Problem 2

2.1 Posterior Distribution: Conjugate Prior (10 pts)

Derive the posterior distribution p(μ, τ |D) for the univariate Gaussian case (both the mean μ and the precision τ are unknown) using the conjugate prior NG(μ, τ |μ0, ν 0, α 0, β 0).

We can say that the posterior distribution can be calculated by saying that:

$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

We can say that the dataset D contains N i.i.d samples: $x_i \sim N(\mu, \sigma^2)$

Let: $\tau = \sigma^{-2}$; we call it the precision

We can say that the univariate Gaussian distribution can be written as:

$$p(x_i|\mu, \tau) = \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{\tau}{2}(x_i - \mu)^2\right) \sim N(\mu, \tau^{-1})$$

We can find out that the conjugate prior for this case is the Normal-Gamma distribution given by:

$$NG(\mu, \tau|\mu_0, v_0, \alpha_0, \beta_0) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}\left(\frac{v_0}{2\pi}\right)^2 \tau^{\alpha_0 - \frac{1}{2}} \exp\left(-\frac{\tau}{2}\left[v_0(\mu - \mu_0)^2 + 2\beta_0\right]\right)$$

We can derive the posterior for the univariate Gaussian case using the conjugate prior above. We know that the posterior must take the same form of the conjugate prior, so that the posterior will be in the same form as: $NG(\mu, \tau|\mu_0, v_0, \alpha_0, \beta_0)$

Posterior: $P(\mu, \tau|X) = P(X|\mu, \tau)P(\mu, \tau|\mu_0, v_0, \alpha_0, \beta_0)$

Looking at: $P(X|\mu, \tau)$

$$P(X|\mu, \tau) = \prod_{i=1}^{n} \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{\tau}{2}(x_i - \mu)^2\right)$$

$$P(X|\mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \exp\left(-\frac{\tau}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

$$P(X|\mu,\tau) = \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \exp\left(-\frac{\tau}{2}\left[n\left(u-\bar{x}\right)^2 + \sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\right]\right)$$

Now: $P\left(\mu,\tau|X\right) = P(X|\mu,\tau)P\left(\mu,\tau|\mu_0,v_0,\alpha_0,\beta_0\right)$

$$= \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \exp\left(-\frac{\tau}{2}\left[n\left(u-\bar{x}\right)^2 + \sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\right]\right) \frac{\beta_0^{\alpha_0}}{\Gamma\left(\alpha_0\right)}\left(\frac{v_0}{2\pi}\right)^2 \tau^{\alpha_0-\frac{1}{2}} \exp\left(-\frac{\tau}{2}\left[v_0\left(\mu-\mu_0\right)^2+2\beta_0\right]\right)$$

Let: $\frac{\beta_0^{\alpha_0}}{\Gamma\left(\alpha_0\right)}\left(\frac{v_0}{2\pi}\right)^2 = C_1$

$$= C_1\left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \tau^{\alpha_0-\frac{1}{2}} \exp\left(-\frac{\tau}{2}\left[n\left(u-\bar{x}\right)^2 + \sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\right]\right) \exp\left(-\frac{\tau}{2}\left[v_0\left(\mu-\mu_0\right)^2+2\beta_0\right]\right)$$

$$= C\tau^{n/2+\alpha_0-1/2} \exp\left(-\frac{\tau}{2}\left[n\left(u-\bar{x}\right)^2 + \sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\right]\right) \exp\left(-\frac{\tau}{2}\left[v_0\left(\mu-\mu_0\right)^2+2\beta_0\right]\right)$$

Let: $C = \frac{\beta_0^{\alpha_0}}{\Gamma\left(\alpha_0\right)}\left(\frac{v_0}{2\pi}\right)^2\left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}}$

$$= C\tau^{n/2+\alpha_0-1/2} \exp\left(-\frac{\tau}{2}\left[n\left(\mu-\bar{x}\right)^2 + \sum_{i=1}^{n}(x_i-\bar{x})^2 + v_0\left(\mu-\mu_0\right)^2 + 2\beta_0\right]\right)$$

$$= C\tau^{n/2+\alpha_0-1/2} \exp\left(-\frac{\tau}{2}\left[n\left(\mu^2-2\mu\bar{x}+\bar{x}^2\right) + v_0\left(\mu^2-2\mu\mu_0+\mu_0^2\right) + \sum_{i=1}^{n}(x_i-\bar{x})^2 + 2\beta_0\right]\right)$$

Let's just consider: $n\left(\mu^2-2\mu\bar{x}+\bar{x}^2\right) + v_0\left(\mu^2-2\mu\mu_0+\mu_0^2\right)$

$$n\mu^2 - 2\mu n\bar{x} + n\bar{x}^2 + v_0\mu^2 - 2\mu\mu_0 v_0 + v_0\mu_0^2$$

$$\left(n+v_0\right)\mu^2 - 2\left(n\bar{x}+\mu_0 v_0\right)\mu + \left(n\bar{x}^2+v_0\mu_0^2\right)$$

$$\left(n+v_0\right)\left[\mu^2 - \frac{2\left(n\bar{x}+\mu_0 v_0\right)}{\left(n+v_0\right)}\mu\right] + \left(n\bar{x}^2+v_0\mu_0^2\right)$$

$$\left(n+v_0\right)\left[\mu^2 - \frac{2\left(n\bar{x}+\mu_0 v_0\right)}{\left(n+v_0\right)}\mu + \left[\frac{\left(n\bar{x}+\mu_0 v_0\right)}{\left(n+v_0\right)}\right]^2\right] + \left(n\bar{x}^2+v_0\mu_0^2\right) - \frac{\left(n\bar{x}+\mu_0 v_0\right)^2}{\left(n+v_0\right)}$$

$$\left(n+v_0\right)\left[\mu - \frac{n\bar{x}+\mu+0v_0}{\left(n+v_0\right)}\right]^2 + \frac{\left(n\bar{x}^2+v_0\mu_0^2\right)\left(n+v_0\right) - \left(n\bar{x}+\mu_0 v_0\right)^2}{\left(n+v_0\right)}$$

$$(n+v_0)\left[\mu-\frac{n\bar{x}+\mu+0v_0}{(n+v_0)}\right]^2+\frac{(n^2\bar{x}^2+v_0n\bar{x}^2+nv_0\mu_0^2+v_0^2\mu_0^2)-(n^2\bar{x}^2+2\mu_0v_0n\bar{x}+\mu_0^2v_0^2)}{(n+v_0)}$$

$$(n+v_0)\left[\mu-\frac{n\bar{x}+\mu+0v_0}{(n+v_0)}\right]^2+\frac{v_0n\bar{x}^2+nv_0\mu_0^2-2\mu_0v_0n\bar{x}}{(n+v_0)}$$

$$(n+v_0)\left[\mu-\frac{n\bar{x}+\mu+0v_0}{(n+v_0)}\right]^2+\frac{nv_0\left(\bar{x}^2-2\mu_0\bar{x}+\mu_0^2\right)}{(n+v_0)}$$

$$(n+v_0)\left[\mu-\frac{n\bar{x}+\mu_0v_0}{(n+v_0)}\right]^2+\frac{nv_0\left(\bar{x}-\mu_0\right)^2}{(n+v_0)}$$

Then we bring this back into the original equation:

$$=C\tau^{n/2+\alpha_0-1/2}\exp\left(-\frac{\tau}{2}\left[(n+v_0)\left[\mu-\frac{n\bar{x}+\mu_0v_0}{(n+v_0)}\right]^2+\frac{nv_0\left(\bar{x}-\mu_0\right)^2}{(n+v_0)}+\sum_{i=1}^{n}(x_i-\bar{x})^2+2\beta_0\right]\right)$$

$$=C\tau^{n/2+\alpha_0-1/2}\exp\left(-\frac{\tau}{2}\left[(n+v_0)\left[\mu-\frac{n\bar{x}+\mu_0v_0}{(n+v_0)}\right]^2+\sum_{i=1}^{n}(x_i-\bar{x})^2+2\beta_0+\frac{nv_0\left(\bar{x}-\mu_0\right)^2}{(n+v_0)}\right]\right)$$

$$=C\tau^{n/2+\alpha_0-1/2}\exp\left(-\frac{\tau}{2}\left[(n+v_0)\left[\mu-\frac{n\bar{x}+\mu_0v_0}{(n+v_0)}\right]^2+\sum_{i=1}^{n}(x_i-\bar{x})^2+2\left[\beta_0+\frac{nv_0\left(\bar{x}-\mu_0\right)^2}{2\left(n+v_0\right)}\right]\right]\right)$$

$$=C\tau^{n/2+\alpha_0-1/2}\exp\left(-\frac{\tau}{2}\left[(n+v_0)\left[\mu-\frac{n\bar{x}+\mu_0v_0}{(n+v_0)}\right]^2+2\left[\beta_0+\frac{nv_0\left(\bar{x}-\mu_0\right)^2}{2\left(n+v_0\right)}+\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]\right]\right)$$

Now that we have the final equation, we can compare it to the conjugate prior to get the form:

$$P\left(\mu,\tau|X\right)NG\left(\mu,\tau|\frac{n\bar{x}+\mu_0v_0}{(n+v_0)},n+v_0,\alpha_0+\frac{n}{2},\beta_0+\frac{nv_0\left(\bar{x}-\mu_0\right)^2}{2\left(n+v_0\right)}+\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)$$

2.2 Decision Boundary: Linear or Non-Linear? (10 pts)

Consider a dataset D containing N i.i.d samples which are d dimensional (xi ∈ R d). There are 2 classes (denote as c and c 0) and the class prior follows the categorical distribution p(y) ~ Categorical(π). Note that in Gaussian Bayes classifier we assume the conditional distribution for each class is a multivariate Gaussian, e.g. p(x|y = c) ~ N (μc, Σ c)

The distribution for each class can then be identified as:

$$p(\mathbf{x}|y=c) = \frac{1}{\sqrt{(2\pi)^d |\sum_c|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_c)^T \sum_c^{-1} (\mathbf{x}-\mu_{\mathbf{c}})\right)$$

$$p(\mathbf{x}|y=c') = \frac{1}{\sqrt{(2\pi)^d |\sum_{c'}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_{c'})^T \sum_{c'}^{-1} (\mathbf{x}-\mu_{\mathbf{c'}})\right)$$

Using Baye's Rule, we can write the following posterior distribution:

$$p(y=c|\mathbf{x}) \propto \pi_c |\sum_c|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_c)^T \sum_c^{-1} (\mathbf{x}-\mu_c)\right)$$

$$p(y=c'|\mathbf{x}) \propto \pi_{c'} |\sum_{c'}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_{c'})^T \sum_{c'}^{-1} (\mathbf{x}-\mu_{c'})\right)$$

The Prior is Categorical(π):

$$f(x|\mathbf{P}) = \prod_{i=1}^{k} p_i^{x_i}$$

1. Consider 2 classes, show that if Σ c = Σ c', then the decision boundary for Gaussian Bayes classifier is linear.

By considering the two classes, I can show that if $\sum_c = \sum_{c'}$, then the decision boundary for Gaussian Bayes classifier is linear

The Gaussian Bayes Classifier can be written as:

$$f^*(\mathbf{x}) = \underset{y=c}{\operatorname{argmax}} P(\mathbf{x}|y=c) P(c)$$

We can also say that:

$$P\left(y=c|\mathbf{x}\right)=\frac{P\left(\mathbf{x}|y=c\right)P\left(y=c\right)}{\sum_c P\left(\mathbf{x}|y=c\right)P\left(y=c\right)}$$

$$P\left(y=c|\mathbf{x}\right)=\frac{P\left(\mathbf{x}|y=c\right)P\left(y=c\right)}{\sum_c P\left(\mathbf{x}|y=c\right)P\left(y=c\right)}=P(\mathbf{x}|y=c)P\left(y=c\right)$$

Fleshing this out:

$$P\left(y=c|\mathbf{x}\right)=\frac{P\left(\mathbf{x}|y=c\right)P\left(y=c\right)}{P\left(\mathbf{x}|y=c\right)P\left(y=c\right)+P\left(\mathbf{x}|y=c'\right)P\left(y=c'\right)}$$

$$P\left(y=c'|\mathbf{x}\right)=\frac{P\left(\mathbf{x}|y=c'\right)P\left(y=c'\right)}{P\left(\mathbf{x}|y=c\right)P\left(y=c\right)+P\left(\mathbf{x}|y=c'\right)P\left(y=c'\right)}$$

If we take the form of logistic regression (a linear classifier):

$$P\left(y=c|\mathbf{x}\right)=\frac{P\left(\mathbf{x}|y=c\right)P\left(y=c\right)}{P\left(\mathbf{x}|y=c\right)P\left(y=c\right)+P\left(\mathbf{x}|y=c'\right)P\left(y=c'\right)}\left(\frac{\frac{1}{P(\mathbf{x}|y=c)P(y=c)}}{\frac{1}{P(\mathbf{x}|y=c)P(y=c)}}\right)$$

$$P\left(y=c|\mathbf{x}\right)=\frac{1}{1+\frac{P(y=c')P(\mathbf{x}|y=c')}{P(y=c)P(\mathbf{x}|y=c)}}$$

We can always say that: $x=\exp\ln\left(x\right)$

$$P\left(y=c|\mathbf{x}\right)=\frac{1}{1+\exp\ln\left(\frac{P(y=c')P(\mathbf{x}|y=c')}{P(y=c)P(\mathbf{x}|y=c)}\right)}$$

$$P\left(y=c|\mathbf{x}\right)=\frac{1}{1+\exp\left(\ln\left(\frac{P(y=c')}{P(y=c)}\right)+\ln\left(\frac{P(\mathbf{x}|y=c')}{P(\mathbf{x}|y=c)}\right)\right)}$$

Because the second natural log term is the only one that relies on x, we can solely look at it for our classifier function:

$$f\left(\mathbf{x}\right)=P\left(y=c|\mathbf{x}\right)$$

$$f\left(x\right)=\ln\left(\frac{P\left(y=c'|\mathbf{x}\right)}{P\left(y=c|\mathbf{x}\right)}\right)$$

$$f(\mathbf{x}) = \ln\left(\frac{\pi_{c'}|\sum_{c'}|^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu_{c'})^T\sum_{c'}^{-1}(\mathbf{x}-\mu_{c'})\right)}{\pi_c|\sum_c|^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu_c)^T\sum_c^{-1}(\mathbf{x}-\mu_c)\right)}\right)$$

$$f(\mathbf{x}) = \ln\left(\frac{\pi_{c'}|\sum_{c'}|^{-1/2}}{\pi_c|\sum_c|^{-1/2}}\right) + \ln\left(\frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\mu_{c'})^T\sum_{c'}^{-1}(\mathbf{x}-\mu_{c'})\right)}{\exp\left(-\frac{1}{2}(\mathbf{x}-\mu_c)^T\sum_c^{-1}(\mathbf{x}-\mu_c)\right)}\right)$$

Because: $\sum_c = \sum_{c'}$

$$f(\mathbf{x}) = \ln\left(\frac{\pi_{c'}}{\pi_c}\right) + \ln\left(\frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\mu_{c'})^T\sum_{c'}^{-1}(\mathbf{x}-\mu_{c'})\right)}{\exp\left(-\frac{1}{2}(\mathbf{x}-\mu_c)^T\sum_c^{-1}(\mathbf{x}-\mu_c)\right)}\right)$$

Where $\ln\left(\frac{\pi_c'}{\pi_c}\right)$ is the prior distribution, a constant, that we can let equal b

$$f(\mathbf{x}) = b + \ln\left(\frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\mu_{c'})^T\sum_{c'}^{-1}(\mathbf{x}-\mu_{c'})\right)}{\exp\left(-\frac{1}{2}(\mathbf{x}-\mu_c)^T\sum_c^{-1}(\mathbf{x}-\mu_c)\right)}\right)$$

$$f(\mathbf{x}) = b + \left[-\frac{1}{2}(\mathbf{x}-\mu_{c'})^T\sum_{c'}^{-1}(\mathbf{x}-\mu_{c'})\right] - \left[-\frac{1}{2}(\mathbf{x}-\mu_c)^T\sum_c^{-1}(\mathbf{x}-\mu_c)\right]$$

Because: $\sum_c = \sum_{c'}$

$$f(\mathbf{x}) = b - \frac{1}{2}\sum_c^{-1}\left[\left[(\mathbf{x}-\mu_c)^T(\mathbf{x}-\mu_c)\right] - \left[(\mathbf{x}-\mu_{c'})^T(\mathbf{x}-\mu_{c'})\right]\right]$$

Because the parts inside the square brackets are dot products, we can say that:

$$f(\mathbf{x}) = b - \frac{1}{2}\sum_c^{-1}\sum_{i=1}^{n}\left[(x_i-\mu_{ic})^2 - (x_i-\mu_{ic'})^2\right]$$

$$f(\mathbf{x}) = b - \frac{1}{2}\sum_c^{-1}\sum_{i=1}^{n}\left[x_i^2 - 2\mu_{ic}x_i + \mu_{ic}^2 - \left(x_i^2 - 2\mu_{ic'}x_i + \mu_{ic'}^2\right)\right]$$

$$f\left(\mathbf{x}\right)=b-\frac{1}{2}\sum\nolimits_{c}^{-1}\sum_{i=1}^{n}\left[-2\mu_{ic}x_i+\mu_{ic}^2+2\mu_{ic'}x_i-\mu_{ic'}^2\right]$$

$$f\left(\mathbf{x}\right)=b-\frac{1}{2}\sum\nolimits_{c}^{-1}\sum_{i=1}^{n}\left[2x_i\left(\mu_{ic'}-u_{ic}\right)+\left(\mu_{ic}^2-\mu_{ic'}^2\right)\right]$$

This is the form of the logistic regressor, and the equation is obviously linear.

2. Give out one condition that results non-linear decision boundary and briefly explain why.

Naive Bayes performing as a linear classifier (as shown here) depends on multiple things:

- Y is a boolean with parameter π
- X is a vector
- The distribution is Gaussian
- Values for the X vector are conditionally dependent on the class Y

The easiest to show, however is the case where standard deviations depend on the class, Y

If $\sum_{c}\neq\sum_{c'}$:

Then:

$$f\left(\mathbf{x}\right)=b+\left[-\frac{1}{2}\left(\mathbf{x}-\mu_{c'}\right)^T\sum\nolimits_{c'}^{-1}\left(\mathbf{x}-\mu_{c'}\right)\right]-\left[-\frac{1}{2}\left(\mathbf{x}-\mu_{c}\right)^T\sum\nolimits_{c}^{-1}\left(\mathbf{x}-\mu_{c}\right)\right]$$

Can not be reduce to:

$$f\left(\mathbf{x}\right)=b-\frac{1}{2}\sum\nolimits_{c}^{-1}\left[\left[\left(\mathbf{x}-\mu_{c}\right)^T\left(\mathbf{x}-\mu_{c}\right)\right]-\left[\left(\mathbf{x}-\mu_{c'}\right)^T\left(\mathbf{x}-\mu_{c'}\right)\right]\right]$$

Because of this, you get a situation where:

$$f\left(\mathbf{x}\right)=b-\frac{1}{2}\sum_{i=1}^{n}\left[\sum\nolimits_{c}^{-1}\left(x_i^2-2\mu_{ic}x_i+\mu_{ic}^2\right)-\sum\nolimits_{c'}^{-1}\left(x_i^2-2\mu_{ic'}x_i+\mu_{ic'}^2\right)\right]$$

Here you can see that the x_i^2 will not cancel, and the decision boundary will not be linear.