6.7900 - Assignment 1

1.1: We have

$$\frac{\partial}{\partial \mu} \log p(\mathcal{D}|\mu, \sigma^2) = \sum_{n=1}^{N}(x^{(n)} - \mu) \overset{\text{set}}{=} 0 \implies \sum_{n=1}^{N} \mu_{\mathbf{ml}} = \sum_{n=1}^{N} x^{(n)} \implies \mu_{\mathbf{ml}} = \frac{1}{N}\sum_{n=1}^{N} x^{(n)}$$

$$\frac{\partial}{\partial \sigma^2} \log p(\mathcal{D}|\mu, \sigma^2) = \sigma^{-3}\sum_{n=1}^{N}(x^{(n)} - \mu)^2 - \sigma^{-1}N \overset{\text{set}}{=} 0 \implies \sigma_{\mathbf{ml}}^{-2}\sum_{n=1}^{N}(x^{(n)} - \mu_{\mathbf{ml}})^2 = N$$

$$\implies \sigma_{\mathbf{ml}}^2 = \frac{1}{N}\sum_{n=1}^{N}(x^{(n)} - \mu_{\mathbf{ml}})^2.$$

1.2: An example is $\mathcal{D} = \{1\}$, which has mean 1 and variance 0. The term $-\frac{1}{2\sigma^2}(x - \mu)^2 = 0$, while the term $-\frac{1}{2}\log \sigma^2$ and thus the entire log-likelihood are unbounded from above.

1.3: Assuming that the logarithm function is in base e,

$$\begin{aligned}
\mu_{\mathbf{ml}} =& \frac{1}{6}(0.9 + 1 + 1.1 + 1.2 + 3 + 3.1) = 1.7167 \\
\log p(\mathcal{D}_0|\mu_{\mathbf{ml}}) =& -2[(0.9 - 1.7167)^2 + (1 - 1.7167)^2 + (1.1 - 1.7167)^2 + (1.2 - 1.7167)^2 \\
& + (3 - 1.7167)^2 + (3.1 - 1.7167)^2] - 3\log(0.5\pi) \\
=& -12.1314.
\end{aligned}$$

1.4: $\mu_{\mathbf{ml}} = 1.7167$ as before, while

$$\sigma_{\mathbf{ml}}^2 = \frac{1}{6}[\sum_{n=1}^{6}(x^{(n)} - 1.7167)^2] = 0.8981$$

$$\log p(\mathcal{D}_{\prime}|\mu_{\mathbf{ml}}, \sigma_{\mathbf{ml}}^2) = -2[\sum_{n=1}^{6}(x^{(n)} - 1.7167)^2] - 3\log(1.7961\pi) = -15.9677.$$

An advantage is that $\sigma_{\mathbf{ml}}^2$ is likely more reflective of the observed data, while the associated disadvantages are a lower log-likelihood and extra computing cost compared to using the provided variance.

2.1: We can model the data using a Bernoulli distribution; in particular, $x^{(n)} \overset{iid}{\sim} \text{Bernoulli}(\theta)$ for $n \in \{1, 2, 3\}$, where $x^{(n)} = \begin{cases} 1 & \text{if } a^{(n)} = \text{“}H\text{”} \\ 0 & \text{otherwise} \end{cases}$ and $\theta \in [0, 1]$ is the probability of a heart attack. Then, the MLE is $\theta_{\mathbf{ml}} = \frac{1}{3}\sum_{n=1}^{3} x^{(n)} = 0$, or that the probability of having a heart attack is 0.

2.6: The final posterior is $p(Q|y_1, y_2) = \frac{p(y_1,y_2|Q)p(Q)}{p(y_1,y_2)} = \frac{p(y_2|Q)}{p(y_2|y_1)}\frac{p(y_1|Q)p(Q)}{p(y_1)} = \frac{p(y_2|Q,y_1)}{p(y_2|y_1)}p(Q|y_1)$, where $p(Q|y_1)$ is the posterior of observing y_1 first. This is also equal to $\frac{p(y_1,y_2|Q)p(Q)}{p(y_1,y_2)} = \frac{p(y_1|Q)}{p(y_1|y_2)}\frac{p(y_2|Q)p(Q)}{p(y_2)} = \frac{p(y_1|Q)}{p(y_1|y_2)}p(Q|y_2)$, where $p(Q|y_2)$ is the posterior of observing y_2 first, showing that the order of the patients does not affect the final posterior.

3.1: For $y = 1500$, the prior is $p(\theta) = \mathcal{N}(\mu_0, \sigma_0^2)$ and the likelihood is $p(y|\theta) = \mathcal{N}(\theta, \sigma_D^2)$, so the posterior is

$$\begin{aligned}
p(\theta|y) \propto_\theta p(\theta)p(y|\theta) \propto_\theta \exp[-\frac{1}{2\sigma_0^2}(\theta-\mu_0)^2 - \frac{1}{2\sigma_D^2}(y-\theta)^2] \\
= \exp[-(\frac{1}{2\sigma_0^2}+\frac{1}{2\sigma_D^2})\theta^2 + (\frac{\mu_0}{\sigma_0^2}+\frac{y}{\sigma_D^2})\theta - (\frac{\mu_0^2}{2\sigma_0^2}+\frac{y^2}{2\sigma_D^2})] \\
= \exp[-\frac{\sigma_0^2+\sigma_D^2}{2\sigma_0^2\sigma_D^2}(\theta - \frac{\mu_0\sigma_D^2+y\sigma_0^2}{\sigma_0^2+\sigma_D^2})^2]
\end{aligned}$$

which is the distribution $\mathcal{N}(\frac{\mu_0\sigma_D^2+y\sigma_0^2}{\sigma_0^2+\sigma_D^2}, \frac{\sigma_0^2\sigma_D^2}{\sigma_0^2+\sigma_D^2})$. Plugging the numerical values into the posterior yields $\mathcal{N}(1260, 30)$.

3.2: The posterior mean is $\frac{\mu_0\sigma_D^2+y\sigma_0^2}{\sigma_0^2+\sigma_D^2} = \mu_0(\frac{\sigma_D^2}{\sigma_0^2+\sigma_D^2}) + y(\frac{\sigma_0^2}{\sigma_0^2+\sigma_D^2})$, showing that it is indeed a weighted average.

3.3: The prior variance is larger than the posterior variance since $50(cc)^2 > 30(cc)^2$.

4.2: Define $f(x) = \frac{p(y=1|x)}{p(y=0|x)} = \frac{p(x|y=1)p(y=1)}{p(x|y=0)p(y=0)}$ such that $p(y=1|x) = f(x)p(y=0|x)$. Then,

$$\begin{aligned}
p(y=0|x) + p(y=1|x) = 1 \iff p(y=0|x) + f(x)p(y=0|x) = 1 \\
\iff p(y=0|x)(1+f(x)) = 1 \iff p(y=0|x) = \frac{1}{1+f(x)} \text{ and } p(y=1|x) = \frac{f(x)}{1+f(x)}
\end{aligned}$$

for $f(x) = \frac{p(y=1)}{p(y=0)}\sqrt{\frac{|\Sigma_0|}{|\Sigma_1|}}\exp[-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)]$.

4.3: The decision boundary is

$$\begin{aligned}
0 =& \log p(y=1|x) - \log p(y=0|x) \\
=& \log p(x|y=1) + \log p(y=1) - \log p(x) - \log p(x|y=0) - \log p(y=0) + \log p(x) \\
=& \log p(x|y=1) - \log p(x|y=0) + \log p(y=1) - \log p(y=0) \\
=& -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_1| - \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + \frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma_0| \\
&+ \frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) + \log\frac{p(y=1)}{p(y=0)} \\
=& -\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) + \frac{1}{2}\log\frac{|\Sigma_0|}{|\Sigma_1|} + \log\frac{p(y=1)}{p(y=0)}
\end{aligned}$$

which is a quadratic function of x. If $\Sigma_0 = \Sigma_1$, this function becomes

$$\begin{aligned}
0 =& -\frac{1}{2}(x-\mu_1)^T\Sigma_0^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) + \frac{1}{2}\log\frac{|\Sigma_0|}{|\Sigma_0|} + \log\frac{p(y=1)}{p(y=0)} \\
=& -x^T\Sigma_0^{-1}x + 2\mu_1^T\Sigma_0^{-1}x - \mu_1^T\Sigma_0^{-1}\mu_1 + x^T\Sigma_0^{-1}x - 2\mu_0^T\Sigma_0^{-1}x + \mu_0^T\Sigma_0^{-1}\mu_0 + \log\frac{p(y=1)}{p(y=0)} \\
=& 2(\mu_1^T\Sigma_0^{-1} - \mu_0^T\Sigma_0^{-1})x - \mu_1^T\Sigma_0^{-1}\mu_1 + \mu_0^T\Sigma_0^{-1}\mu_0 + \log\frac{p(y=1)}{p(y=0)}
\end{aligned}$$

which is a linear function of x.

4.4: For $x = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^T$, the numerical form is $g(x) = -3.3307\cdot 10^{-16}x_1^2 - 2.8479\cdot 10^{-17}x_1x_2 + 5x_1 + 5.0903\cdot 10^{-16}x_2 - 12.5$ and the associated decision rule is $y = \begin{cases} 1 & \text{if } g(x) \geq 0 \\ 0 & \text{otherwise} \end{cases}$. Based on this prediction, 500 points have $y = 0$.