CMPS 242: Homework #2

Due: *October 25, 2016*

S V N Vishwanathan

Kostas Zampetakis	Panos Karagiannis
1567380	1309484

Contents

Question 1

Consider the following 1-d dataset with 5 points $X = \{-1, 1, 10, -0.5, 0\}$, on which we are going to perform Gaussian density estimation. For the exercise below, you may use Python for plotting but all the calculations have to be done by hand.

(a) Compute the Maximum Likelihood Estimate (MLE) of the mean and variance. For the variance, compute both the unbiased and biased versions. Comment on what you observe. In particular, how does the presence of an outlier affect your estimates.

(b) Assume that you have a $\mathcal{N}(0, 1)$ prior over the mean parameter and set the standard deviation $\sigma^2 = 1$. Compute the posterior distribution of the mean parameter and plot both the prior and the posterior distributions. Comment on what you observe.

(c) Now suppose we change the prior over the mean parameter to $\mathcal{N}(10, 1)$. Compute the new posterior distribution, plot it, and contrast it with what you observed previously.

(d) Suppose 2 more data points get added to your dataset: $X = \{-1, 1, 10, -0.5, 0, 2, 0.5\}$ Using the same $\mathcal{N}(0, 1)$ prior over the mean parameter, compute and plot the posterior. How does observing new data points affect the posterior?

Answer:

(a) Let $X = \{x_1 \dots x_n\}$ and we are asked to assume that $P(X|\mu, \sigma^2) = N(\mu, \sigma^2)$. Then from Homework 1 Exercise 2 we know that for the mean:

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{1}{N}[-1 + 1 + 10 + 0 - 0.5] = \frac{1}{5}(9.5) = 1.9$$

and that for the biased variance:

$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2 =$$

$$= \frac{1}{5}\left[(-1 - 1.9)^2 + (1 - 1.9)^2 + (10 - 1.9)^2 + (0 - 1.9)^2 + (-0.5 - 1.9)^2 + \right] = \frac{1}{5}84.2 = 16.84$$

By definition a statistic $\bar{y}$ is an unbiased estimate of a parameter y being estimated if $E[\bar{y}] = y$. But from Homework 1 Exercise 2 we know that:

$$E[\sigma^2_{ML}] = \frac{N-1}{N}\sigma^2$$

As a result an unbiased estimate of σ^2 must be

$$\sigma^2_{UB} = \frac{N}{N-1}\sigma^2_{ML} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{ML})^2 = \frac{1}{4}84.2 = 21.05$$

In the above we see that the presence of an outlier, affects significantly the variance in both the biased and unbiased case. The variance in either cases becomes larger since the value $X = 10$ introduces more uncertainty around the mean of the population. This is a drawback of the MLE method, that a Bayesian approach eliminates by introducing a prior distribution over the mean.

(b) Now we need to assume that $p(\mu) = N(\mu_0 = 0, \sigma_0^2 = 1)$ and also that $\forall x_i \in X \implies P(x_i \mid \mu) = N(\mu, \sigma^2 = 1)$. Then we can simply find the posterior by:

$$P(\mu \mid X) = \frac{P(X \mid \mu)P(\mu)}{P(X)} = \frac{P(X \mid \mu)P(\mu)}{\int_{-\infty}^{\infty} P(X \mid \mu)P(\mu)d\mu} \propto P(X \mid \mu)P(\mu)$$

Now we work further on the product:

$$P(X \mid \mu)P(\mu) = N(\mu, \sigma^2)N(\mu_0, \sigma_0^2)$$

Assuming i.i.d. on the observations X we get:

$$P(X \mid \mu) = \prod_{i=1}^{n} P(x_i \mid \mu) = \prod_{i=1}^{n} N(x_i \mid \mu, \sigma^2)$$

Hence after making the i.i.d. assumption:

$$P(X|\mu)P(\mu) = \prod_{i=1}^{n} P(x_i|\mu) = \prod_{i=1}^{n} N(x_i|\mu,\sigma^2)N(\mu_0,\sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\frac{1}{\sqrt{2\pi\sigma_0^2}} exp\left\{-\left[\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2} + \frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right]\right\}$$

Manipulate the exponent in order to complete the square in terms of μ:

$$\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}+\frac{(\mu-\mu_0)^2}{2\sigma_0^2} = \sum_{i=1}^{n}\frac{x_i^2-2\mu x_i+\mu^2}{2\sigma^2}+\frac{\mu^2-2\mu\mu_0+\mu_0^2}{2\sigma_0^2} = \sum_{i=1}^{n}\frac{x_i^2}{2\sigma^2}-2\mu\sum_{i=1}^{n}\frac{x_i}{2\sigma^2}+n\frac{\mu}{2\sigma^2}-2\frac{\mu\mu_0}{2\sigma_0^2}+\frac{\mu_0^2}{2\sigma_0^2}$$

Since we focus on μ we can drop all terms not involving the unknown, since they will only contribute to the normalization constant, hence:

$$\sum_{i=1}^{n}\frac{x_i^2}{2\sigma^2}-2\mu\sum_{i=1}^{n}\frac{x_i}{2\sigma^2}+n\frac{\mu}{2\sigma^2}-2\frac{\mu\mu_0}{2\sigma_0^2}+\frac{\mu_0^2}{2\sigma_0^2} = 2\mu\sum_{i=1}^{n}\frac{x_i}{2\sigma^2}+n\frac{\mu}{2\sigma^2}-2\frac{\mu\mu_0}{2\sigma_0^2}+\frac{\mu_0^2}{2\sigma_0^2} = \frac{1}{2}\left(\mu^2\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)-2\mu\left(\sum_{i=1}^{n}\frac{x_i}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)\right)$$

To simplify notation let:

$$\beta = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}$$

$$\gamma = \sum_{i=1}^{n}\frac{x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}$$

Hence we get the following expression:

$$\frac{1}{2}\left(\mu^2\beta - 2\mu\gamma\right) = \frac{\beta}{2}\left(\mu^2 - 2\mu\left(\frac{\gamma}{\beta}\right) + \left(\frac{\gamma}{\beta}\right)^2 - \left(\frac{\gamma}{\beta}\right)^2\right) = \frac{\beta}{2}\left(\mu - \frac{\gamma}{\beta}\right)^2 - \frac{\beta}{2}\left(\frac{\gamma}{\beta}\right)^2$$

Again we can drop the term $\frac{\beta}{2}\left(\frac{\gamma}{\beta}\right)^2$ because it is not a function of μ finally getting:

$$\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2} + \frac{(\mu-\mu_0)^2}{2\sigma_0^2} = \frac{\beta}{2}\left(\mu - \frac{\gamma}{\beta}\right)^2$$

Now we simply have to plug this result back to the exponent to get:

$$P(X \mid \mu)P(\mu) = \frac{1}{\sqrt{2\pi\sigma^2}}\frac{1}{\sqrt{2\pi\sigma_0^2}}e^{-\frac{(\mu-\frac{\gamma}{\beta})^2}{2\frac{1}{\beta}}}$$

If we let $P(\mu \mid X) = N(\mu_n, \sigma_n^2)$ then by equating coefficients we get that:

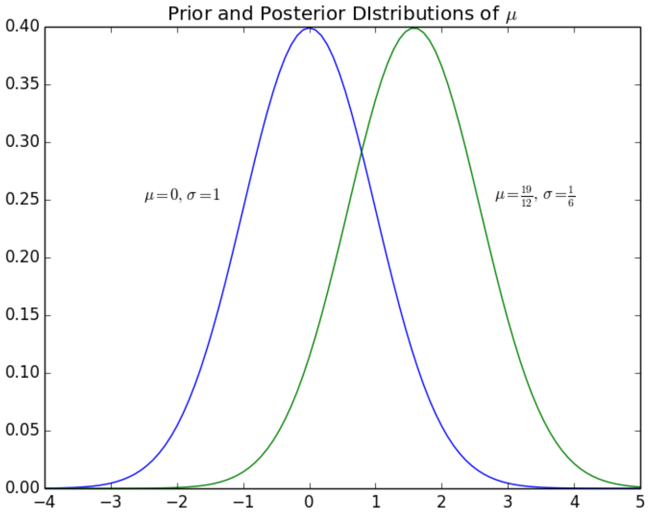
$$\sigma_n^2 = \frac{1}{\beta} = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2}$$

$$\mu_n = \frac{\gamma}{\beta} = \gamma\,\sigma_n^2 = \left(\sum_{i=1}^{n} \frac{x_i}{\sigma^2} + 2\frac{\mu_0}{\sigma_0^2}\right)\sigma_n^2 = \frac{\sigma_n^2}{\sigma^2}\sum_{i=1}^{n} x_i + \frac{\sigma_n^2}{\sigma_0^2}\mu_0$$

In the case of $N(\mu_0 = 0, \sigma_0^2 = 1)$ as prior, $\sigma^2 = 1$ and $n = 5$, $\sum_{i=1}^{5} x_i = -1 + 1 + 10 - 0.5 + 0 = 9.5$ we conclude that:

$$\sigma_n^2 = \frac{1}{5+1} = \frac{1}{6}$$

$$\mu_n = \frac{1}{6}9.5 + 0 = \frac{19}{12}$$



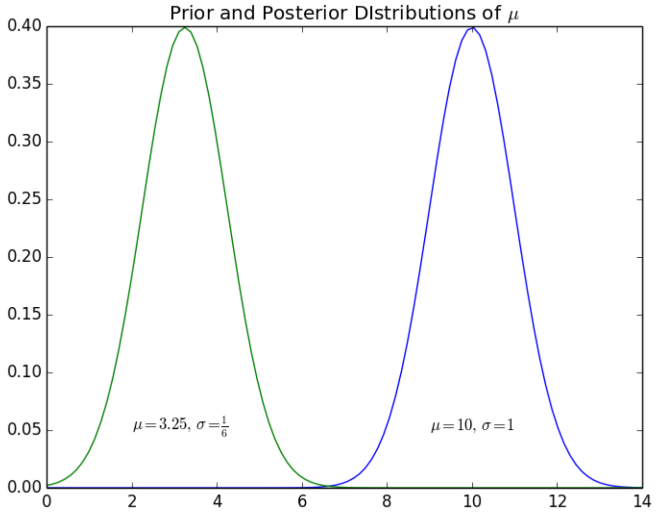
Explanation:
Our prior distribution of the mean μ expresses our belief before observing the data. Therefore our belief before observing the values in X is that $P(\mu) = N(\mu_0 = 0, \sigma_0^2 = 1)$, while after observing X we change our belief saying that $P(\mu \mid X) = N(\mu_n = \frac{19}{12}, \sigma_n^2 = \frac{1}{6})$. Since the mean of the prior distribution is 0 it is logical to observe that the mean of the posterior is very close to the sample mean of the data set X. Moreover, we notice that the variance decreases from our prior belief. This is because our posterior variance is proportional to the variance which we assume our entire data to have ($\sigma^2 = 1$) and our prior belief about the variance ($\sigma_0^2 = 1$) but also because it is inversely proportional to the sample size n. As a result, we are more confident in our posterior distribution that most of our data will be centered around $\mu_n = \frac{19}{12}$. Lastly we should point out that our prior and posterior distributions still intersect signifying that after seeing the data set X our belief of how the x values are distributed did not change dramatically.

(c) Changing the prior to be $N(\mu_0 = 10, \sigma_0^2 = 1)$ we get:

$$\sigma_n = \frac{1}{6}$$

$$\mu_n = \frac{1}{6}9.5 + \frac{1}{6}10 = 3.25$$

Question 1 [(c)] continued on next page. . .



Explanation:
In this case we adjust our prior distribution, claiming that $P(\mu) = N(\mu_0 = 10, \sigma_0^2 = 1)$. As far as the posterior variance is concerned, this example is identical to 1(b) thus our previous argument as to why it became $\sigma_n^2 = \frac{1}{6}$ still applies. Nevertheless, we observe that the posterior mean μ_n has changed a lot compared to our prior belief μ_0, leading almost to no intersection between the graphs of the prior and the posterior distribution. Furthermore it is important to notice that in our model we put a lot more trust in the sample mean that we observe rather than our prior belief about the distribution of the mean. This might not be clear from the current formula for σ_n nonetheless by letting $\bar{x}$ be the sample mean and by employing some algebraic manipulations we can see that:

$$\mu_n = \frac{\sigma_n^2}{\sigma^2}\sum_{i=1}^{n} x_i + \frac{\sigma_n^2}{\sigma_0^2}\mu_0 =$$

$$\frac{\sigma_0^2}{n\sigma_0^2 + \sigma^2}(n\bar{x}) + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 =$$

$$\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{x} + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 =$$

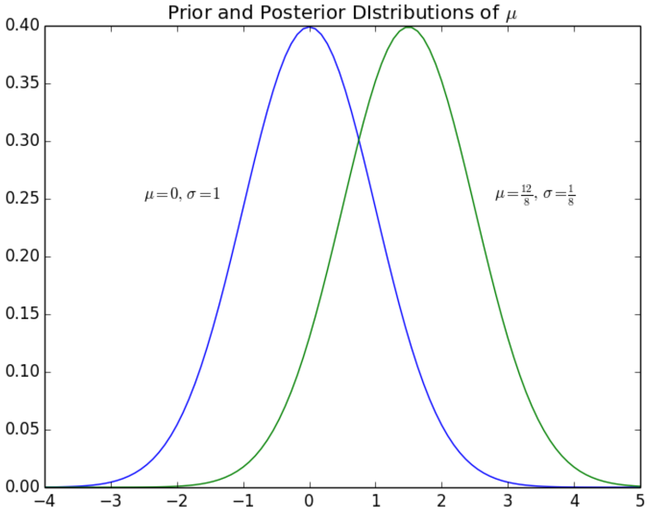
In our case since $\sigma^2 = \sigma_0^2 = 1$, we see that the weight of the sample mean is $\frac{n}{n+1}$ while the weight of the prior belief in the mean is just $\frac{1}{n+1}$. This is the reason why we observe the mean of the posterior distribution to be closer to the sample mean rather than the prior mean.

(d) By adding two more data points to X the only thing that changes is $n = 7$ and $\sum_{i=1}^{7} x_i = -1 + 1 + 10 - 0.5 + 0 + 2 + 0.5 = 12$ resulting to:

$$\sigma_n = \frac{1}{8}$$

$$\mu_n = \frac{7}{8}\frac{12}{7} + \frac{1}{8}0 = \frac{12}{8}$$

Question 1 [(d)] continued on next page. . .



Explanation:
As we saw from calculating the posterior distribution, adding two more data points already close to the pre-existing sample mean decreases the variance of the the posterior distribution making us more certain about our posterior belief. The mean of the posterior μ_n using a similar argument as in parts (b),(c) will tend to be closer to the sample mean. This is also what we observe in the above graph.

Question 2

Generate 100 data points as follows: Draw x uniformly at random1 from $[-100, 100]$. For each x draw t from $\mathcal{N} f(x), 1$ where $f(x) = 0.1 + 2x + x^2 + 3x^3$. In order to fit this curve, we will make use of the following probabilistic model:

$$\mathbb{P}(t|x, w, \beta) = \mathcal{N}(t \mid (x, w), \beta^{-1}).$$

where $y(x, w) = w_0 + w_1 x + w_2 x^2 + w_3 x^3$.

- Perform MLE estimation of w and β. You may use the optimize module from scipy for this task. Comment on how well w and β match the true parameters used to generate the data. How do the estimates change when you use 1000 or $10,000$ data points for your estimates?
- Now suppose that you use $y(x, w) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_5 x^5$ and repeat the above task. Comment on what changed.
- Refer to the slides from the class, where we added a prior over w in order to derive Bayesian linear regression. Assume that we set the hyperparameter $\alpha = 1$ and plot the Bayesian estimate of the curve and the uncertainty around the estimate. How well does it match the observed data. How does the estimate change when you use 1000 or $10,000$ data points?
- Perform Bayesian linear regression but now set $\alpha = 100$. How does this change your estimates? Again, what happens when you use 1000 or $10,000$ data points?

(a) Assume $T = \{t_1 \dots t_n\}$ and $X = \{x_1 \dots x_n\}$. Then we are given that $P(T|X, \mathbf{w}, \beta) = N(T|y(X, \mathbf{w}), \beta^{-1})$ and by assuming i.i.d on X we get:

$$P(T \mid X, \mathbf{w}, \beta) = \prod_{i=1}^{N} N(t_i \mid y(x_i, \mathbf{w}), \beta^{-1}) \tag{1}$$

Where $y(x_i, \mathbf{w}) = w_0 + w_1 x_i + w_2 x_i^2 + w_3 x_i^3$

We know that (1) is the likelihood therefore we need to compute:

$$\arg\max_{\mathbf{w},\beta} \prod_{i=1}^{N} N(t_i \mid y(x_i, \mathbf{w}), \beta^{-1})$$

Manipulating the product in (1) , and since $log(x)$ is an increasing function we can equivalently maximize:

$$\begin{aligned}
\arg\max_{\mathbf{w},\beta} log\left(\prod_{i=1}^{N} N(t \mid y(x_i, \mathbf{w}), \beta^{-1})\right) &= \\
\arg\max_{\mathbf{w},\beta} \sum_{i=1}^{N} log\left(N(t \mid y(x_i, \mathbf{w}), \beta^{-1})\right) &= \\
\arg\max_{\mathbf{w},\beta} \sum_{i=1}^{N} log\left(\frac{\sqrt{\beta}}{\sqrt{2\pi}}\right) exp\left\{\frac{\beta\left(t_i - y(x_i, \mathbf{w})\right)^2}{2}\right\} &= \\
\arg\max_{\mathbf{w},\beta} \frac{N}{2} log(\beta) - \frac{N}{2} log(2\pi) - \frac{\beta}{2} \sum_{i=1}^{N} \left(t_i - y(x_i, \mathbf{w})\right)^2 &
\end{aligned}$$

Equivalently we can minimize with respect to $\mathbf{w}$ and β:

$$\arg\min_{\mathbf{w},\beta} -\frac{N}{2} log(\beta) + \frac{N}{2} log(2\pi) + \frac{\beta}{2} \sum_{i=1}^{N} \left(t_i - y(x_i, \mathbf{w})\right)^2 \tag{2}$$

Taking the partial derivative of (2) with respect to $\mathbf{w}$ we get that we need to minimize the error function:

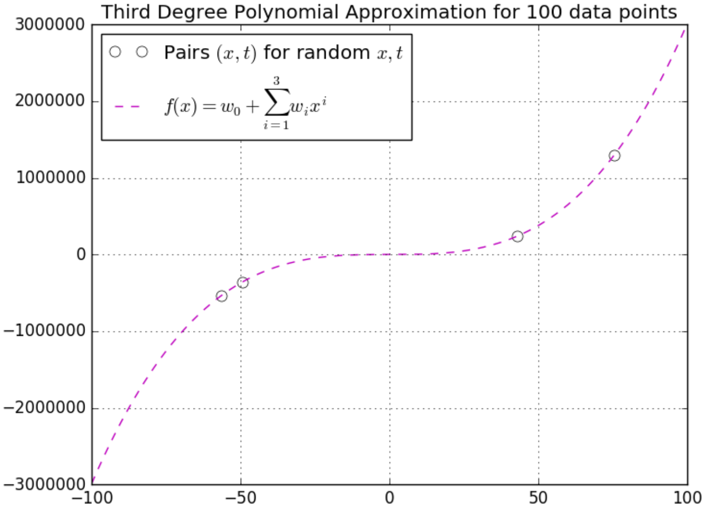
$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{N} \left(t_i - y(x_i, \mathbf{w})\right)^2 = \frac{1}{2} \sum_{i=1}^{N} \left[t_i - (w_0 + w_1 x_i + w_2 x_i^2 + w_3 x_i^3)\right]^2 \tag{3}$$

Also taking the partial derivative of (2) with respect to β:

$$\begin{aligned}
-\frac{N}{\beta_{ML}} + \frac{1}{2} 2 \sum_{i=1}^{N} \left(t_i - y(x_i, \mathbf{w}_{ML})\right) = 0 \Leftrightarrow \\
\frac{1}{\beta_{ML}} = \frac{1}{N} \sum_{i=1}^{N} \left(t_i - y(x_i, \mathbf{w}_{ML})\right)
\end{aligned}$$

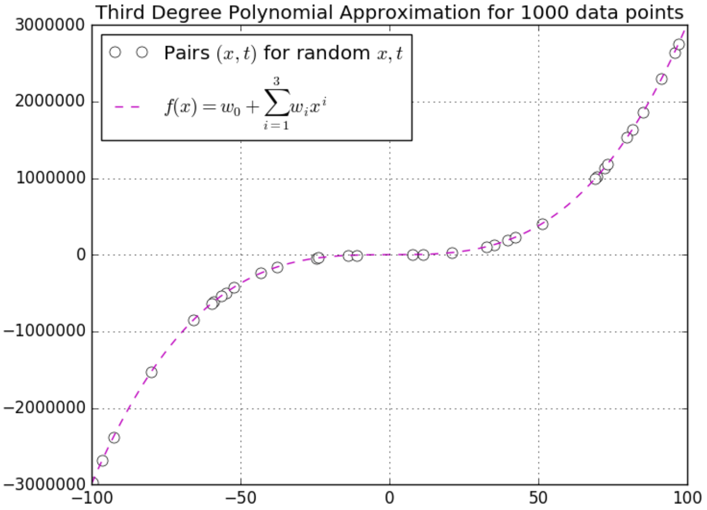
Using the optimize module and 100 data points we get the values of $\mathbf{w}_{ML} = [w_0, w_1, w_2, w_3] = [0.03142213, 2.00023587, 1.00000$ and $\beta_{ML} = 1.331857$

[(a)] continued on next page...



*only 1 in 30 points appears in the graph for illustration purposes

Next for 1,000 data points we get the values of $\mathbf{w}_{ML} = [w_0, w_1, w_2, w_3] = [0.11227895, 2.00154478, 0.99999288, 2.99999972]$ while $\beta_{ML} = 0.981139$.

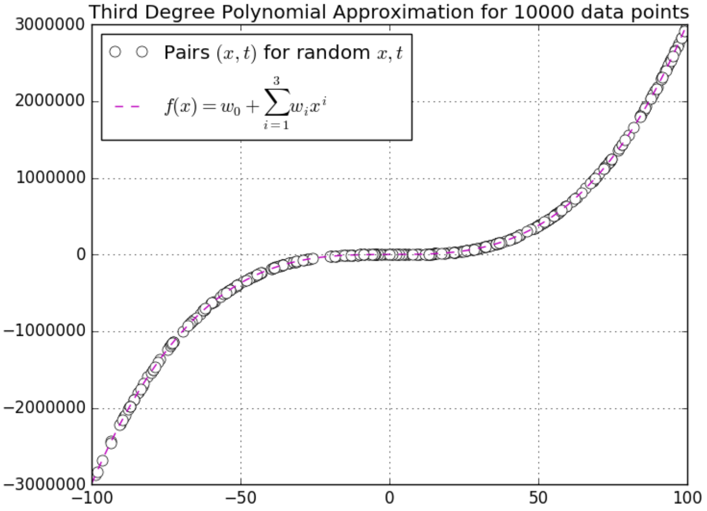


*only 1 in 30 points appears in the graph for illustration purposes

Increasing the data points to 10,000 we get that:

$\mathbf{w}_{ML} = [w_0, w_1, w_2, w_3] = [0.10604444, 1.9997676, 0.99999551, 3.00000003]$ while $\beta_{ML} = 1.012934$.

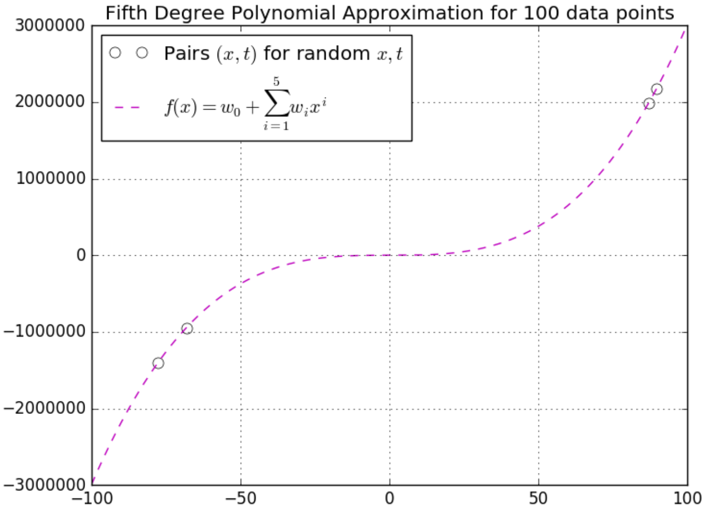
[(a)] continued on next page. . .



*only 1 in 30 points appears in the graph for illustration purposes

As we observe even with 100 data points our estimates for $\mathbf{w}_{ML}$ seem very accurate. Of course by increasing the number of data, our estimates get better and at $10,000$ points the predicted coefficients practically coincide with the true values. Our precision for 100 data points has 0.3 units of difference from our true value. At $1,000$ data points the difference decreases to 0.02, which is significantly small to give us great confidence to our estimates. Actually judging by our plots we observe that the estimations seem to be accurate enough so that no distinction can be easily be spotted with the naked eye. This is a logical result since the number of points in each case, is significantly larger from the degree of our polynomial and variance is very small compared to the the output range.

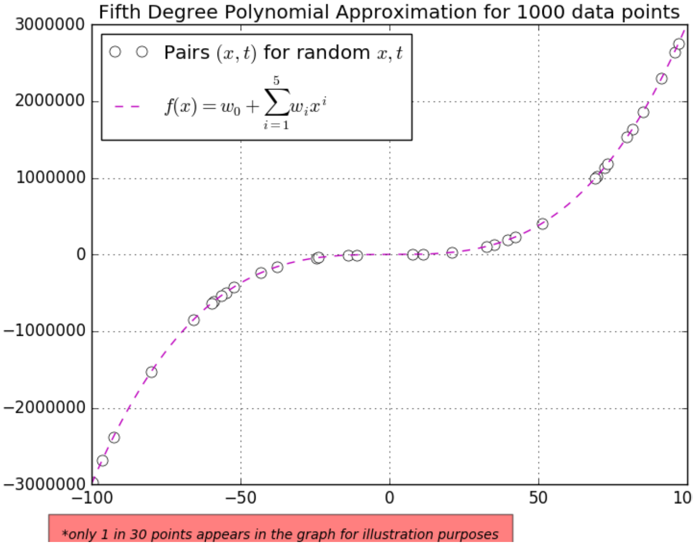
(b) Using a polynomial of higher degree to fit the data, $y(x_i, \mathbf{w}) = w_0 + w_1 x_i + w_2 x_i^2 + w_3 x_i^3 + w_4 x_i^4 + w_5 x_i^5$ and following similar steps as in part (a) to find the values of $\mathbf{w}_{ML}$ and β_{ML} we see that, for 100 data points $\mathbf{w}_{ML} = [w_0, w_1, w_2, w_3, w_4, w_5] = [-0.978830324, 2.00064173, 1.00014741, 2.99999972e + 00, -1.67936917 \times 10^{-8} 0.106228218 \times 10^{-11}]$ and $\beta = 1.360520$



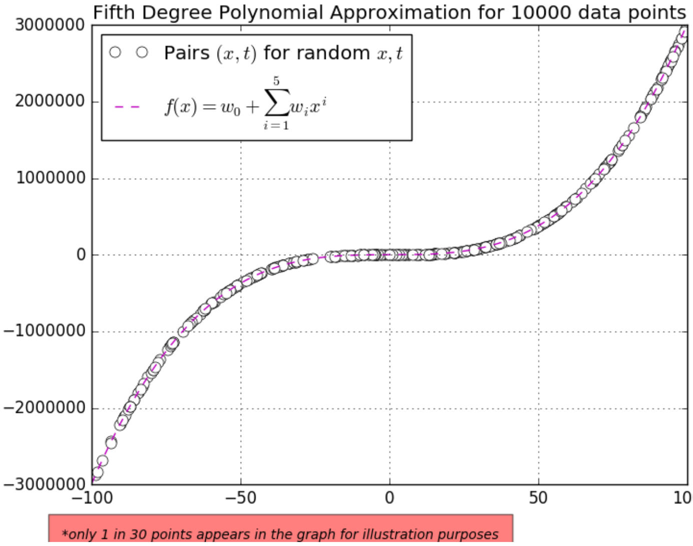
*only 1 in 30 points appears in the graph for illustration purposes

For $1,000$ data points:

$\mathbf{w}_{ML} = [w_0, w_1, w_2, w_3, w_4, w_5] = [0.203860795, 2.00343920, 0.999967466, 2.99999883, 3.12086803\times10^{-9}, 8.17001552\times 10^{-11}]$ while $\beta_{ML} = 0.982455$.



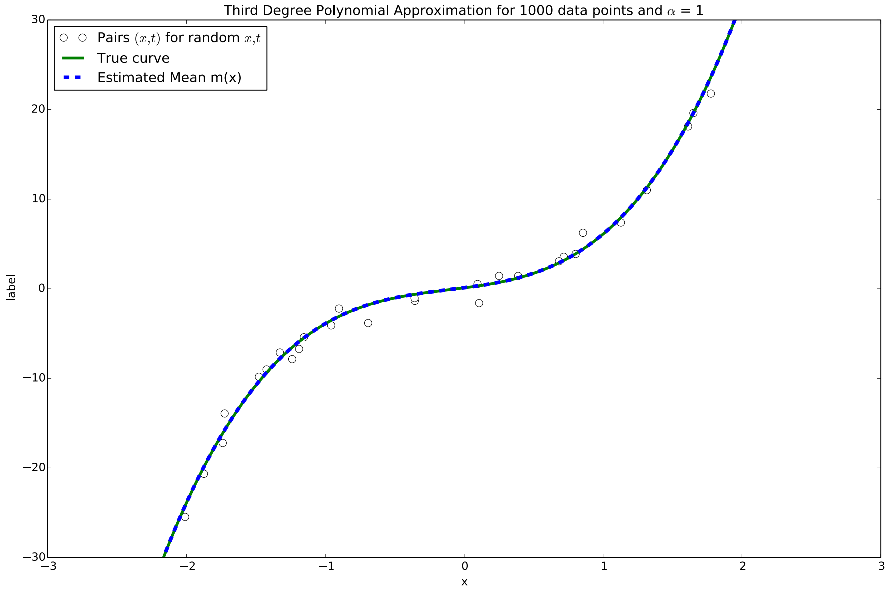
Using 10,000 data points we see:
$\mathbf{w}_{ML} = [w_0, w_1, w_2, w_3, w_4, w_5] = [0.115558057, 1.99823554, 1.00000070, 3.00000072, 2.17624484\times10^{-10}, 6.45810050\times 10^{-11}]$ while $\beta_{ML} = 0.985238$.

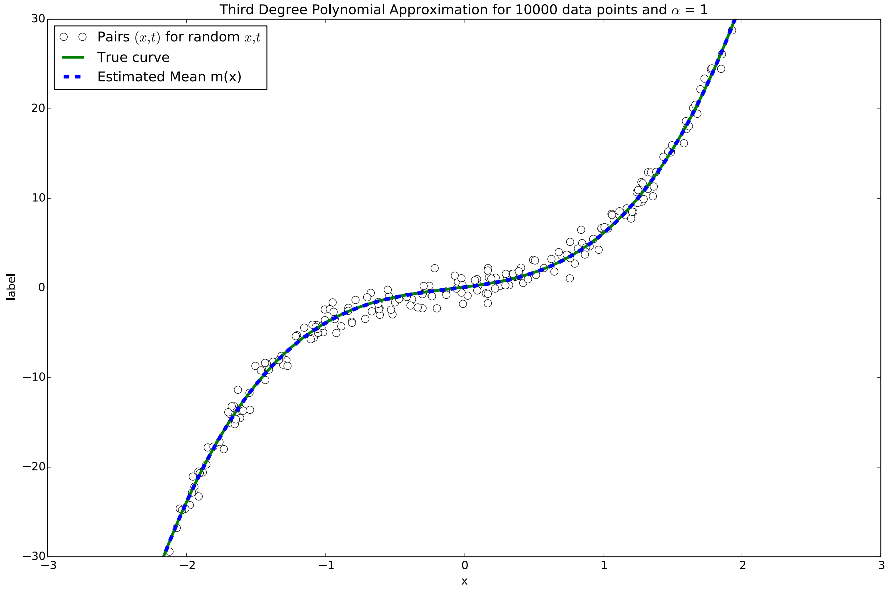


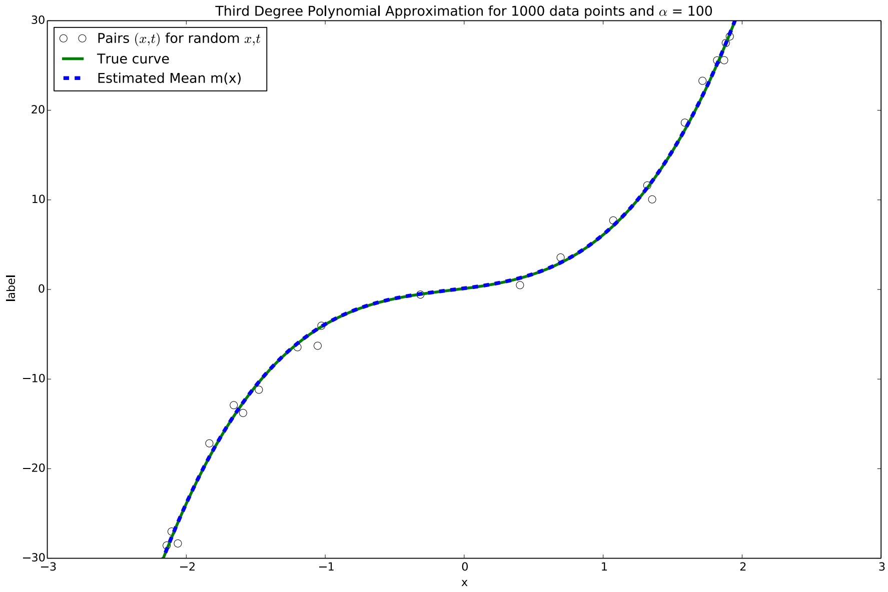
Again, following a similar argument as in part (a) we cannot see any significant difference while varying the number of data points. Also, as in part (a) the precision decreases as we increase the number of data points, in quite the same way as before. We also observe that both estimations with polynomials of degrees 3 and 5, are almost the same. The only difference we observe is the existence of very small coefficients in the x^4,x^5 terms which can actually be ignored as floating point precision errors. This is a result that we anticipated because increasing the degree of the polynomial from 3 to 5 is almost negligent when considering such a large amount of data points.

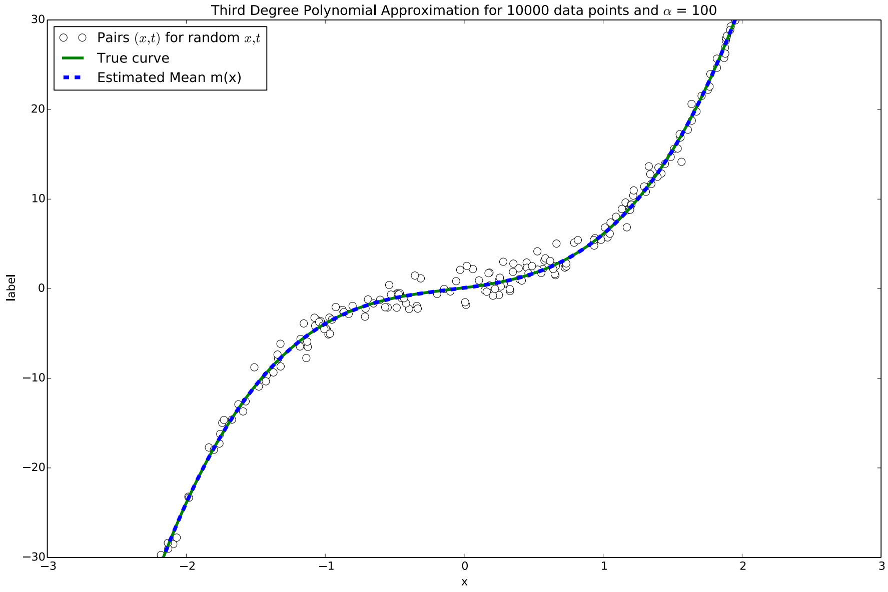
continued on next page...

(c,d) Below follow the plots that we obtained for Bayesian Linear Regression while varying the parameter α as well as the number of sample data points.









Question 3

Let us assume the following generative model for the data:

$$P(\mathbf{x} \mid C_k) = \prod_{i=1}^{D} N(x_i \mid \mu_{ki}, \sigma_k^2)$$

First derive the MLE estimate μ_{ki} and then:

- Assume that k (the number of classes) is equal to 2, and derive the decision rule of the Naive Bayes classifier.
- Suppose the misclassification costs are not equal. In particular, if the cost of misclassifying class 1 is 10 times more than the cost of misclassifying class 2, how does the decision rule change?
- How does the decision rule change when $k = 3$?

Answer:

- On what follows we denote with k the total number of classes, with N the size of our data set and with D the number of components in each given data. $\mathcal{C}$ is a N-vector of classes and $[\mathcal{C}]_i$ its i-th component. Also $\mathcal{X}$ stands for a vector of N data, i.e. each $\mathcal{X}$'s component contains a D dimensional vector $\mathbf{x}_i$. Finally with π_k we denote $\mathbb{P}(\mathcal{C}_k)$ We assuming that we have N data $\mathbf{x}_i$ and for each of them a corresponding label $[\mathcal{C}]_i$, where $i \in \{1, 2, \cdots, N\}$. The likelihood in our case is:

$$\mathbb{P}(\mathcal{C} \mid \mathcal{X}, \{\mu_{k,j}\}_{k\in[m], j\in[N]}, \{\sigma_k^2\}_{k\in[m]})$$

(To simplify our notation in what follows, we will not denote the range of the above indexes). We want to maximize the likelihood with respect to μ_{kj}'s. Equivalently we have:

$$\underset{\mu_{k,j}}{\arg\max}\ \mathbb{P}(\mathcal{C} \mid \mathcal{X}, \{\mu_{k,j}\}, \{\sigma_k^2\}) \overset{\text{i.i.d}}{=} \underset{\mu_{k,j}}{\arg\max} \prod_{i=1}^{N} \mathbb{P}([\mathcal{C}]_i, \mathbf{x}_i \mid \mu_{k,j}, \sigma_k^2) =$$

$$\underset{\mu_{k,j}}{\arg\max} \prod_{i=1}^{N} \left(\prod_{l=1}^{D} \mathbb{P}(x_{il} \mid [\mathcal{C}]_i, \mu_{[\mathcal{C}]_i,l}, \sigma^2_{[\mathcal{C}]_i}) \right) \cdot \mathbb{P}([\mathcal{C}]_i) =$$

$$\underset{\mu_{k,j}}{\arg\max} \prod_{i=1}^{N} \left(\prod_{l=1}^{D} \frac{1}{\sqrt{2\pi\sigma^2_{[\mathcal{C}]_i}}} \exp\left(-\frac{(x_{il} - \mu_{[\mathcal{C}]_i,l})^2}{2\sigma^2_{[\mathcal{C}]_i}} \right) \right) \cdot \pi_{[\mathcal{C}]_i} =$$

$$\underset{\mu_{k,j}}{\arg\max} \prod_{i=1}^{N} \left[\frac{1}{\sqrt{2\pi\sigma^2_{[\mathcal{C}]_i}}} \right]^D \left(\prod_{l=1}^{D} \exp\left(-\frac{(x_{il} - \mu_{[\mathcal{C}]_i,l})^2}{2\sigma^2_{[\mathcal{C}]_i}} \right) \right) \cdot \pi_{[\mathcal{C}]_i} =$$

$$\underset{\mu_{k,j}}{\arg\max} \sum_{i=1}^{N} \left[-\frac{D}{2} \ln\left(2\pi\sigma^2_{[\mathcal{C}]_i}\right) + \ln\left(\prod_{l=1}^{D} \exp\left(-\frac{(x_{il} - \mu_{[\mathcal{C}]_i,l})^2}{2\sigma^2_{[\mathcal{C}]_i}} \right) \right) + \ln(\pi_{[\mathcal{C}]_i}) \right] =$$

$$\underset{\mu_{k,j}}{\arg\max}\ -\frac{D}{2} \sum_{i=1}^{N} \ln\left(2\pi\sigma^2_{[\mathcal{C}]_i}\right) + \sum_{i=1}^{N} \sum_{l=1}^{D} \left(-\frac{(x_{il} - \mu_{[\mathcal{C}]_i,l})^2}{2\sigma^2_{[\mathcal{C}]_i}} \right) + \sum_{i=1}^{N} \ln(\pi_{[\mathcal{C}]_i})$$

In order to find the values of $\mu_{k,j}$'s maximizing the above we take its partial derivatives with respect to each one of them, to be equal to 0. In each case the only surviving term is the middle one, so we equivalently take:

$$\frac{\partial}{\partial(\mu_{[\mathcal{C}]_i,l})}\sum_{j=1}^{N}\sum_{r=1}^{D}\left(-\frac{(x_{jr}-\mu_{[\mathcal{C}]_j,r})^2}{2\sigma^2_{[\mathcal{C}]_j}}\right)=0,\ \text{for every } i\in[N], l\in[D]$$

The terms of first sum that will survive are precisely those with indexes j that satisfy $[\mathcal{C}]_j=[\mathcal{C}]_i$, (that is those data belonging on the same class), and among them only those with $r=l$. Thus, defining $Id(i)=\{j\in[N]:[\mathcal{C}]_j=[\mathcal{C}]_i\}$ the above equations for every $i\in[N], l\in[D]$ take the form:

$$\frac{\partial}{\partial(\mu_{[\mathcal{C}]_i,l})}\sum_{j\in Id(i)}\left(-\frac{(x_{jl}-\mu_{[\mathcal{C}]_j,l})^2}{2\sigma^2_{[\mathcal{C}]_j}}\right)=0\Rightarrow\sum_{j\in Id(i)}\frac{2(x_{jl}-\mu_{[\mathcal{C}]_j,l})}{2\sigma^2_{[\mathcal{C}]_j}}=0\Rightarrow$$

$$\sum_{j\in Id(i)}(x_{jl}-\mu_{[\mathcal{C}]_j,l})=0\Rightarrow\sum_{j\in Id(i)}x_{jl}=|Id(i)|\mu_{[\mathcal{C}]_i,l}\Rightarrow$$

$$\boxed{\mu_{[\hat{\mathcal{C}}]_i,l}=\sum_{j\in Id(i)}\frac{x_{jl}}{|Id(i)|}}$$

At this point we observe that the above estimator, derived with full rigourosity, is actually the essential one. Our hypothesis tells us that data are i.i.d., and each component of them is independent from the others. Thus, estimating $\mu_{k,j}$ can be done as follows: keep every $\mathbf{x}_i$ that belongs to $\mathcal{C}_k$, (i.e. $Id(i)$ with the above notation), and among them keep only their j-th component (as other components do not give additional information), then perform MLE for them. A we know that MLE coincides with the sample mean in the case of i.i.d 1-d data, we get exactly the same estimation.

- We denote by π the $\mathbb{P}([\mathcal{C}]_1)$, and we assume that each class missclasification has the same cost. As we explain in the next exercise the optimal rule will be the following:

$$\mathbb{P}(\mathcal{C}_1\mid\mathbf{x})\ \geq\ \mathbb{P}(\mathcal{C}_2\mid\mathbf{x})$$

In our case we have:

$$\frac{\mathbb{P}(\mathbf{x}\mid\mathcal{C}_1)\mathbb{P}(\mathcal{C}_1)}{\mathbb{P}(\mathbf{x})}\ \geq\ \frac{\mathbb{P}(\mathbf{x}\mid\mathcal{C}_2)\mathbb{P}(\mathcal{C}_2)}{\mathbb{P}(\mathbf{x})}\iff$$

$$\ln(\mathbb{P}(\mathbf{x}\mid\mathcal{C}_1))+\ln(\pi)\ \geq\ \ln(\mathbb{P}(\mathbf{x}\mid\mathcal{C}_2))+\ln(1-\pi)\iff$$

$$\ln(\prod_{i=1}^{D}\mathcal{N}(x_i\mid\mu_{1i},\sigma_1^2))+\ln(\pi)\ \geq\ \ln(\prod_{i=1}^{D}\mathcal{N}(x_i\mid\mu_{2i},\sigma_2^2))+\ln(1-\pi)\iff$$

$$\sum_{i=1}^{D}\ln(\mathcal{N}(x_i\mid\mu_{1i},\sigma_1^2))+\ln(\pi)\ \geq\ \sum_{i=1}^{D}\ln(\mathcal{N}(x_i\mid\mu_{2i},\sigma_2^2))+\ln(1-\pi)\iff$$

$$\sum_{i=1}^{D}\left[\ln\left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(x_i-\mu_{1i})^2}{2\sigma_1^2}\right)\right)-\ln\left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left(-\frac{(x_i-\mu_{2i})^2}{2\sigma_2^2}\right)\right)\right]+\ln\left(\frac{\pi}{1-\pi}\right)\geq 0\iff$$

$$\boxed{\sum_{i=1}^{D} x_i^2\left(\frac{\sigma_1^2-\sigma_2^2}{2\sigma_1^2\sigma_2^2}\right)+\sum_{i=1}^{D} x_i\left(\frac{\mu_{1i}}{\sigma_1^2}-\frac{\mu_{2i}}{\sigma_2^2}\right)+\sum_{i=1}^{D}\left(\frac{\mu_{2i}^2}{2\sigma_2^2}-\frac{\mu_{1i}^2}{2\sigma_1^2}\right)+\left(\ln\left(\frac{\pi}{1-\pi}\right)-\frac{D}{2}\ln\left(\frac{\sigma_1^2}{\sigma_2^2}\right)\right)\geq 0}$$

- First let us denote with $C_{1\to 2}$ and $C_{2\to 1}$ the costs of missclasiffying 1 and 2, respectively. Let's also call $\mathcal{R}_1$ and $\mathcal{R}_2$ the regions of $\mathbb{R}^D$ where we classify to $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively, (so $\mathcal{R}_1, \mathcal{R}_2$ disjoint and $\mathcal{R}_1 \cup \mathcal{R}_2 = \mathbb{R}^D$). Then the total expected cost of our choices can be given by:

$$\int_{\mathcal{R}_1} \left(\mathbb{P}(\mathbf{x}, \mathcal{C}_2)\right) C_{2\to 1} + \int_{\mathcal{R}_2} \left(\mathbb{P}(\mathbf{x}, \mathcal{C}_1)\right) C_{1\to 2}$$

Our goal is to find the optimal regions $\mathcal{R}_1, \mathcal{R}_2$ in the sense of minimizing the cost function. We claim that the following partition is optimal:

$$\begin{aligned}\mathcal{R}_1 &= \{\mathbf{x}\in\mathbb{R}^D : \mathbb{P}(\mathcal{C}_1 \mid \mathbf{x})C_{1\to 2} \ \geq\ \mathbb{P}(\mathcal{C}_2 \mid \mathbf{x})C_{2\to 1}\}\\ \mathcal{R}_2 &= \{\mathbf{x}\in\mathbb{R}^D : \mathbb{P}(\mathcal{C}_1 \mid \mathbf{x})C_{1\to 2} \ <\ \mathbb{P}(\mathcal{C}_2 \mid \mathbf{x})C_{2\to 1}\}\end{aligned}$$

In fact the above statement is obvious. Each $\mathbf{x} \in \mathbb{R}^D$ should be taken in one of the integrals, (remember $\mathcal{R}_1, \mathcal{R}_2$ is partition of $\mathbb{R}^D$) and the rule simply says to choose the one contributing the least cost. (This argument is consist a proof in discrete case, and can be easily modified to a proof in continuous case by considering ϵ-neighbourhoods of $\mathbf{x}$'s).

Doing exactly the same procedure as above, but taking into account the multiplied costs, we take the decision rule:

$$\sum_{i=1}^{D} x_i^2\left(\frac{\sigma_1^2-\sigma_2^2}{2\sigma_1^2\sigma_2^2}\right)+\sum_{i=1}^{D} x_i\left(\frac{\mu_{1i}}{\sigma_1^2}-\frac{\mu_{2i}}{\sigma_2^2}\right)+\sum_{i=1}^{D}\left(\frac{\mu_{2i}^2}{2\sigma_2^2}-\frac{\mu_{1i}^2}{2\sigma_1^2}\right)+\left(\ln\left(\frac{\pi}{1-\pi}\right)-\frac{D}{2}\ln\left(\frac{\sigma_1^2}{\sigma_2^2}\right)\right)\geq \ln\left(\frac{C_{2\to 1}}{C_{1\to 2}}\right)$$

As we have $\frac{C_{2\to 1}}{C_{1\to 2}} = 1/10$, we get:

$$\boxed{\sum_{i=1}^{D} x_i^2\left(\frac{\sigma_1^2-\sigma_2^2}{2\sigma_1^2\sigma_2^2}\right)+\sum_{i=1}^{D} x_i\left(\frac{\mu_{1i}}{\sigma_1^2}-\frac{\mu_{2i}}{\sigma_2^2}\right)+\sum_{i=1}^{D}\left(\frac{\mu_{2i}^2}{2\sigma_2^2}-\frac{\mu_{1i}^2}{2\sigma_1^2}\right)+\left(\ln\left(\frac{\pi}{1-\pi}\right)-\frac{D}{2}\ln\left(\frac{\sigma_1^2}{\sigma_2^2}\right)\right)\geq -\ln\left(10\right)}$$

Which seems reasonable, as now we are less inclined to classify a $\mathbf{x}$ to $\mathcal{C}_2$.

- Let's call our classes $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$, and restrict ourselves to the case of equal missclassification costs. In the same way of thinking as above, our cost function is:

$$\int_{\mathcal{R}_1} \left(\mathbb{P}(\mathbf{x}, \mathcal{C}_2) + \mathbb{P}(\mathbf{x}, \mathcal{C}_3)\right) + \int_{\mathcal{R}_2} \left(\mathbb{P}(\mathbf{x}, \mathcal{C}_1) + \mathbb{P}(\mathbf{x}, \mathcal{C}_3)\right) + \int_{\mathcal{R}_3} \left(\mathbb{P}(\mathbf{x}, \mathcal{C}_1) + \mathbb{P}(\mathbf{x}, \mathcal{C}_2)\right)$$

Where again $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3$ consist a partition of $\mathbb{R}^D$. We claim that the optimal partition is:

$$\begin{aligned}\mathcal{R}_1 &= \{\mathbf{x}\in\mathbb{R}^D : \mathbb{P}(\mathcal{C}_1 \mid \mathbf{x}) \ \geq\ \max\{\mathbb{P}(\mathcal{C}_2 \mid \mathbf{x}), \mathbb{P}(\mathcal{C}_3 \mid \mathbf{x})\}\}\\ \mathcal{R}_2 &= \{\mathbf{x}\in\mathbb{R}^D : \mathbb{P}(\mathcal{C}_2 \mid \mathbf{x}) \ \geq\ \max\{\mathbb{P}(\mathcal{C}_1 \mid \mathbf{x}), \mathbb{P}(\mathcal{C}_3 \mid \mathbf{x})\}\}\\ \mathcal{R}_1 &= \{\mathbf{x}\in\mathbb{R}^D : \mathbb{P}(\mathcal{C}_3 \mid \mathbf{x}) \ \geq\ \max\{\mathbb{P}(\mathcal{C}_1 \mid \mathbf{x}), \mathbb{P}(\mathcal{C}_2 \mid \mathbf{x})\}\}\end{aligned}$$

(As we have defined the sets above do not form a partition of $\mathbb{R}^D$ as they are not disjoint. However their intersections integrate to zero no matter the integrand, so they can be consider to belong to any of them, without any difference.) Again, our choice is justified by the fact that each $\mathbf{x}$ should be should be taken in one of the integrals, and we choose to put it in that which gives it the minimum cost. To get an explicit formula, we use the above result for two classes and perform it to each of the 6 inequalities we have.