6140 Machine Learning Assignment 2

Shreeti Shrestha

Due Date: 18 October 2025

1. Logistic Regression: In the logistic regression for binary classification ($y \in \{0, 1\}$), we defined $p(y = 1|x) = \sigma(w^\top \phi(x))$, where the sigmoid function is defined as

$$\sigma(z) \triangleq \frac{1}{1 + e^{-z}}.$$

Assume we have trained the logistic regression model using a given dataset and have learned w. Let x^n be a test sample.

1. Assume $w^\top \phi(x^n) > -0.4$. What can you conclude about the class to which x^n belongs? Provide details of your justification or derivations.

   Answer: Let $z = \mathrm{w}^\top \phi(x^n)$. Then,

$$p(y = 1|x) = \sigma(w^\top \phi(x^n)) = \frac{1}{1 + e^{-w^\top \phi(x^n)}}.$$

At $(w^\top \phi(x^n)) = -0.4$,
$p(y = 1|x) = \frac{1}{1+e^{-(-0.4)}}$

$= \frac{1}{1+e^{0.4}}$

$= \frac{1}{1+1.4918}$

$= \frac{1}{2.4918}$

≈ 0.401

Since the sigmoid function is monotonically increasing, if $z_1 > z_2$, then $\sigma(z_1) > \sigma(z_2)$. Therefore, $p(y = 1|x) > 0.401$ for $w^\top \phi(x^n) > -0.4$. The sigmoid function intersects the y-axis at 0.5. So, in binary logistic regression:

$$\hat{y} = \begin{cases} 1, & \text{if } p(y = 1 \mid x) \geq 0.5, \\ 0, & \text{otherwise.} \end{cases}$$

$p(y = 1 \mid x_n) > 0.401$ but we do not know the exact value (could still be less than 0.5). Since we don't know if it is greater than or equal to 0.5, so we cannot guarantee the model predicts class 1. So, we cannot conclude which class x^n belongs to.

2. Assume $\frac{1}{1+e^{w^\top \phi(x^n)}} = 0.72$. What is the probability that x^n belongs to class 1? Explain.

   Answer:

   For binary logistic regression, $\frac{1}{1+e^{w^\top \phi(x^n)}} = p(y = 0|x)$.

   So, $p(y = 0|x) = 0.72$. Since the probabilities must add up to 1 for this binary classification, $p(y = 1|x) = 1 - 0.72 = 0.28$.

   So, the probability that x_n belongs to class 1 is 0.28.

2. Probability: Consider three random variables X, Y, Z. Show that the following hold.

1. $P(X_1, X_2, \ldots, X_N) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \times \cdots \times P(X_N|X_1, X_2, \ldots, X_{N-1})$.

Answer:

Conditional probability for 2 random variables X and Y is:
$P(Y|X) = \frac{P(X,Y)}{P(X)}$
or, $P(X, Y) = P(Y|X) * P(X) : (equation1)$

Using $equation1$ for $P(X_1, X_2, \ldots, X_N)$ where $(X_1, X_2, \ldots, X_{N-1}) = X$ and $X_N = Y$, we get:

$P(X_1, X_2, \ldots, X_N) = P(X_N|X_1, X_2, \ldots, X_{N-1}) \times P(X_1, X_2, \ldots, X_{N-1})$ (now, $Y = X_{N-1}$)

$= P(X_N|X_1, X_2, \ldots, X_{N-1}) \times P(X_{N-1}|X_1, X_2, \ldots, X_{N-2}) \times P(X_1, X_2, \ldots, X_{N-2})$

doing this recursively for $Y = X_{N-2}, Y = X_{N-3} ... Y = X_1$, we get:

$= P(X_N|X_1, X_2, \ldots, X_{N-1}) \times (X_{N-1}|X_1, X_2, \ldots, X_{N-2}) \times \cdots P(X_3|X_1, X_2) \times$
$P(X_2|X_1) \times P(X_1)$ (chain rule)

$= P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \times \cdots \times P(X_N|X_1, X_2, \ldots, X_{N-1})$ (which is the right hand side of the solution)

Therefore, $P(X_1, X_2, \ldots, X_N) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \times \cdots \times P(X_N|X_1, X_2, \ldots, X_{N-1})$**.**

2. Assume $P(X, Y) = P(X) \times P(Y)$. Show that $P(Y|X) = P(Y)$.

   Answer:

   From conditional probability of 2 random variables, X and Y, we have:

   $P(Y|X) = \frac{P(X,Y)}{P(X)}$
   or, $P(Y|X) = \frac{P(X) \times P(Y)}{P(X)}$
   or, $P(Y|X) = P(Y)$

   Therefore, $P(Y|X) = P(Y)$ which means that X and Y are independent and have nothing to do with each other (probability of X doesn't affect the probability of Y)

3. Maximum Likelihood Estimation: Assume $X_1, X_2, \ldots, X_N$ are i.i.d. random variables each taking a real value, where

$$p_\theta(X_i = x_i) = e^{-(\theta^4 - \frac{2\theta^2}{x_i} + 3)}.$$

Here, $\theta > 0$ is the parameter of the distribution (assume we know true θ must be positive). Assume, we observe $X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N$.

a) Write down the likelihood function $L(\theta)$ and the log-likelihood function $\ell(\theta)$.

Answer:

The likelihood function is the joint distribution:
$L(\theta) = P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N)$

$= \Pi_{i=1}^N P_\theta(X_i = x_i)$ (for i.i.d. random vars)

$= \Pi_{i=1}^N e^{-(\theta^4 - \frac{2\theta^2}{x_i} + 3)}$

$= e^{\sum_{i=1}^N (-(\theta^4 - \frac{2\theta^2}{x_i} + 3))}$

$= e^{\sum_{i=1}^N (-\theta^4 + \frac{2\theta^2}{x_i} - 3)}$

$= e^{-N\theta^4 + 2\theta^2 \sum_{i=1}^N \frac{1}{x_i} - 3N}$

$\text{Ł}(\theta) = \mathbf{e}^{\mathbf{-N\theta^4 + 2\theta^2 \sum_{i=1}^N \frac{1}{x_i} - 3N}}$

Taking natural log for $L(\theta)$ gives the log-likelihood function:
$\ell(\theta) = \log(L(\theta))$

$= \log(e^{-N\theta^4 + 2\theta^2 \sum_{i=1}^N \frac{1}{x_i} - 3N})$

$= (-N\theta^4 + 2\theta^2 \sum_{i=1}^N \frac{1}{x_i} - 3N) log(e) => (because \log(a^b) = b \log(a))$

$= -N\theta^4 + 2\theta^2 \sum_{i=1}^N \frac{1}{x_i} - 3N =>$(because naturallog$(e) = 1$)

$\ell(\theta) = \mathbf{-N\theta^4 + 2\theta^2 \sum_{i=1}^N \frac{1}{x_i} - 3N}$

b) Derive the maximum log-likelihood estimation of θ for the given observations. Provide all steps of derivations.

Answer:

To find the MLE of θ, we have:

$\theta^*_{MLE} = argmax_\theta \ \ \ell(\theta)$

$= argmax_\theta \ \ (-N\theta^4 + 2\theta^2 \sum_{i=1}^{N} \frac{1}{x_i} - 3N)$

To find the maximum value, we take the derivative of this with respect to θ :

$\frac{\delta(\ell(\theta))}{\delta(\theta)} = \frac{\delta(-N\theta^4 + 2\theta^2 \sum_{i=1}^{N} \frac{1}{x_i} - 3N)}{\delta(\theta)}$

$= -4N\theta^3 + 4\theta \sum_{i=1}^{N} \frac{1}{x_i} => (eq^n 1)$

Now, setting the derivative to 0, we get:

$-4N\theta^3 + 4\theta \sum_{i=1}^{N} \frac{1}{x_i} = 0$

or, $4\theta \sum_{i=1}^{N} \frac{1}{x_i} = 4N\theta^3$

or, $\sum_{i=1}^{N} \frac{1}{x_i} = N\theta^2$

or, $\frac{1}{N} \sum_{i=1}^{N} \frac{1}{x_i} = \theta^2$

or, $\theta^2 = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{x_i}$

or, $\theta^*_{MLE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \frac{1}{x_i}}$

To test, if this in indeed the maximum, we can take a second derivate (i.e. derivate of $eq^n 1$ with respect to θ:

$\frac{\delta^2 \ell(\theta)}{\delta(\theta^2)} = \frac{\delta(-4N\theta^3 + 4\theta \sum_{i=1}^{N} \frac{1}{x_i})}{\delta(\theta^2)}$

$= -12N\theta^2 + 4 \sum_{i=1}^{N} \frac{1}{x_i}$

At $\theta = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \frac{1}{x_i}}$:

$\frac{\delta^2 \ell(\theta)}{\delta(\theta^2)} = -12N \frac{1}{N} \sum_{i=1}^{N} \frac{1}{x_i} + 4 \sum_{i=1}^{N} \frac{1}{x_i}$

$= -12 \sum_{i=1}^{N} \frac{1}{x_i} + 4 \sum_{i=1}^{N} \frac{1}{x_i}$

$= -8 \sum_{i=1}^{N} \frac{1}{x_i}$

Since $\sum_{i=1}^{N} \frac{1}{x_i}$ has to be positive for there to be a real root, $-8 \sum_{i=1}^{N} \frac{1}{x_i}$ is always negative, so the second derivative being negative confirms this to be a local minimum.

Therefore, $\theta^*_{MLE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \frac{1}{x_i}}$ (provided the quantity under ths square root is non-negative).

4. Maximum Likelihood Estimation for Gaussian Random Variables: Consider a dataset $\mathcal{D} = \{X_1 = x_1, \ldots, X_N = x_N\}$ with IID samples from a Gaussian distribution with mean μ and variance σ^2. The Gaussian distribution for each sample is given by

$$p_\theta(X_i = x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}},$$

where the parameter vector θ consists of two entries, $\theta = (\mu, \sigma^2)$. In the class, we formed the log-likelihood function and found the MLE for the mean, μ^*_{MLE}. Derive the maximum likelihood estimation for σ^2. Provide details of your derivations.

Answer:

From class, we derived the following:

- $\text{L}(\theta) = (2\pi\sigma^2)^{-\frac{N}{2}} \times e^{-\sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}}$
- $\ell(\theta) = \log(L(\theta)) = \log((2\pi\sigma^2)^{-\frac{N}{2}} \times e^{-\sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}}) = -\frac{N}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}$
- $\frac{\delta(\ell(\theta))}{\delta(theta)} = \begin{pmatrix} \frac{\delta(\ell)}{\delta(\mu)} \\ \\ \frac{\delta(\ell)}{\delta(\sigma^2)} \end{pmatrix}$ since θ is a 2-dimensional parameter with μ and σ^2.
- $\frac{\delta(\ell)}{\delta(\mu)} \ldots \mu^*_{MLE} = \sum_{i=1}^{N} \frac{x_i}{N}$ (from class notes)

Here, we will differentiate the log-likelihood with respect to σ^2 to derive the MLE for σ^2:

$$\begin{aligned}
\frac{\delta(\ell)}{\delta(\sigma^2)} &= \frac{\delta(-\frac{N}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2})}{\delta(\sigma^2)} \\
&= \frac{\delta(-\frac{N}{2}(\log(2\pi) + \log(\sigma^2)) - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2})}{\delta(\sigma^2)} \Rightarrow (\text{from } log(ab) = loga + logb) \\
&= \frac{\delta(-\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^2) - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2}(\sigma^2)^{-1})}{\delta(\sigma^2)} \\
&= -\frac{N}{2}\frac{1}{\sigma^2} + \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2}(\sigma^2)^{-2} \Rightarrow (\text{from } \frac{\delta(loga)}{\delta(a)} = \frac{1}{a}) \\
&= -\frac{1}{2}(\frac{N}{\sigma^2} - \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{(\sigma^2)^2})
\end{aligned}$$

Setting the derivative to 0, for estimating MLE, we get:

$-\frac{1}{2}(\frac{N}{\sigma^2} - \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{(\sigma^2)^2}) = 0$

$or, \frac{N}{\sigma^2} - \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{(\sigma^2)^2} = 0$

$or, \frac{N\sigma^2 - \sum_{i=1}^{N}(x_i-\mu)^2}{(\sigma^2)^2} = 0$

$or, N\sigma^2 - \sum_{i=1}^{N}(x_i-\mu)^2 = 0$

$or, (\sigma^2)^*_{MLE} = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$

Since this is a gaussian distribution, this is the global maximum (we don't need to take second derivative to verify). So, $(\sigma^2)^*_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$

5. MAP Estimation for Gaussian Random Variables: Consider a dataset $\mathcal{D} = \{X_1 = x_1, \ldots, X_N = x_N\}$ with IID samples from a Gaussian distribution with mean μ and variance 1. The Gaussian distribution for each sample is given by

$$p_\mu(X_i = x_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2}}.$$

Let's assume we believe that μ must be close to 2. We can characterize this by assuming a Gaussian distribution (prior) on μ with mean 2 and variance σ^2. Thus, the prior distribution is given by

$$p(\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\mu-2)^2}{2\sigma^2}}.$$

1. Compute the MAP estimation for μ. Show all steps of your derivations.

   Answer:

   We're given $\sigma^2 = 1$ and prior estimation for that μ is close to 2. With respect to μ:

   The likelihood function is the joint distribution:
   $L_(\mu) = P_\mu(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N)$

   $= \Pi_{i=1}^N P_\mu(X_i = x_i)$ (for i.i.d. random vars)

   $= \Pi_{i=1}^N \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2}}$

   $= \frac{1}{N\sqrt{2\pi}} e^{-\sum_{i=1}^N \frac{(x_i-\mu)^2}{2}}$

   $\ell_{(\mu)} = \log(L_{(\mu)})$

   $= log(\frac{1}{N\sqrt{2\pi}} e^{-\sum_{i=1}^N \frac{(x_i-\mu)^2}{2}})$

   $= log(\frac{1}{N\sqrt{2\pi}}) + \log(e^{-\sum_{i=1}^N \frac{(x_i-\mu)^2}{2}}) => (eq^n 1)$

   $\mu^*_{MAP} = \arg\max_\mu \ \log(P(\mu|D))$

   $= \arg\max_\mu \ log(L_{(\mu)} \times P_{(\mu)})$
   $= \arg\max_\mu \ (log(L_\mu) + log(P_\mu))$
   $= \arg\max_\mu \ (\ell_\mu + log(P_\mu))$

   Substituting the value of ℓ_μ from $eq^n 1$, and the prior distribution from p_μ we get:

   $\mu^*_{MAP} = \arg\max_\mu \ (log(\frac{1}{N\sqrt{2\pi}}) + \log(e^{-\sum_{i=1}^N \frac{(x_i-\mu)^2}{2}}) + log(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\mu-2)^2}{2\sigma^2}}))$

   $= \arg\max_\mu \ (log(\frac{1}{N\sqrt{2\pi}}) + \log(e^{-\sum_{i=1}^N \frac{(x_i-\mu)^2}{2}}) + log(\frac{1}{\sqrt{2\pi\sigma^2}}) + \log(e^{-\frac{(\mu-2)^2}{2\sigma^2}}))$

$$= \arg\max_{\mu} \ (log(\frac{1}{N\sqrt{2\pi}}) - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2} + log(\frac{1}{\sqrt{2\pi\sigma^2}}) - \frac{(\mu-2)^2}{2\sigma^2})$$

To maximize, we take derivative with respect to μ and set it to 0:

$$\frac{\delta(log(\frac{1}{N\sqrt{2\pi}}) - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2} + log(\frac{1}{\sqrt{2\pi\sigma^2}}) - \frac{(\mu-2)^2}{2\sigma^2})}{\delta(\mu)} = 0$$

or, $-\frac{1}{2}\frac{\delta(\sum_{i=1}^{N}(\mu-x_i)^2)}{\delta(\mu)} - \frac{1}{2\sigma^2}\frac{\delta((\mu-2)^2}{\delta(\mu)} = 0$

or, $-\frac{1}{2}(2\sum_{i=1}^{N}(\mu - x_i)) - \frac{1}{2\sigma^2}2(\mu-2) = 0$

or, $-\sum_{i=1}^{N}(\mu - x_i) - \frac{1}{\sigma^2}(\mu-2) = 0$

or, $-\sum_{i=1}^{N}\mu + \sum_{i=1}^{N}x_i - \frac{\mu}{\sigma^2} + \frac{2}{\sigma^2} = 0$

or, $-\sum_{i=1}^{N}\mu - \frac{\mu}{\sigma^2} + \sum_{i=1}^{N}x_i + \frac{2}{\sigma^2} = 0$

or, $-\mu(N + \frac{1}{\sigma^2}) + \sum_{i=1}^{N}x_i + \frac{2}{\sigma^2} = 0$

or, $\mu(N + \frac{1}{\sigma^2}) = \sum_{i=1}^{N}x_i + \frac{2}{\sigma^2}$

Therefore, $\mu^*_{\mathbf{MAP}} = \frac{\sum_{\mathbf{i=1}}^{\mathbf{N}}\mathbf{x_i} + \frac{\mathbf{2}}{\sigma^{\mathbf{2}}}}{(\mathbf{N} + \frac{\mathbf{1}}{\sigma^{\mathbf{2}}})}$

2. Assume $N = 5$, $\sum x_i = 0.5$, $\sigma^2 = 0.01$. What is the Maximum Likelihood Estimate (MLE) for μ? What is the MAP estimate for μ?

$\ell_{(\mu)} = log(\frac{1}{N\sqrt{2\pi}}) + \log(e^{-\sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2}}) =>$ (from $eq^n 1$ in Q1)

$$\mu^*_{MLE} = \arg\max_{\mu} \ \log(\ell_{(\mu)})$$

$$= \arg\max_{\mu} \ (log(\frac{1}{N\sqrt{2\pi}}) + \log(e^{-\sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2}})$$

To maximize, we take derivative with respect to μ and set it to 0:

$\frac{\delta(log(\frac{1}{N\sqrt{2\pi}}) - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2})}{\delta(\mu)} = 0$

or, $-\frac{1}{2}\frac{\delta(\sum_{i=1}^{N}(\mu - x_i)^2)}{\delta(\mu)} = 0$

or, $-\frac{1}{2}(2\sum_{i=1}^{N}(\mu - x_i)) = 0$

or, $-\sum_{i=1}^{N}(\mu - x_i) = 0$

or, $-\sum_{i=1}^{N}\mu + \sum_{i=1}^{N} x_i = 0$

or, $\sum_{i=1}^{N}\mu = \sum_{i=1}^{N} x_i$

or, $\mu^*_{\mathbf{MLE}} = \frac{\sum_{i=1}^{N} x_i}{(N)}$

$= \frac{0.5}{5}$

$= 0.1$

$\mu^*_{\mathbf{MAP}} = \frac{\sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{x_i} + \frac{\mathbf{2}}{\sigma^{\mathbf{2}}}}{(\mathbf{N} + \frac{\mathbf{1}}{\sigma^{\mathbf{2}}})}$

$= \frac{0.5 + \frac{2}{0.01}}{(5 + \frac{1}{0.01})}$

$= \frac{0.5+200}{(5+100)}$

$= \frac{200.5}{105}$

$= 1.9095$

$\mu^*_{\mathbf{MLE}} = \mathbf{0.1}, \quad \mu^*_{\mathbf{MAP}} = \mathbf{1.9095}$

6. Convex Functions and Sets: Show if each of the following statements is true or false. If it is true, prove it by providing all steps of the proof. If it is false, you can show it using a counter example or discussion.

1. Show that the set $S = \{x \in \mathbb{R}^n : Ax \leq c\}$ is convex, where $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$ and $c \in \mathbb{R}^m$ (Hint: you can use the fact that the set $\{x \in \mathbb{R}^n : a^\top x \leq c\}$ is convex).
   Answer:
   A set $S \subseteq \mathbb{R}^n$ is convex if for any $x, y \in S$ and any $\alpha \in [0, 1]$: $\alpha x + (1 - \alpha)y \in S$

   Given, matrix A: $\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}_{m \times n}$ and x: $\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}_{n \times 1}$ and c: $\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix}_{m \times 1}$

   Ax is a vector in $\mathbb{R}^m$ where each component of $(Ax)_i \leq c_i$ for $i = 1, ..., m$. So, the inequality $Ax \leq c$ is component wise:

   $$(Ax)_i = a_i^\top x \leq c_i, \quad i = 1, \dots, m$$

   where $a_i^\top$ is the i-th row of A. So S can be written as the intersection of half-spaces:

   $$S = \bigcap_{i=1}^{m} \{x \in \mathbb{R}^n : a_i^\top x \leq c_i\}$$

   Using the hint, each set $\{x : a_i^\top x \leq c_i\}$ is a half-space, which is convex. The intersection of convex sets is convex. **Therefore, the intersection of m convex half-spaces: $S = \bigcap_{i=1}^{m} \{x : a_i^\top x \leq c_i\}$ is convex.**

2. The function $f(x) = c_1 e^{-\alpha x} + c_2 \log(x)$ is convex. Here, $\alpha, c_1, c_2 \in \mathbb{R}$ with $c_1 \geq 0$, $c_2 \leq 0$ and the domain is $x \in (0, +\infty)$.

   Answer:
   $f(x) = c_1 e^{-\alpha x} + c_2 \log(x)$
   $f^{'}(x) = -\alpha c_1 e^{-\alpha x} + \frac{c_2}{x}$
   $f^{''}(x) = \alpha^2 c_1 e^{-\alpha x} - 1 c_2 x^{-2}$
   $= \frac{\alpha^2 c_1}{e^{\alpha x}} - \frac{c_2}{x^2} => (eq^n 1)$
   We're given: $c_1 \geq 0, c_2 \leq 0, x \in (0, +\infty)$
   $\alpha^2 \geq 0$ (squared value)
   $x^2 \geq 0$ (squared value)
   $e^{\alpha x} \geq 0$ (since base of exponential is positive)

   Since $c_2 \leq 0$, $eq^n 1$, that is the second derivative is always positive. For a twice differentiable function, since $f''(x) \geq 0$ for all $x > 0$, the function is convex on domain $(0, +\infty)$.

3. The function $f(x) = x^3$ is convex.

   Answer:

   $f(x) = x^3$

   $f^{'}(x) = 3x^2$

   $f^{''}(x) = 6x$

   By definition of convex functions, a twice differentiable function, $f(x)$ is convex if $f^{''}(x) \geq 0 \ \forall x$. In this case, if $x = 1$, $f''(x) = -6 \leq 0$ and thus, there exist x values for which $f''(x) < 0$. **Therefore, $f(x)$ is not convex.**

7. Hyperplane Normal: Consider the hyperplane $w^\top x + b = 0$. Show that the normal vector to the hyperplane is w (Hint: show that for any two points in the hyperplane, the line connecting the two points is orthogonal to w).

Answer:

Consider an arbitrary point $z \in \mathbb{R}^n$. We want to project z onto the plane, or find the point x on the hyperplane $w^\top x + b = 0$ closest to z, i.e.,

$$\min_x \|x - z\|^2 \quad \text{s.t.} \quad w^\top x + b = 0.$$

Using the lagrangian function, we get:
$\mathcal{L}(x, \lambda) = \|x - z\|^2 + \lambda(w^\top x + b)$ Now, we take the gradient with respect to x and set it to 0:
$\frac{\delta(L(x,\alpha))}{\delta(x)} = 2(x - z) + \lambda w$
$or, 0 = 2(x - z) + \lambda w$
$or, x - z = -\frac{\lambda}{2} w$
$or, x^* = z - \frac{\lambda}{2} w$

Using the hyperplane constraint and replacing x, we get:
$\frac{\delta(L(x,\alpha))}{\delta(\alpha)} = 0$
$w^\top x + b = 0$
$or, w^\top \left(z - \frac{\lambda}{2} w\right) + b = 0$
$or, \quad \lambda^* = \frac{2(w^\top z + b)}{\|w\|^2}$

$x^* = z - \frac{w^\top z + b}{\|w\|^2} w$
$or, x^* - z = -\frac{w^\top z + b}{\|w\|^2} w$
The displacement vector from z to the hyperplane is parallel to w, so w is orthogonal to all directions lying in the hyperplane. **Hence, w is the normal vector to the hyperplane.**

Optionally, if we were to use distance For the hyperplane $w^\top x + b = 0$, let $x, y \in \mathbb{R}^n$ be 2 points that lie on the hyperplane. Then, by definition of the hyperplane:

$$w^\top x + b = 0 \quad \text{and} \quad w^\top y + b = 0$$

Subtracting the equation 1 by equation 2 to eliminate the bias term, b, we get:
$(w^\top x + b) - (w^\top y + b) = 0 - 0$
$w^\top x + b - w^\top y - b = 0 - 0$
$w^\top (x - y) = 0$
The equation $w^\top (x - y) = 0$ means the 2 vectors, w and (x-y) are perpendicular to each other (a 90 degree). The vector $x - y$ is a direction vector on the hyperplane since it connects the 2 points x and y on the plane. For any 2 points in the hyperplane, w is then perpendicular to each direction vector that connects them. **Therefore, the normal vector to the hyperplane is w.**

8. Logistic Regression Implementation: The goal of this assignment is to give you hands-on experience with **logistic regression (binary)** and **softmax regression (multi-class)**. Please download the provided Iris dataset (120 training samples, 30 test samples). Each sample has two features and a label in $\{0, 1, 2\}$, corresponding to: 0: *Iris setosa*; 1: *Iris versicolor*; 2: *Iris virginica*;

The files `X_train.csv` and `X_test.csv` contain the **features** (two columns), and `y_train.csv` and `y_test.csv` contain the **labels** (one column).

1. **Logistic Regression for Binary Classification:** Train a model to classify setosa vs. non-setosa samples. First, convert all labels of 1 and 2 into 1 (so 0 remains 0; 1 and 2 become 1). Then train a logistic regression model using `LogisticRegression` in scikit-learn on the training set. Report the learnt coefficients. Report classification accuracies on the training set and on the test set. Plot the training samples (different colors for the two classes) and plot the decision boundary.

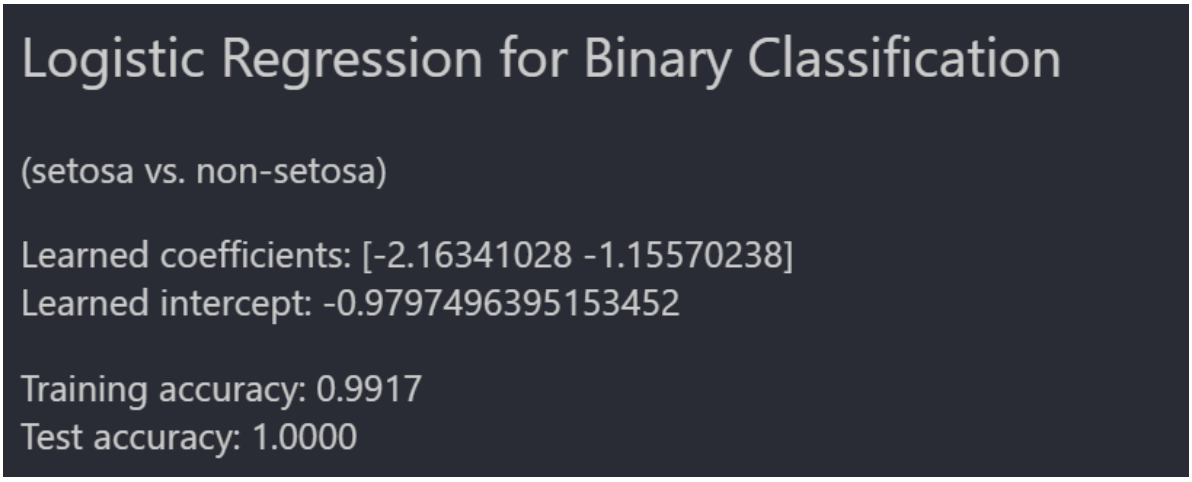

Figure 1: Regression coefficients for binary classification using logistic regression.

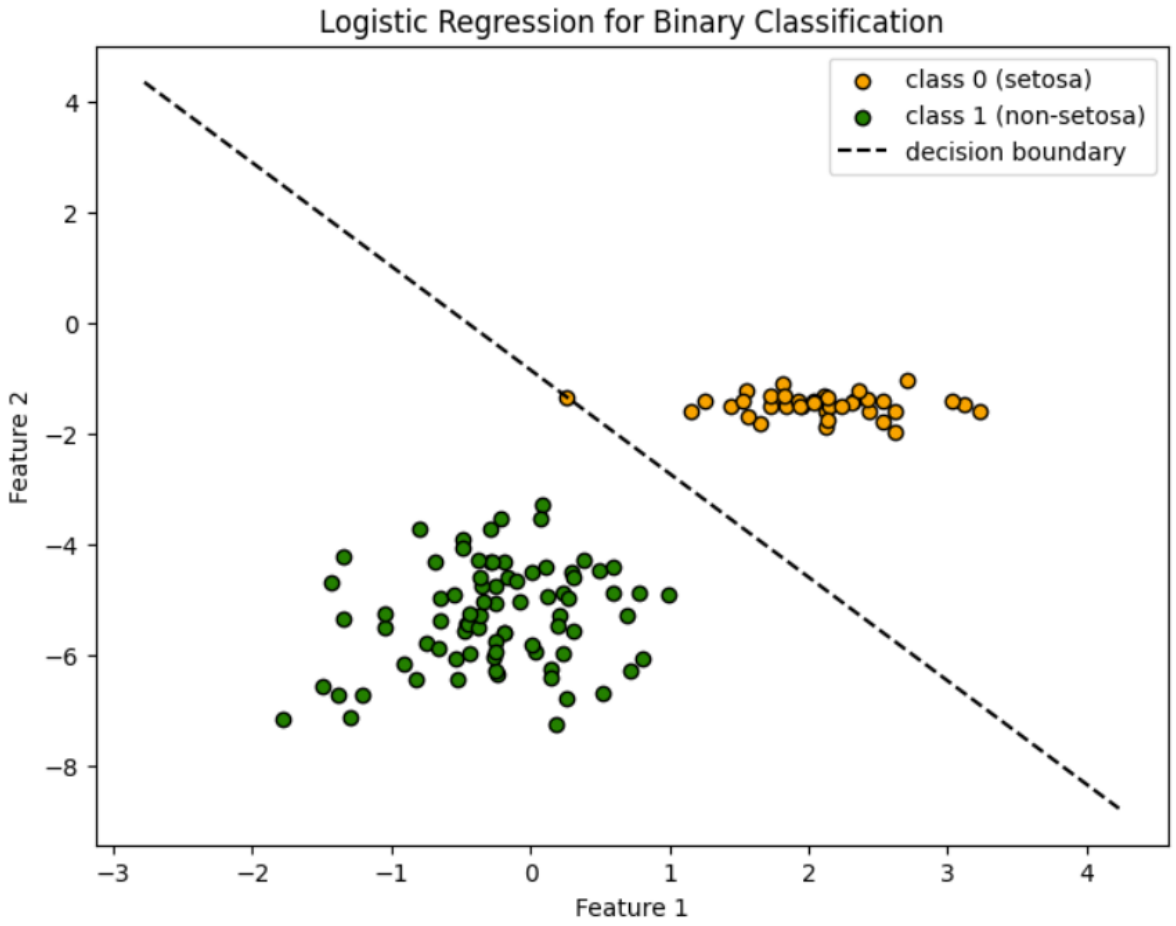

Figure 2: Plotting training samples and decision boundary.

2. **Logistic Regression with Softmax for Multi-class Classification:** Use the original 3-class labels $\{0, 1, 2\}$ and train a logistic regression model with `multi_class='multinomial'` (softmax) on the training set. Report the learnt coefficients. Report classification accuracies on the training set and on the test set. Plot the training samples (three distinct colors for the three classes) and plot the decision boundaries (you will have three boundaries, one for each pair of classes).

```
Logistic Regression with Softmax for Multi-class Classification

(0: Iris setosa; 1: Iris versicolor ; 2: Iris virginica)

Class 0: coefficients = [1.12393584 2.22919459], intercept = 8.245104825434854
Class 1: coefficients = [-0.55527035 0.60345908], intercept = 4.974294036206994
Class 2: coefficients = [-0.56866549 -2.83265367], intercept = -13.219398861641801

Classification Training accuracy: 0.9583
Classification Test accuracy: 1.0000
```

Figure 3: Regression coefficients for multi-class classification using logistic regression (softmax).

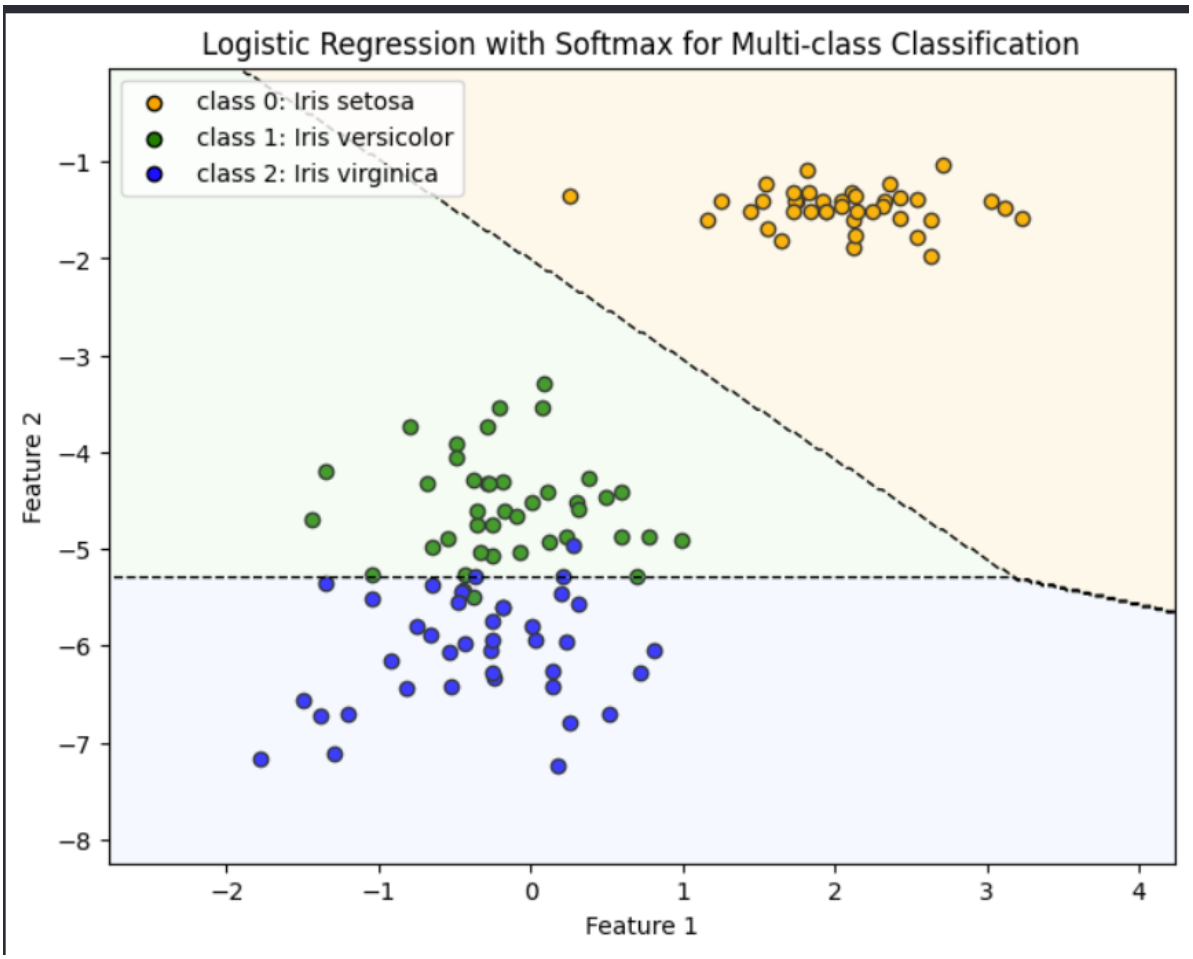

Figure 4: Plotting training samples and decision boundary.