MACHINE LEARNING: EXERCISE 1

Linxi Wang 587032
Lusine Nazaretyan 513624
Thomas Herold .....
Karen Nazaretyan .....
Ya Qian 518902

Exercise 1

a)

$$p(x, y) = \lambda\eta e^{-\lambda x - \eta y}$$

$$p(x) = \int_0^\infty \lambda\eta e^{-\lambda x - \eta y} \, \mathrm{d}y. \tag{1}$$

we assume that $P(w_1 \mid x) \geq P(w_2 \mid x)$, that is the function $P(error \mid x) = P(w_2 \mid x)$. Now with (1), (2) and (3) we have:

$$\int P(w_2 \mid x)p(x)dx \leq \int \frac{2}{\frac{1}{P(w_1|x)} + \frac{1}{P(w_2|x)}} p(x)dx$$

Because both sides are integrating over the same variable we can simplify the term to:

$$\begin{aligned}
&P(w_2 \mid x)p(x) \leq \frac{2}{\frac{1}{P(w_1|x)} + \frac{1}{P(w_2|x)}} p(x) \\
&\Leftrightarrow P(w_2 \mid x) \leq \frac{2}{\frac{1}{P(w_1|x)} + \frac{1}{P(w_2|x)}} \\
&\Leftrightarrow \frac{1}{P(w_1 \mid x)} + \frac{1}{P(w_2 \mid x)} \leq \frac{2}{P(w_2 \mid x)} \\
&\Leftrightarrow \frac{1}{P(w_1 \mid x)} \leq \frac{1}{P(w_2 \mid x)} \\
&\Leftrightarrow P(w_1 \mid x) \geq P(w_2 \mid x)
\end{aligned}$$

This holds true with the assumptions we made earlier.

b)

With this result, we now show that:

$$P(error) \leq \frac{2P(w_1)P(w_2)}{\sqrt{P(w_1)^2 + (4\mu^2 + 2)P(w_1)P(w_2) + P(w_2)^2}}$$

While using the univariate probability distribution:

$$p(x \mid w_1) = \frac{\pi^{-1}}{1 + (x - \mu)^2} \text{ and } p(x \mid w_2) = \frac{\pi^{-1}}{1 + (x + \mu)^2}$$

With the rule of bayes we have $P(w_1 \mid x) = \frac{p(x|w_1)P(w_1)}{p(x)}$:

$$
\begin{aligned}
P(error) &\leq \int \frac{2}{\frac{1}{P(w_1|x)} + \frac{1}{P(w_2|x)}} p(x)dx \\
\Leftrightarrow P(error) &\leq \int \frac{2}{\frac{1}{\frac{p(x|w_1)P(w_1)}{p(x)}} + \frac{1}{\frac{p(x|w_2)P(w_2)}{p(x)}}} p(x)dx \\
\Leftrightarrow P(error) &\leq \int \frac{2}{\frac{p(x)}{p(x|w_1)P(w_1)} + \frac{p(x)}{p(x|w_2)P(w_2)}} p(x)dx \\
\Leftrightarrow P(error) &\leq \int \frac{2}{\frac{1}{p(x|w_1)P(w_1)} + \frac{1}{p(x|w_2)P(w_2)}} dx \\
\Leftrightarrow P(error) &\leq \int \frac{2}{\frac{1}{\frac{\pi^{-1}}{1+(x-\mu)^2}P(w_1)} + \frac{1}{\frac{\pi^{-1}}{1+(x+\mu)^2}P(w_2)}} dx \\
\Leftrightarrow P(error) &\leq \int \frac{2}{\frac{1}{\frac{\pi^{-1}P(w_1)}{1+(x-\mu)^2}} + \frac{1}{\frac{\pi^{-1}P(w_2)}{1+(x+\mu)^2}}} dx \\
\Leftrightarrow P(error) &\leq \int \frac{2}{\frac{1+(x-\mu)^2}{\pi^{-1}P(w_1)} + \frac{1+(x+\mu)^2}{\pi^{-1}P(w_2)}} dx \\
\Leftrightarrow P(error) &\leq \int \frac{2}{\frac{(1+(x-\mu)^2)P(w_2)}{\pi^{-1}P(w_1)P(w_2)} + \frac{(1+(x+\mu)^2)P(w_1)}{\pi^{-1}P(w_2)P(w_1)}} dx \\
\Leftrightarrow P(error) &\leq \int \frac{2\pi^{-1}P(w_2)P(w_1)}{(1+(x+\mu)^2)P(w_1) + (1+(x-\mu)^2)P(w_2)} dx \\
\Leftrightarrow P(error) &\leq \int \frac{2\pi^{-1}P(w_2)P(w_1)}{(x^2+2x\mu+\mu^2+1)P(w_1) + (x^2-2x\mu+\mu^2+1)P(w_2)} dx \\
\Leftrightarrow P(error) &\leq \int \frac{2\pi^{-1}P(w_2)P(w_1)}{(P(w_1)+P(w_2))x^2 + (P(w_1)-P(w_2))2\mu x + (P(w_1)\mu^2 + P(w_2)\mu^2 + P(w1) + P(w_2))} dx
\end{aligned}
$$

We can now take out the numerator of the integral and use the following equation:

$$
\int \frac{1}{ax^2+bx+c} dx = \frac{2\pi}{\sqrt{4ac-b^2}} \tag{2}
$$

with:

$$
\begin{aligned}
a &= P(w_1) + P(w_2) \\
b &= (P(w_1) - P(w_2))2\mu \\
c &= P(w_1)\mu^2 + P(w_2)\mu^2 + P(w1) + P(w_2)
\end{aligned}
$$

because

$$
\begin{aligned}
& b^2 < 4ac \\
\Leftrightarrow\ & 0 < 4ac - b^2 \\
\Leftrightarrow\ & 0 < 4(P(w_1) + P(w_2))(P(w_1)\mu^2 + P(w_2)\mu^2 + P(w1) + P(w_2)) - ((P(w_1) - P(w_2))2\mu)^2 \\
\Leftrightarrow\ & 0 < 4P(w_1)^2\mu^2 + 4P(w_1)P(w_2)\mu^2 + 4P(w_1)^2 + 4P(w_1)P(w_2) \\
& \quad + 4P(w_1)P(w_2)\mu^2 + 4P(w_2)^2\mu^2 + 4P(w_2)P(w_1) + 4P(w_2)^2 \\
& \quad - (4P(w_1)^2\mu^2 - 8P(w_1)P(w_2)\mu^2 + 4P(w_2)^2\mu^2) \\
\Leftrightarrow\ & 0 < 16P(w_1)P(w_2)\mu^2 + 8P(w_1)P(w_2) + 4P(w_1)^2 + 4P(w_2)^2
\end{aligned}
$$

and this holds since $P(w_1), P(w_2) \in [0, 1]$ and $P(w_1) + P(w_2) = 1$.
We now already calculated $4ac - b^2$ in the steps before and just need to use (4) to proceed where we stopped before introducing equation 4:

$$
\begin{aligned}
\Leftrightarrow\ & P(error) \leq \int \frac{2\pi^{-1}P(w_2)P(w_1)}{(P(w_1) + P(w_2))x^2 + (P(w_1) - P(w_2))2\mu x + (P(w_1)\mu^2 + P(w_2)\mu^2 + P(w1) + P(w_2))} dx \\
\Leftrightarrow\ & P(error) \leq 2\pi^{-1}P(w_2)P(w_1)\frac{2\pi}{\sqrt{16P(w_1)P(w_2)\mu^2 + 8P(w_1)P(w_2) + 4P(w_1)^2 + 4P(w_2)^2}} \\
\Leftrightarrow\ & P(error) \leq \frac{4P(w_1)P(w_2)}{\sqrt{4((4\mu^2 + 2)P(w_1)P(w_2) + P(w_1)^2 + P(w_2)^2)}} \\
\Leftrightarrow\ & P(error) \leq \frac{4P(w_1)P(w_2)}{\sqrt{4}\sqrt{((4\mu^2 + 2)P(w_1)P(w_2) + P(w_1)^2 + P(w_2)^2)}} \\
\Leftrightarrow\ & P(error) \leq \frac{2P(w_1)P(w_2)}{\sqrt{P(w_1)^2 + (4\mu^2 + 2)P(w_1)P(w_2) + P(w_2)^2}}
\end{aligned}
$$

c)

According to informations form chapter 2.8 in Pattern Classification:
For both cases we can find the upper-bound numericaly via the Chernoff Bound or the Bhattacharyya Bound, this may lead to a tighter upper bound than the analytical version, which is even harder to approximate due to the discontinues nature of the integrals. For the high-dimensional space the Bhattacharyya Bound might be better, because it is computationally less expensive than the Chernoff Bound.
Besides from that, it's possible to compute an approximation if we have a data-set that is large (in terms of datapoints, not dimensions) enough to contain the class characteristics. We could use for example nested cross validation to compute and evaluate multiple classifiers with different parameters and then take the mean error (on the test set) as the approximation. This is especially practical for low dimensional data but becomes computationally expensive if the number of dimensions grows, because the number of possible decision boundaries grows and the need of data to cover the classes characteristics even grows exponentially.
For high dimensional data one could try to estimate the prior by analyzing the dataset and use the smaller prior as an approximation of the error. But this would rather be the highest error a classifier should have, since only picking the class with the higher prior should yield the approximated error rate. So this would almost always be a poor approximation for the bayes error, but at least give a lower bound of the error that a classifier should produce.

Exercise 2

a)

The data is generated by the univariate Laplacian Distrubution:

$$p(x \mid w_1) = \frac{1}{2\sigma} \exp\left(-\frac{\mid x-\mu \mid}{\sigma}\right) \text{ and } p(x \mid w_2) = \frac{1}{2\sigma} \exp\left(-\frac{\mid x+\mu \mid}{\sigma}\right)$$

To get the optimal decision boundary we have to solve $P(w_1 \mid x) = P(w_2 \mid x)$ for x.

$$\begin{aligned}
& P(w_1 \mid x) = P(w_2 \mid x) \\
\Leftrightarrow \quad & \frac{p(x \mid w_1)P(w_1)}{p(x)} = \frac{p(x \mid w_2)P(w_2)}{p(x)} \\
\Leftrightarrow \quad & p(x \mid w_1)P(w_1) = p(x \mid w_2)P(w_2) \\
\Leftrightarrow \quad & \frac{1}{2\sigma} \exp\left(-\frac{\mid x-\mu \mid}{\sigma}\right) P(w_1) = \frac{1}{2\sigma} \exp\left(-\frac{\mid x+\mu \mid}{\sigma}\right) P(w_2) \\
\Leftrightarrow \quad & \exp\left(-\frac{\mid x-\mu \mid}{\sigma}\right) P(w_1) = \exp\left(-\frac{\mid x+\mu \mid}{\sigma}\right) P(w_2) \\
\Leftrightarrow \quad & \ln(\exp\left(-\frac{\mid x-\mu \mid}{\sigma}\right) P(w_1)) = \ln(\exp\left(-\frac{\mid x+\mu \mid}{\sigma}\right) P(w_2)) \\
\Leftrightarrow \quad & -\frac{\mid x-\mu \mid}{\sigma} + \ln(P(w_1)) = -\frac{\mid x+\mu \mid}{\sigma} + \ln(P(w_2)) \\
\Leftrightarrow \quad & \ln(P(w_1)) - \ln(P(w_2)) = \frac{\mid x-\mu \mid}{\sigma} - \frac{\mid x+\mu \mid}{\sigma} \\
\Leftrightarrow \quad & \ln\left(\frac{P(w_1)}{P(w_2)}\right) = \frac{\mid x-\mu \mid - \mid x+\mu \mid}{\sigma} \\
\Leftrightarrow \quad & \sigma \ln\left(\frac{P(w_1)}{P(w_2)}\right) = \mid x-\mu \mid - \mid x+\mu \mid
\end{aligned}$$

There are now 4 cases to look at:

1. $x < \mu$ and $-x < \mu$

$$\begin{aligned}
\sigma \ln\left(\frac{P(w_1)}{P(w_2)}\right) &= -(x-\mu) - (x+\mu) \\
&= -2x \\
-\frac{\sigma}{2} \ln\left(\frac{P(w_1)}{P(w_2)}\right) &= x
\end{aligned}$$

2. $x > \mu$ and $-x > \mu$

$$\begin{aligned}
\sigma \ln\left(\frac{P(w_1)}{P(w_2)}\right) &= x-\mu+x+\mu \\
&= 2x \\
\frac{\sigma}{2} \ln\left(\frac{P(w_1)}{P(w_2)}\right) &= x
\end{aligned}$$

3. $x \leq \mu$ and $-x \geq \mu$

$$\begin{aligned}
\sigma \ln\left(\frac{P(w_1)}{P(w_2)}\right) &= -(x-\mu)+x+\mu \\
&= 2\mu
\end{aligned}$$

4. $x \geq \mu$ and $-x \leq \mu$

$$\begin{aligned}\sigma \ln\left(\frac{P(w_1)}{P(w_2)}\right) &= (x-\mu) - (x+\mu) \\ &= -2\mu\end{aligned}$$

Equations (3) and (4) tell us, that the boundary must lie between the mean of both classes (there is no boundary for the range of x outside of that). The first equation gives us the boundary if we have $\mu > 0$ and the second one if we have $\mu < 0$.

b)

To satisfy the constrain of $P(w_2 \mid x) = P(error \mid x)$ for all x, w_1 and w_2 must have the same density functions and thus

$$P(x \mid w_1) = P(x \mid w_2)$$

Visually it would mean that the area of the curve of $P(x \mid w_2)$ is completly included inside the area of the curve of $P(x \mid w_1)$. When varying the mean μ, both curves move away from each other, which means it always has to stay 0, else it's not possible to satisfy the constraint. The scale σ can take any value that is larger than zero, since it has the same effect on both distributions. $P(w_1)$ must always be larger than $P(w_2)$ (or else $P(w_1 \mid x) \leq P(w_2 \mid x)$ for some x and so $P(error \mid x) \neq P(w_2 \mid x)$ for some x):

$$\begin{aligned} P(error \mid x) &= P(w_2 \mid x) \\ \Leftrightarrow P(w_2 \mid x) &= min[P(w_1 \mid x), P(w_2 \mid x)] \\ \Leftrightarrow P(w_2 \mid x) &< P(w_1 \mid x) \\ \Leftrightarrow \frac{P(x \mid w_2)P(w_2)}{P(x)} &< \frac{P(x \mid w_1)P(w_1)}{P(x)} \\ \Leftrightarrow P(x \mid w_2)P(w_2) &< P(x \mid w_1)P(w_1) \\ \Leftrightarrow P(w_2) &< P(w_1) \end{aligned}$$

Formally all possible value-tuples are:

$$\{(P(w_1), P(w_2), \mu, \sigma) \mid P(w_1) > P(w_2) \wedge \mu = 0 \wedge \sigma \in \mathbb{R}^+\}$$

c)

The data is generated by the univariate Gaussian Distrubution:

$$p(x \mid w_1) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \text{ and } p(x \mid w_2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x+\mu)^2}{2\sigma^2}\right)$$

To get the optimal decision boundary we have to solve $P(w_1 \mid x) = P(w_2 \mid x)$ for x.

$$
\begin{aligned}
& P(w_1 \mid x) = P(w_2 \mid x) \\
\Leftrightarrow \quad & \frac{p(x \mid w_1)P(w_1)}{p(x)} = \frac{p(x \mid w_2)P(w_2)}{p(x)} \\
\Leftrightarrow \quad & \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) P(w_1) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x+\mu)^2}{2\sigma^2}\right) P(w_2) \\
\Leftrightarrow \quad & \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) P(w_1) = \exp\left(-\frac{(x+\mu)^2}{2\sigma^2}\right) P(w_2) \\
\Leftrightarrow \quad & -\frac{(x-\mu)^2}{2\sigma^2} + \ln\left(P(w_1)\right) = -\frac{(x+\mu)^2}{2\sigma^2} + \ln\left(P(w_2)\right) \\
\Leftrightarrow \quad & \ln\left(\frac{P(w_1)}{P(w_2}\right) = \frac{(x-\mu)^2 - (x+\mu)^2}{2\sigma^2} \\
\Leftrightarrow \quad & 2\sigma^2 \ln\left(\frac{P(w_1)}{P(w_2)}\right) = x^2 - 2x\mu + \mu^2 - x^2 - 2x\mu - \mu^2 \\
\Leftrightarrow \quad & 2\sigma^2 \ln\left(\frac{P(w_1)}{P(w_2)}\right) = -4x\mu \\
\Leftrightarrow \quad & \frac{2\sigma^2}{-4\mu} \ln\left(\frac{P(w_1)}{P(w_2)}\right) = x
\end{aligned}
$$

These are the optimal decision boundaries under the restriction that $\mu \neq 0$, which leads us to the second part of the exercise.
Just as in case of the Laplacian distribution, the mean μ has to be zero and $P(w_1)$ has to be larger than $P(w_2)$ in order that $P(error \mid x) = P(w_2 \mid x)$ for all x. Again we can obtain these possible value tuples:

$$\{(P(w_1), P(w_2), \mu, \sigma) \mid P(w_1) > P(w_2) \wedge \mu = 0 \wedge \sigma \in \mathbb{R}^+\}$$