Problem 1: Logistic Regression

a) The the negative log likelihood is

$$\begin{aligned}\mathcal{L}(\mathbf{w}) &= -\log(\prod_{i=1}^{n} P(Y = y_i|\mathbf{X} = \mathbf{x_i})) \\ &= -\sum_{i=1}^{n} \log(P(Y = y_i|\mathbf{X} = \mathbf{x_i})) \\ &= -\sum_{i=1}^{n}(y_i \log(\sigma(\mathbf{w}^T\mathbf{x_i})) + (1-y_i)\log((1-\sigma(\mathbf{w}^T\mathbf{x_i})))\end{aligned} \tag{1}$$

b) The first derivative is

$$\frac{\partial \mathcal{L}(\mathbf{w})}{\partial \mathbf{w}} = \sum_{i=1}^{n}(\sigma(\mathbf{w}^T\mathbf{x_i}) - y_i)\mathbf{x_i} \tag{2}$$

Thus the Gradient Descent Update Rule is:

$$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta\sum_{i=1}^{n}(\sigma(\mathbf{w}^T\mathbf{x_i}) - y_i)\mathbf{x_i} \tag{3}$$

And update rule can find the global minimum, since the Hessian is semi-definite, which is proved as follows

$$H = \frac{\partial^2 \mathcal{L}(\mathbf{w})}{\partial \mathbf{w}\partial \mathbf{w}^T} = \sum_{i=1}^{n}\sigma(\mathbf{w}^T\mathbf{x_i})(1-\sigma(\mathbf{w}^T\mathbf{x_i}))\mathbf{x_i}\mathbf{x_i}^T \tag{4}$$

Then for any vector $\mathbf{u}$, we have

$$\begin{aligned}\mathbf{u}^T H\mathbf{u} &= \sum_{i=1}^{n}(\sigma(\mathbf{w}^T\mathbf{x_i}) - y_i)\mathbf{u}^T\mathbf{x_i}\mathbf{x_i}^T\mathbf{u} \\ &= \sum_{i=1}^{n}\sigma(\mathbf{w}^T\mathbf{x_i})(1-\sigma(\vec{w}^T\mathbf{x_i}))||\mathbf{x}_i^T\mathbf{u}||_2^2 \geq 0\end{aligned} \tag{5}$$

c) Let I_{lk} be an indicator function, where $I_{lk} = 1$ if $Y^l = k$, otherwise $I_{lk} = 0$. Then the negative log likelihood function is

$$\begin{aligned}\mathcal{L}(\mathbf{w_1}, ..., \mathbf{w_k}) &= -\log[\prod_{l=1}^{n}\prod_{k=1}^{K} P(Y^l = k|\mathbf{X}^l = \mathbf{x})^{I_{lk}}] \\ &= -\sum_{l=1}^{n}\sum_{k=1}^{K} I_{lk}[\mathbf{w}_k^T\mathbf{x}^l - \log(\sum_r \exp(\mathbf{w}_r^T\mathbf{x}^l))]\end{aligned} \tag{6}$$

d) Taking derivative with pespective to $\mathbf{w}_i$:

$$\partial\frac{\mathcal{L}(\mathbf{w_1}, ..., \mathbf{w_k})}{\partial \mathbf{w}_i} = -\sum_{l=1}^{D}[I_{li}\mathbf{x}^l - \frac{\mathbf{x}^l\exp(\mathbf{w}_i^T\mathbf{x}^l)}{\sum_r \exp(\mathbf{w}_i^T\mathbf{x}^l)}] \tag{7}$$

which can be simplified as

$$\partial \frac{\mathcal{L}(\mathbf{w_1}, ..., \mathbf{w_k})}{\partial \mathbf{w}_i} = -\sum_{l=1}^{D} [I_{li} - P(Y^l = i | \mathbf{X}^l)] \mathbf{x}^l \tag{8}$$

And the update rule is

$$\mathbf{w}_i \leftarrow \mathbf{w}_i + \eta \sum_{l=1}^{D} [I_{li} - P(Y^l = i | \mathbf{X}^l)] \mathbf{x}^l \tag{9}$$

Problem 2: Linear/Gaussian Discriminant Analysis

(a) The log likelihood function is

$$\begin{aligned}
\log P(\mathcal{D}) &= \sum_n \log p(x_n, y_n) \\
&= \sum_{n:y_n=1} \log(p_1 \frac{1}{\sqrt{2\pi}\sigma_1} \exp(-\frac{(x_n - \mu_1)^2}{2\sigma_1^2})) \\
&+ \sum_{n:y_n=2} \log(p_2 \frac{1}{\sqrt{2\pi}\sigma_2} \exp(-\frac{(x_n - \mu_2)^2}{2\sigma_2^2}))
\end{aligned} \tag{10}$$

Take partial derivative of $\log P(\mathcal{D})$ w.r.t. p_1, p_2, μ_1, μ_2, σ_1, σ_2 respectively, and set it to 0, we have

$$p_1^* = \frac{\sum_n I(y_n = 1)}{N}$$

$$p_2^* = \frac{\sum_n I(y_n = 2)}{N}$$

$$\mu_1^* = \frac{\sum_{n:y_n=1} x_n}{\sum_n I(y_n = 1)}$$

$$\mu_2^* = \frac{\sum_{n:y_n=2} x_n}{\sum_n I(y_n = 2)}$$

$$\sigma_1^{2*} = \frac{\sum_{n:y_n=1} (x_n - \mu_1^*)^2}{\sum_n : I(y_n = 1)}$$

$$\sigma_2^{2*} = \frac{\sum_{n:y_n=2} (x_n - \mu_2^*)^2}{\sum_n : I(y_n = 2)}$$

(b) **Gaussian and Linear Discriminant Analysis**

Note: since the variances in these two classes are the same, both Gaussian and Linear Discriminant Analysis are the same, which is a linear boundary.

(Method 1) For class 1, Gaussian: $\mu_1 = (0, 0, \cdots, 0)$, $\mathbf{\Sigma}_1$ is a diagonal matrix with $\mathbf{\Sigma}_1(i, i) = \sigma^2$. For class 2, Gaussian: $\mu_2 = (0, 0, \cdots, 0, \delta, \delta, \cdots, \delta)$, $\mathbf{\Sigma}_2 = \mathbf{\Sigma}_1$. The solution of LDA is $\mathbf{w} = \mathbf{\Sigma}_1^{-1}(\mu_1 - \mu_2)$. When δ changes, the solution does change. But the direction of $\mathbf{w}$ remains the same.

(Method 2) Let Y be a variable representing two class labels 0 and 1. Then given a sample $\vec{x}$, we know

$$\begin{aligned}
\log\frac{p(Y=0|\vec{x})}{p(Y=1|\vec{x})} &= \log\frac{p(\vec{x}|Y=0)P(Y=0)}{p(\vec{x}|Y=1)P(Y=1)} \\
&= \log\frac{P(Y=0)\prod_{i=1}^{2D} p(x_i|Y=0)}{P(Y=1)\prod_{i=1}^{2D} p(x_i|Y=1)} \\
&= (\log P(Y=0) + \sum_{i=1}^{2D}\log p(x_i|Y=0)) - (\log P(Y=1) + \sum_{i=1}^{2D}\log p(x_i|Y=1)) \\
&= \log P(Y=0) - \log P(Y=1) + \sum_{i=1}^{2D}\log\frac{1}{\sqrt{2\pi}\sigma}exp(\frac{-x_i^2}{2\sigma^2}) \\
&\quad - \sum_{i=1}^{D}\log\frac{1}{\sqrt{2\pi}\sigma}exp(\frac{-x_i^2}{2\sigma^2}) - \sum_{i=D+1}^{2D}\log\frac{1}{\sqrt{2\pi}\sigma}exp(\frac{-(x_i-\delta)^2}{2\sigma^2}) \\
&= \log P(Y=0) - \log P(Y=1) + \sum_{i=D+1}^{2D}\log\frac{1}{\sqrt{2\pi}\sigma}exp(\frac{-x_i^2}{2\sigma^2}) \\
&\quad - \sum_{i=D+1}^{2D}\log\frac{1}{\sqrt{2\pi}\sigma}exp(\frac{-(x_i-\delta)^2}{2\sigma^2}) \\
&= \log P(Y=0) - \log P(Y=1) + \sum_{i=D+1}^{2D}\frac{(x_i-\delta)^2}{2\sigma^2} - \sum_{i=D+1}^{2D}\frac{x_i^2}{2\sigma^2} \\
&= \log\frac{P(Y=0)}{P(Y=1)} + \frac{D\delta^2}{2\sigma^2} - \frac{\delta}{\sigma^2}\sum_{i=D+1}^{2D} x_i
\end{aligned} \tag{11}$$

Thus, if $\log\frac{P(Y=0)}{P(Y=1)} + \frac{D\delta^2}{2\sigma^2} - \frac{\delta}{\sigma^2}\sum_{i=D+1}^{2D} x_i >= 0$, predict as $Y = 0$; Otherwise, predict as $Y = 1$. Also, note this is a linear decision boundary.

(c) Let $p_1 = p(y=1), p_2 = p(y=2)$ be the prior distributions of classes.

$$
\begin{aligned}
& p(y=1|\mathbf{x}) && (12)\\
=& \frac{p(\mathbf{x}|y=1)p_1}{p(\mathbf{x}|y=1)p_1 + p(\mathbf{x}|y=2)p_2} && (13)\\
=& \frac{\exp\left[-(\mathbf{x}-\mu_1)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_1)\right] p_1}{\exp\left[-(\mathbf{x}-\mu_1)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_1)\right] p_1 + \exp\left[-(\mathbf{x}-\mu_2)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_2)\right] p_2} && (14)\\
=& \frac{1}{1+\exp\left[-(\mathbf{x}-\mu_1)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_1) + (\mathbf{x}-\mu_2)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_2)\right] p_2/p_1} && (15)\\
=& \frac{1}{1+\exp\left[(\mu_2-\mu_1)^\top \mathbf{\Sigma}^{-1}\mathbf{x} + \mu_1^\top \mathbf{\Sigma}^{-1}\mu_2/2 - \mu_2^\top \mathbf{\Sigma}^{-1}\mu_2/2\right] p_2/p_1} && (16)\\
=& \frac{1}{1+\exp\left[(\mu_2-\mu_1)^\top \mathbf{\Sigma}^{-1}\mathbf{x} + \mu_1^\top \mathbf{\Sigma}^{-1}\mu_2/2 - \mu_2^\top \mathbf{\Sigma}^{-1}\mu_2/2 + \log p_2 - \log p_1\right]} && (17)\\
=& \frac{1}{1+\exp\left[\theta^\top \mathbf{x} + \gamma\right]} && (18)
\end{aligned}
$$

where $\theta = (\mu_2-\mu_1)^\top \mathbf{\Sigma}^{-1}$, $\gamma = \mu_1^\top \mathbf{\Sigma}^{-1}\mu_1/2 - \mu_2^\top \mathbf{\Sigma}^{-1}\mu_2/2 + \log p_2 - \log p_1$.

Problem 3: Perceptron and Online Learning

Conditions: From any current step parameters $\boldsymbol{w}_i$ we want to update the classifier such that $\text{sign}(\boldsymbol{w}_{i+1}^\top \boldsymbol{x}_{i+1}) = y_{i+1}$. However, we also would like $\|\boldsymbol{w}_{i+1} - \boldsymbol{w}_i\|_2$ to be small.

Solution: If $y_{i+1} = \text{sign}(\boldsymbol{w}_i^\top \boldsymbol{x}_{i+1})$, then let $\boldsymbol{w}_{i+1} = \boldsymbol{w}_i$ (do nothing). Otherwise, we know $\boldsymbol{w}_i^\top \boldsymbol{x}_{i+1} y_{i+1} < 0$. Now we need the *smallest* amount of movement such that then point $\boldsymbol{x}_{i+1}$ is on the correct side of the plane:

$$\boldsymbol{w}_{i+1} = \arg\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}_i\|_2^2 \quad \text{s.t.} \quad \boldsymbol{w}^\top \boldsymbol{x}_{i+1} y_{i+1} = 0$$

Tthe equality constraint is due to the smallest amount of movement we need.

Writing the Lagrangian, yields

$$\mathcal{L}(\boldsymbol{w}, \lambda) = \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_i)^\top (\boldsymbol{w} - \boldsymbol{w}_i) + \lambda \boldsymbol{w}^\top \boldsymbol{x}_{i+1} y_{i+1}$$

Take a derivative w.r.t. $\boldsymbol{w}$

$$\frac{\partial}{\partial \boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, \lambda) = (\boldsymbol{w} - \boldsymbol{w}_i) - \lambda \boldsymbol{x}_{i+1} y_{i+1} = 0$$

$$\boldsymbol{w} = \lambda \boldsymbol{x}_{i+1} y_{i+1} + \boldsymbol{w}_i$$

Transpose and multiply by $\boldsymbol{x}_{i+1} y_{i+1}$ on both sides, then apply the equality $\boldsymbol{w}^\top \boldsymbol{x}_{i+1} y_{i+1} = 0$:

$$\boldsymbol{w}^\top \boldsymbol{x}_{i+1} y_{i+1} = 0 = \lambda (\boldsymbol{x}_{i+1} y_{i+1})^\top (\boldsymbol{x}_{i+1} y_{i+1}) + \boldsymbol{w}_i^\top (\boldsymbol{x}_{i+1} y_{i+1})$$

$$\lambda = -\frac{\boldsymbol{w}_i^\top(\boldsymbol{x}_{i+1}y_{i+1})}{\|\boldsymbol{x}_{i+1}\|_2^2}$$

Plug back in, and let this be the update rule.

$$\boldsymbol{w}_{i+1} = \boldsymbol{w}_i - \frac{\boldsymbol{w}_i^\top \boldsymbol{x}_{i+1}}{\|\boldsymbol{x}_{i+1}\|_2^2}\boldsymbol{x}_{i+1}$$

Geometrically this is the same as finding some vector that is perpendicular to $\boldsymbol{x}_{i+1}$ and projecting $\boldsymbol{w}_i$ onto it, taking the projection as the new normal vector.