.t to $\boldsymbol{\theta}_j$ and find their LMEs separately.

Notice that the inner summation's terms will be zero, when we take partial derivative, when $i \neq j$.

$$\frac{\varphi L}{\varphi \boldsymbol{\theta}_j} = \sum_{k=1}^{n} \boldsymbol{x}_{\boldsymbol{k}j} \frac{1}{\boldsymbol{\theta}_j} + (1-\boldsymbol{x}_{\boldsymbol{k}j}) \frac{-1}{1-\boldsymbol{\theta}_j} = 0$$
$$\sum_{k=1}^{n} \boldsymbol{x}_{\boldsymbol{k}j} \frac{1}{\boldsymbol{\theta}_j} = \sum_{k=1}^{n} (1-\boldsymbol{x}_{\boldsymbol{k}j}) \frac{1}{1-\boldsymbol{\theta}_j}$$
$$\sum_{k=1}^{n} \boldsymbol{x}_{\boldsymbol{k}j} \frac{1}{\boldsymbol{\theta}_j} = \sum_{k=1}^{n} \frac{1}{1-\boldsymbol{\theta}_j} - \frac{\boldsymbol{x}_{\boldsymbol{k}j}}{1-\boldsymbol{\theta}_j}$$

$\sum_{k=1}^{n} \boldsymbol{x_{kj}} - \boldsymbol{x_{kj}}\boldsymbol{\theta}_j = \sum_{k=1}^{n} \boldsymbol{\theta}_j - \boldsymbol{\theta}_j\boldsymbol{x_{kj}}$

$\sum_{k=1}^{n} \boldsymbol{x_{kj}} = \sum_{k=1}^{n} \boldsymbol{\theta}_j = n\boldsymbol{\theta}_j$

$\frac{1}{n}\sum_{k=1}^{n} \boldsymbol{x_{kj}} = \boldsymbol{\theta}_j$

is the best estimate for $\boldsymbol{\theta}_j$. Since the same can be shown for each j, we can generalize this to the following vector notation:

$\hat{\boldsymbol{\theta}} = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{x_k}$

Which was what wanted to be showed.

Answer 7

The given equation is just going to become $\frac{1}{n}(X-\mu)(X-\mu)^T$ which is basically just $\frac{1}{n}\begin{bmatrix} x_1-\mu_1 \\ \vdots \\ x_n-\mu_n \end{bmatrix}\begin{bmatrix} x_1-\mu_1 \cdots x_n-\mu_n \end{bmatrix}$ The result of this vector mıltiplication becomes a matrix where the i,jth entry is $x_i - \mu_i x_j - \mu_j$ which is basically $Cov(x_i, x_j)$ which means we have attained the covariance matrix.

Answer 8

Let us define $D = \{\mathbf{x_1}, \cdots \mathbf{x_n}\}$ where each $\mathbf{x_i} \in R^d$.

In a MAP we will try to maximize $P(D|\boldsymbol{\mu})P(\boldsymbol{\mu})$.

Note that:

$$P(\boldsymbol{x}|\boldsymbol{\mu}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T)\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}))$$

Also, let us define,

$$a = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}$$

Which is a scalar independent of $\boldsymbol{\mu}$

.

And the prior is:

$$P(\mu) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}_0|^{1/2}} exp(-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \mathbf{\Sigma}_0^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0))$$

Also, let us define,

$$b = \frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}_0|^{1/2}}$$

And in our book Pattern Classification section 3.2.2 equation 9, it is given that;

$$lnP(\mathbf{x}|\mu) = -\frac{1}{2} ln[(2\pi)^d |\Sigma|] - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})$$

The log-likelihood function L is:

$$L(\boldsymbol{\mu}) = ln(P(D|\boldsymbol{\mu})P(\boldsymbol{\mu})) = ln[P(\boldsymbol{x}_1|\boldsymbol{\mu}) \cdots P(\boldsymbol{x}_n|\boldsymbol{\mu})P(\boldsymbol{\mu})] = ln[(\textstyle\prod_{k=1}^n P(\boldsymbol{x}_k|\boldsymbol{\mu}))P(\boldsymbol{\mu})]$$

$$= \textstyle\sum_{k=1}^n lnP(\boldsymbol{x}_k|\boldsymbol{\mu})P(\boldsymbol{\mu})$$

$$= [\textstyle\sum_{k=1}^n ln(a) - \frac{1}{2}(\mathbf{x_k} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{x_k} - \boldsymbol{\mu})] + ln(b) - \frac{1}{2}(\mu - \boldsymbol{\mu}_0)^T \mathbf{\Sigma}_0^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)$$

Now, let's take derivative of L and equate to zero:

$$\frac{dL}{d\boldsymbol{\mu}} = [\textstyle\sum_{k=1}^n \mathbf{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})] - \mathbf{\Sigma}_0^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0) = 0$$

$$\textstyle\sum_{k=1}^n \mathbf{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}) = \mathbf{\Sigma}_0^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)$$

$$\textstyle\sum_{k=1}^n \mathbf{\Sigma}^{-1}\mathbf{x}_k - n\mathbf{\Sigma}^{-1}\mu = \mathbf{\Sigma}_0^{-1}\boldsymbol{\mu} - \mathbf{\Sigma}_0^{-1}\boldsymbol{\mu}_0$$

$$\mathbf{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \mathbf{\Sigma}^{-1}\textstyle\sum_{k=1}^n \mathbf{x}_k = \mathbf{\Sigma}_0^{-1}\boldsymbol{\mu} + n\mathbf{\Sigma}^{-1}\mu = (\mathbf{\Sigma}_0^{-1} + n\mathbf{\Sigma}^{-1})\mu$$

So;

$$\hat{\mu} = \frac{\mathbf{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \mathbf{\Sigma}^{-1}\sum_{k=1}^n \mathbf{x}_k}{(\mathbf{\Sigma}_0^{-1} + n\mathbf{\Sigma}^{-1})}$$

Is our MAP estimator.

b) First let's find the transformed values;
$\boldsymbol{\mu}' = E(\mathbf{x}') = E(\mathbf{A}\mathbf{x}) = \mathbf{A}E(\mathbf{x}) = \mathbf{A}\boldsymbol{\mu}$

and

$$\mathbf{\Sigma}' = E[(\mathbf{x}' - \boldsymbol{\mu}')(\mathbf{x}' - \boldsymbol{\mu}')^t]$$

$= E[(A\mathbf{x} - A\boldsymbol{\mu})(A\mathbf{x} - A\boldsymbol{\mu})^t]$

$= E[\mathbf{A}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t\mathbf{A}^t]$

$= \mathbf{A}E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t]\mathbf{A}^t$

$= \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^t$

Now,even tough the question lacks a definition for what is an "appropriate estimate", we can try recalculating MAP for the transformed space. We need to find $P(\mathbf{x}'|\mu')$ and a new prior $P(\mu')$

$$P(\boldsymbol{x'}|\boldsymbol{\mu'}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma'}|^{1/2}} exp(-\frac{1}{2}(\mathbf{x}' - \mu')^T)\boldsymbol{\Sigma'}^{-1}(\mathbf{x}' - \mu'))$$
$$= \frac{1}{(2\pi)^{d/2}|\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^t|^{1/2}} exp(-\frac{1}{2}(\mathbf{A}\mathbf{x} - \mathbf{A}\mu)^T(\mathbf{A}^t)^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{A}^{-1}(\mathbf{A}\mathbf{x} - \mathbf{A}\boldsymbol{\mu}))$$
$$= \frac{1}{(2\pi)^{d/2}|\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^t|^{1/2}} exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T\mathbf{A}^t(\mathbf{A}^t)^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{A}^{-1}\mathbf{A}(\mathbf{x} - \mu))$$
$$= \frac{a}{|\mathbf{A}\mathbf{A}^t|^{1/2}} exp(-\frac{1}{2}(\mathbf{x}' - \boldsymbol{\mu'})^T)\boldsymbol{\Sigma'}^{-1}(\mathbf{x}' - \boldsymbol{\mu'}))$$

Remember how we defined a from part (a). The derivation will be the same for $P(\mu)$:

$$P(\boldsymbol{\mu}) = \frac{b}{|\mathbf{A}\mathbf{A}^t|^{1/2}} exp(-\frac{1}{2}(\mu - \mu_0)^T\boldsymbol{\Sigma}_0^{-1}(\mu - \mu_0))$$

So by the transformation of the space, we only scaled the constants a and b. It is easy to see that our new log-likelihood $L(\mu)$ will be:

$$L(\boldsymbol{\mu}) = [\sum_{k=1}^{n} ln(\frac{a}{|\mathbf{A}\mathbf{A}^t|^{1/2}}) - \frac{1}{2}(\mathbf{x_k} - \mu)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x_k} - \mu)] + ln(\frac{b}{|\mathbf{A}\mathbf{A}^t|^{1/2}}) - \frac{1}{2}(\mu - \mu_0)^T\boldsymbol{\Sigma}_0^{-1}(\mu - \mu_0)$$

And when we take the derivative, these scaled terms of a and b will vanish, since they are just constants. The final MAP estimator will be the same as the old one:

$$\hat{\boldsymbol{\mu'}} = \hat{\boldsymbol{\mu}} = \frac{\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\Sigma}^{-1}\sum_{k=1}^{n}\mathbf{x}_k}{(\boldsymbol{\Sigma}_0^{-1} + n\boldsymbol{\Sigma}^{-1})}$$

So it turns out, yes, our old MAP estimator gives an appropriate estimate for the transformed space.

Answer 9

a)We use the linearity of expectation to show that $E[\hat{\mu}] = \frac{1}{n}(E[x_1]+E[x_2]+E[x_3]+...+E[x_n])$ which is $x_1 = \frac{n \cdot x_1}{n}$ so the estimator is unbiased b)The ML estimate for $\hat{\mu}$ is $\frac{x_1+x_2+...+x_n}{n}$ which

	Gaussian	Cauchy	Uniform
Random Sample	This is an overall bad estimator for gaussian, yet since the samples closer to the mean are more likely to be selected (and even to be sampled in the first place),gaussian is the most appropriate for this kind of sampling out of three.	Same as Gaussian, the only difference is that for Cauchy this sampling is sligthly worse since the tails are fatter and the middle bell is thinner compared to gaussian, so sampling closer to mean is a little less likely.	Uniform is the worst distribution for this sampling since all values regardless of their distances to mean are equally likely to be selected.
Avg. of Samples	This is a good estimator for all of them, satisfying many wanted properties, and also coincides with some know estimators.	-	-
Avg. of Extremes	This is an ok estimate, since the distrubution is symetric around mean. Yet comparing between three, this is most apropriate for the uniform case, since it is more likely that the choosen extremes will have the same distance from mean. For gaussian and cauchy, since they have tails, it is more likely that one extreme "over-throws" the other one, which means is further away from the mean.	As always, cauchy is between uniform and gaussian w.r.t how much helpfull this sampling method is.	-
Trimmed Mean	This method helps for gaussian, since both gaussian and cauchy has tails which means for a small chance a very big or small sample may impact the average disproportionally. By trimming, we reduce the chance of this happening.	-	For uniform tough i believe this does not make a difference when compared with Avg. of Samples. Since The trimmed values are equally likely to appear compared to non-trimmed values, intiutively speaking this doesn't change our "sampling bias".

maskes the variance $\hat{\sigma} = \sum_i x_i - \hat{\mu}$ for the mean in part a we would have a zero term which would distort the variance undesirably.
c)Again we use the linearity of expectation $(n-1)E[\hat{\sigma}]$=$E[\sum_{i=1}^{n} x_i^2 - 2\mu x_i + \mu^2]$
Which will obiously become $(n-1)E[\hat{\sigma}]$=$n(E[x_i^2] - E[\mu^2])$ which is $\frac{n-1}{n}E[\hat{\sigma}]$=$(E[x_i^2] - E[\mu^2])$

Answer 10

Answer 11

a)If we take the limit of both equations as they go to infinity we see that μ_n approaches $\hat{\mu_n}$. We see that the coefficient of the first term approaches 1 as the denominator of the second term goes to infinity making the coefficient 0. For σ_n^2 approaches 0, since again the denominator will outgrow the expression, this makes quite a lot of sense since as the sample size increases it is bound to be more predictable.
b)The limit of μ_n as n approaches 0 is μ_0. Since we know the limit at infinity which is $\hat{\mu_n}$, we can easily say that μ_n lies between these two.
c)So basically σ_0^2=0 which makes μ_n=μ_0 and σ_n^2=0. İt also means that we have guessed the parameters of the distribution exactly. Which is why neither σ_n^2 nor μ_n will not change as the sample size grows

d)μ_n becomes $\hat{\mu_n}$, and σ_n^2 becomes $\frac{\sigma^2}{n}$

Answer 12