Problem 1

The Bayesian equations for classification

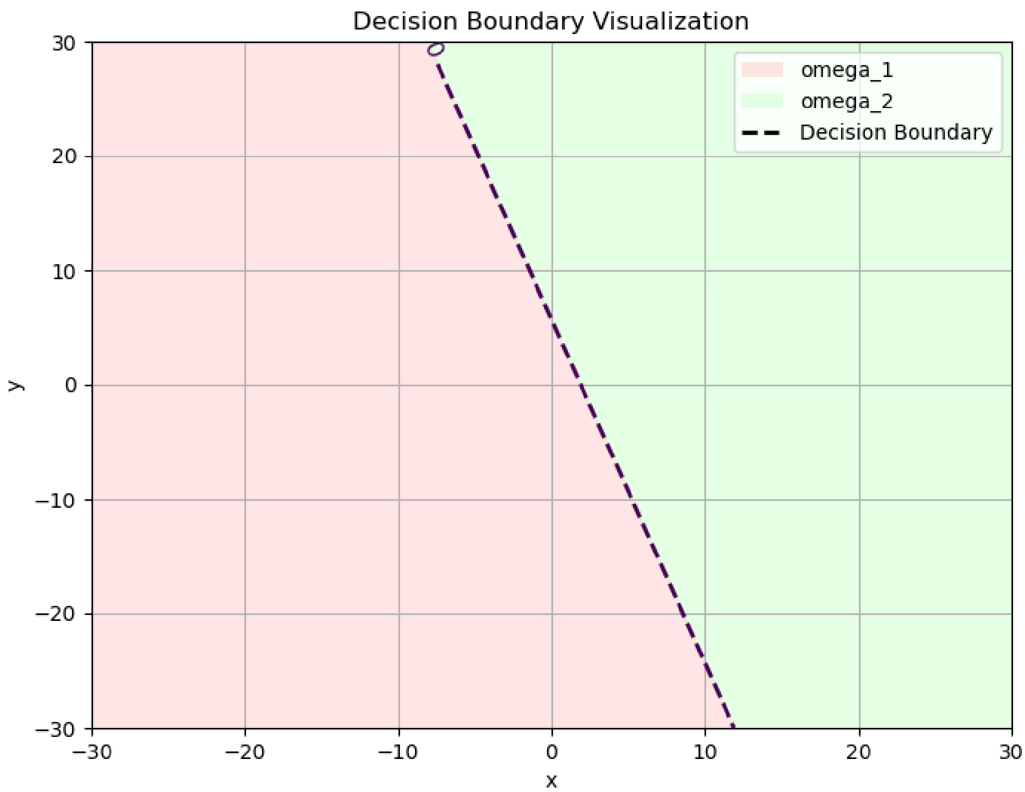
$$
\begin{aligned}
P(\omega=\omega_1|x,y) &= \frac{P(\omega=\omega_1)P(x,y|\omega=\omega_1)}{P(x,y)} \\
&= \frac{P(\omega=\omega_1)P(x,y|\omega=\omega_1)}{P(\omega=\omega_1)P(x,y|\omega=\omega_1)+P(\omega=\omega_2)P(x,y|\omega=\omega_2)} \\
&= \frac{1}{1+\frac{1}{2}e^{-\frac{10-6x-2y}{2}}}
\end{aligned} \tag{1}
$$

$$
\begin{aligned}
P(\omega=\omega_2|x,y) &= \frac{P(\omega=\omega_2)P(x,y|\omega=\omega_2)}{P(x,y)} \\
&= \frac{P(\omega=\omega_2)P(x,y|\omega=\omega_2)}{P(\omega=\omega_1)P(x,y|\omega=\omega_1)+P(\omega=\omega_2)P(x,y|\omega=\omega_2)} \\
&= \frac{1}{1+2e^{-\frac{6x+2y-10}{2}}}
\end{aligned} \tag{2}
$$

so the Bayesian optimal classifier:

$$
\begin{aligned}
f(x,y) &= \begin{cases} \omega_1 & \frac{P(\omega=\omega_1|x,y)}{P(\omega=\omega_2|x,y)} \geq 1 \\ \omega_2 & \frac{P(\omega=\omega_1|x,y)}{P(\omega=\omega_2|x,y)} < 1 \end{cases} \\
\implies f(x,y) &= \begin{cases} \omega_1 & e^{6x+2y-10} \leq 4 \\ \omega_2 & e^{6x+2y-10} > 4 \end{cases}
\end{aligned} \tag{3}
$$

which is the classifier wanted. The decision boundary are as below:



Problem 2

If

$$\frac{P(\omega_1|\boldsymbol{x})}{P(\omega_2|\boldsymbol{x})} \geq \frac{\lambda_2}{\lambda_1} \tag{4}$$

we have

$$\begin{aligned}
f^*(\boldsymbol{x}) &= \min_f \int_{\boldsymbol{x}} \sum_{i=1}^{2} \lambda_i \cdot \mathbb{I}(f(x) \neq \omega_i) \cdot P(\omega_i, \boldsymbol{x}) dx \\
&= \min_f \int_{\boldsymbol{x}} P(\boldsymbol{x}) \cdot \sum_{i=1}^{2} \lambda_i \cdot \mathbb{I}(f(x) \neq \omega_i) P(\omega_i|\boldsymbol{x}) dx \\
&= \min_f \int_{\boldsymbol{x}} \sum_{i=1}^{2} \lambda_i \cdot \mathbb{I}(f(x) \neq \omega_i) P(\omega_i|\boldsymbol{x}) dx \\
&= \min_f \int_{\boldsymbol{x}} \mathbb{I}(f(x) \neq \omega_1) \cdot (\lambda_1 P(\omega_1|\boldsymbol{x})) + \mathbb{I}(f(x) \neq \omega_2) \cdot (\lambda_2 P(\omega_2|\boldsymbol{x}))\, dx \\
&= \min\{\lambda_1 P(\omega_1|\boldsymbol{x}), \lambda_2 P(\omega_2|\boldsymbol{x})\} \qquad (f = f^*) \\
&= \lambda_2 P(\omega_2|\boldsymbol{x})
\end{aligned} \tag{5}$$

With the condition, the minimum takes $\lambda_2 P(\omega_2|\boldsymbol{x})$, which is equivalent to

$$\begin{aligned}
&\mathbb{I}(f^*(\boldsymbol{x}) \neq \omega_2) = 1 \\
\iff &f^*(\boldsymbol{x}) = \omega_1
\end{aligned} \tag{6}$$

The sufficiency is now proved. Below is the proof of necessity.

If the final classifier outputs $f^*(\boldsymbol{x}) = \omega_1$, it leads to

$$\begin{aligned}
&\min_f \int_{\boldsymbol{x}} \sum_{i=1}^{2} \lambda_i \cdot \mathbb{I}(f(x) \neq \omega_i) \cdot P(\omega_i, \boldsymbol{x}) dx \\
=&\lambda_2 P(\omega_2, \boldsymbol{x}) P(\boldsymbol{x})
\end{aligned} \tag{7}$$

which is equivalent to

$$\begin{aligned}
\lambda_2 P(\omega_2|\boldsymbol{x}) P(\boldsymbol{x}) \geq &\left[\int_{\boldsymbol{x}} \sum_{i=1}^{2} \lambda_i \cdot \mathbb{I}(f(x) \neq \omega_i) \cdot P(\omega_i, \boldsymbol{x}) dx\right]_{f(\boldsymbol{x})=\omega_2} \\
=&\lambda_1 P(\omega_1|\boldsymbol{x}) P(\boldsymbol{x}) \\
\implies \frac{P(\omega_1|\boldsymbol{x})}{P(\omega_2|\boldsymbol{x})} \geq &\frac{\lambda_2}{\lambda_1}
\end{aligned} \tag{8}$$

Q.E.D.

Problem 3

Firstly, consider 3 datapoints `1`, `2`, and `3` with features $x_1 < x_2 < x_3$:

There will only be 2 situations: all 3 datapoints are in or out of range $[a, b]$; or only one of the datapoints has the different class with the others. The first case is apparently separable by the range $[a, b]$. In the second case, we can always let the range cover only the point with different class, and the choice of $k \in \{0, 1\}$ can fit whatever the label the particular point has. So we get that

$$d_{VC} \geq 3 \tag{9}$$

Now let's consider 4 data points `1`, `2`, `3` and `4` with features $x_1 < x_2 < x_3 < x_4$, and let point `1` and `3` have label `+1`, point `2` and `4` have label `-1`. It's apparent that if $d_{VC} \geq 4$, there should be a range $[a, b]$ that covers point `1` and `3`, meanwhile `2` and `4` not in it. i.e.

$$\begin{aligned} a \leq x_1 < x_3 \leq b \\ (x_2 - a)(x_2 - b) > 0 \end{aligned} \tag{10}$$

whereas

$$\begin{aligned} & x_1 < x_2 < x_3 \\ \implies & a < x_2 < b \end{aligned} \tag{11}$$

so $\forall h \in \mathcal{H}, h(x_1) = h(x_2) = h(x_3)$, i.e. the points can't be correctly classified. Now we get

$$d_{VC} < 4 \tag{12}$$

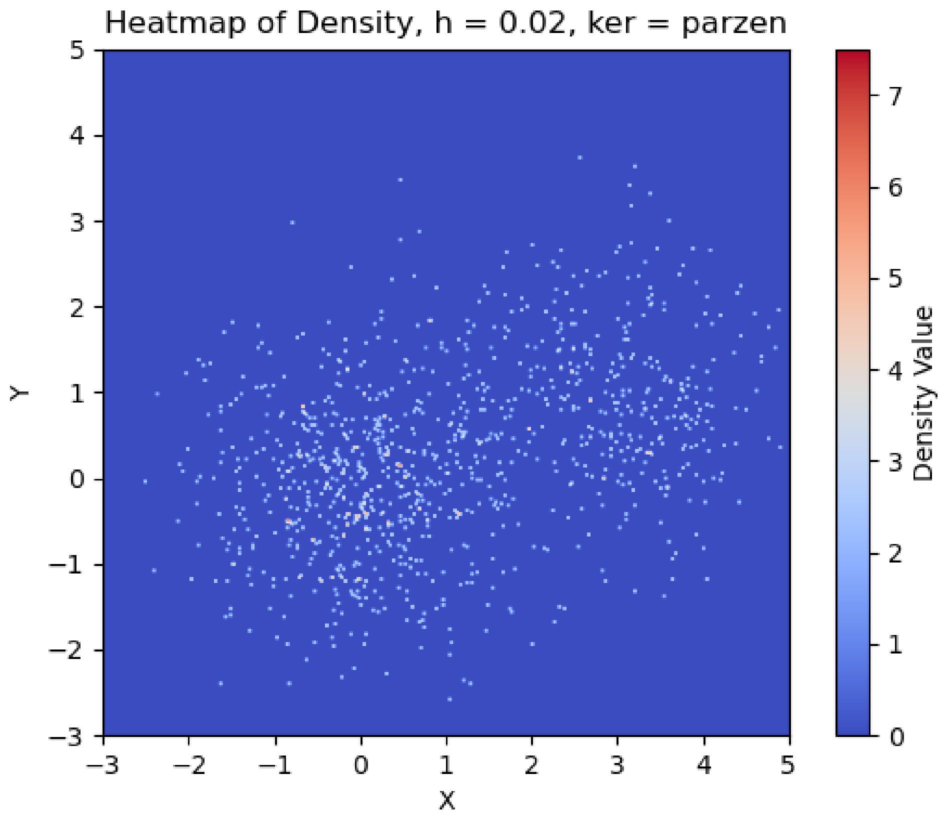
Finally, we have

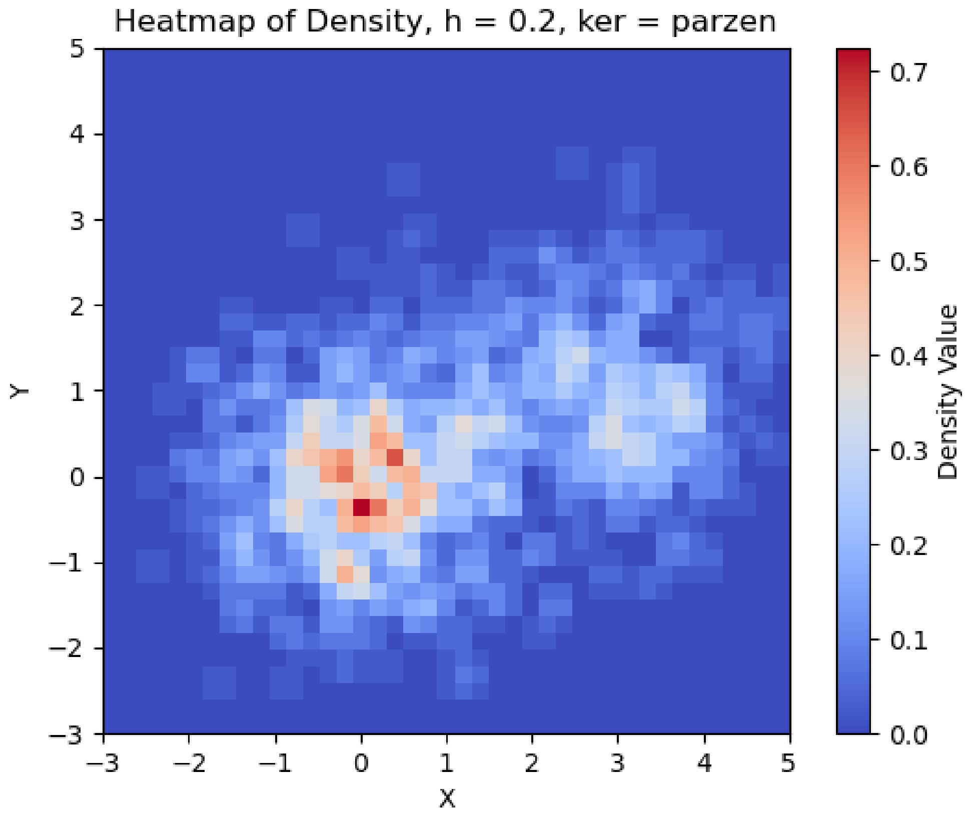
$$d_{VC} = 3 \tag{13}$$

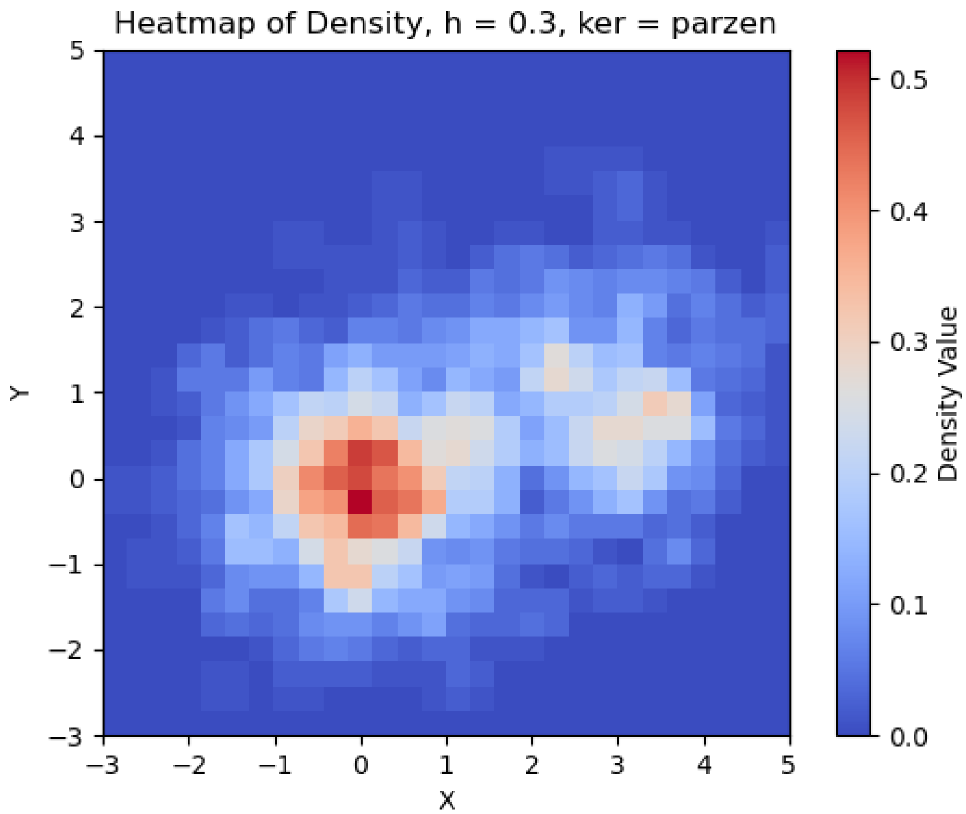
Q.E.D.

Problem 4

The results under `Parzen` window:







The results of error with different kernels and `h` values:

```
error = 4.799405477355697, h = 0.02, ker = parzen
error = 4.239703447154391, h = 0.05, ker = parzen
error = 3.4879346163176645, h = 0.1, ker = parzen
error = 3.091011102695854, h = 0.2, ker = parzen
error = 3.020525044339158, h = 0.3, ker = parzen
error = 2.991360598885944, h = 0.5, ker = parzen
The best composition of kernel parzen is:
 h = 0.5, error = 2.991360598885944
```

```
error = 1.1850893325153076, h = 0.02, ker = gaussian
error = 1.2960466222088645, h = 0.05, ker = gaussian
error = 1.0779539891624839, h = 0.1, ker = gaussian
error = 1.0702003056141813, h = 0.2, ker = gaussian
error = 1.9885279927732376, h = 0.3, ker = gaussian
error = 3.8645331684118123, h = 0.5, ker = gaussian
The best composition of kernel gaussian is:
 h = 0.2, error = 1.0702003056141813
```

```
error = 311.6945435850403, h = 0.02, ker = exp
error = 124.04165583299414, h = 0.05, ker = exp
error = 61.52142573996591, h = 0.1, ker = exp
error = 30.207931963714078, h = 0.2, ker = exp
error = 19.75648753562578, h = 0.3, ker = exp
error = 11.264026157710983, h = 0.5, ker = exp
The best composition of kernel exp is:
 h = 0.5, error = 11.264026157710983
```

where `h` is the size of the meshes. It's clear that the best composition is the Gaussian kernel with `h = 0.2`. It's notable that the normalization of the density matrix matters in the error estimation, especially when the sample set are not large enough.