In the name of GOD.



Sharif University of Technology

Machine Learning

Spring 2025

Hamid R. Rabiee, Zahra Dehghanian

Homework 3 — Deadline: 1404/01/19

1. [20] Prove that the logistic regression cost function $J(w)$, defined as follows for the dataset $\{(x_n, y_n)\}_{n=1}^{N}$, is a convex function with respect to the model weights w:

$$J(w) = -\frac{1}{N}\sum_{n=1}^{N}\{y_n \log h_w(x_n) + (1-y_n)\log(1-h_w(x_n))\}$$

where in the above equation, $h_w(x) = \frac{1}{1+\exp(-w^T x)}$.

Hint: To prove that a function is convex, you can compute its Hessian matrix and show that it is positive semi-definite.

Solution:
We can write the loss for a single data point as follows:

$$\begin{aligned}
-L(w) &= y \log \frac{e^{wx}}{1+e^{wx}} + (1-y)\log\frac{1}{1+e^{wx}} \\
&= y\log(e^{wx}) - y\log(1+e^{wx}) + (1-y)\log(1) - (1-y)\log(1+e^{wx}) \\
&= ywx - \log(1+e^{wx})
\end{aligned}$$

Thus, for each data point, we get:

$$L(w) = \log(1+e^{wx}) - ywx$$

Now, we compute the derivative with respect to the weights w:

$$\begin{aligned}
\frac{\partial L}{\partial w} &= \frac{\partial \log(1+e^{wx})}{\partial w} - yx \\
&= \frac{\partial \log(1+e^{wx})}{\partial(1+e^{wx})} \cdot \frac{\partial(1+e^{wx})}{\partial w} - yx \\
&= \frac{e^{wx}x}{1+e^{wx}} - yx \\
&= \frac{x}{1+e^{-wx}} - yx
\end{aligned}$$

Now, we compute the second derivative:

$$\begin{aligned}\frac{\partial^2 L}{\partial w^2} &= \frac{\partial}{\partial w}\left(\frac{x}{1+e^{-wx}} - yx\right) \\ &= \frac{xx^T e^{-wx}}{(1+e^{-wx})^2} \\ &= xx^T C_{wx}\end{aligned}$$

Clearly, the coefficient C_{wx} is always a non-negative number. Also, the matrix xx^T is a positive semi-definite matrix. Therefore, the Hessian matrix of the loss function is positive semi-definite.

It is evident that the sum of multiple positive semi-definite matrices remains positive semi-definite. This fact confirms that summing all the Hessians over the N inputs results in a globally positive semi-definite Hessian matrix.

Thus, the loss function is convex.

2. [15] We consider the following models of logistic regression for a binary classification with a sigmoid function $g(z) = \frac{1}{1+e^{-z}}$:

   - Model 1: $P(Y = 1|X, w_1, w_2) = g(w_1X_1 + w_2X_2)$
   - Model 2: $P(Y = 1|X, w_0, w_1, w_2) = g(w_0 + w_1X_1 + w_2X_2)$

   We have three training examples:

$$x^{(1)} = \begin{bmatrix}1\\1\end{bmatrix}, \quad x^{(2)} = \begin{bmatrix}1\\0\end{bmatrix}, \quad x^{(3)} = \begin{bmatrix}0\\0\end{bmatrix}$$
$$y^{(1)} = 1, \quad y^{(2)} = -1 \quad y^{(3)} = 1$$

   (a) Does it matter how the third example is labeled in Model 1? I.e., would the learned value of $\mathbf{w} = (w_1, w_2)$ be different if we change the label of the third example to -1? Does it matter in Model 2? Briefly explain your answer.

   (b) Now, suppose we train the logistic regression model (Model 2) based on n training examples $x^{(1)}, \ldots, x^{(n)}$ and labels $y^{(1)}, \ldots, y^{(n)}$ by maximizing the penalized log-likelihood of the labels:

$$\sum_i \log P(y^{(i)}|x^{(i)}, \mathbf{w}) - \frac{\lambda}{2}||\mathbf{w}||^2 = \sum_i \log g(y^{(i)}\mathbf{w}^T x^{(i)}) - \frac{\lambda}{2}||\mathbf{w}||^2 \tag{1}$$

   For large λ (strong regularization), the log-likelihood terms will behave as linear functions of $\mathbf{w}$:

$$\log g(y^{(i)}\mathbf{w}^T x^{(i)}) \approx \frac{1}{2} y^{(i)}\mathbf{w}^T x^{(i)} \tag{2}$$

   Express the penalized log-likelihood using this approximation (with Model 1), and derive the expression for MLE $\widehat{\mathbf{w}}$ in terms of λ and training data $\{x^{(i)}, y^{(i)}\}$. Based on this, explain how $\mathbf{w}$ behaves as λ increases. (We assume each $x^{(i)} = (x_1^{(i)}, x_2^{(i)})^T$ and $y^{(i)}$ is either 1 or -1.)

   (c) Now consider the case where the training data is **not linearly separable**. That is, there exists at least one pair $(x^{(i)}, y^{(i)})$ such that the linear decision boundary of logistic regression cannot perfectly classify the dataset. In such a scenario:

      i. Show how the derived MLE solution for $\mathbf{w}$ from part (b) behaves in a non-linearly separable dataset. What does this imply about the learned parameters?

      ii. If we introduce an additional feature $x_3 = x_1^2 + x_2^2$, transforming the data into a higher-dimensional space, how does this affect the interpretability and expressiveness of the logistic regression model?

Solution:

(a) It does not matter in Model 1 because $x^{(3)} = (0, 0)$ makes $w_1x_1 + w_2x_2$ always zero and hence the likelihood of the model does not depend on the value of $\mathbf{w}$. But it does matter in Model 2.

(b)

$$\log L(\mathbf{w}) \approx \sum_i \frac{1}{2} y^{(i)} \mathbf{w}^T x^{(i)} - \frac{\lambda}{2} ||\mathbf{w}||^2$$

$$\frac{\partial}{\partial w_1} \log L(\mathbf{w}) \approx \frac{1}{2} \sum_i y^{(i)} x_1^{(i)} - \lambda w_1 = 0$$

$$\frac{\partial}{\partial w_2} \log L(\mathbf{w}) \approx \frac{1}{2} \sum_i y^{(i)} x_2^{(i)} - \lambda w_2 = 0$$

$$\implies \mathbf{w} = \frac{1}{2\lambda} \sum_i y^{(i)} x^{(i)}$$

(c)

- When the data is not linearly separable, the summation $\sum_i y^{(i)} x^{(i)}$ does not necessarily converge to a meaningful decision boundary. The weight vector $\mathbf{w}$ will be influenced by conflicting samples where positive and negative classes overlap. This means the model will attempt to minimize errors but may still have significant misclassifications depending on the regularization term λ.
- Introducing $x_3 = x_1^2 + x_2^2$ allows the decision boundary to take a quadratic form, making it possible to separate datasets that are not linearly separable in the original space. This transformation increases the expressiveness of the model but also introduces the risk of overfitting if not properly regularized. The interpretability also decreases as the decision boundary is no longer a simple linear hyperplane.

3. [15] The margin of a set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ is defined by:

$$\gamma = \max_{||\mathbf{w}||=1} \min_{(\mathbf{x_i}, y_i)} y_i \mathbf{w}^T \mathbf{x}_i.$$

Prove that if the training data is linearly separable with margin $\gamma > 0$, the Perceptron algorithm will converge after at most $(\frac{R}{\gamma})^2$ mistakes, where R is the radius of the smallest ball containing all training examples (which means that $\max_i ||\mathbf{x}_i|| = R$). Assume the optimal separating hyperplane $\hat{\mathbf{w}}$ satisfies $||\hat{\mathbf{w}}|| = 1$ and $y_i \mathbf{x}_i^T \hat{\mathbf{w}} \geq \gamma$ for all i (by definition).

Solution:
Assume $\hat{\mathbf{w}}$ is the separating hyperplane based on the definintion of margin (in other words, $\gamma = \min_{(\mathbf{x_i}, y_i)} y_i \hat{\mathbf{w}}^T \mathbf{x_i}$). Thus we have:

$$\begin{aligned}
\hat{\mathbf{w}}^T \mathbf{w}^{(k)} &= \hat{\mathbf{w}}^T (\mathbf{w}^{(k-1)} + y_i \mathbf{x}_i) && \text{update rule} \\
&= \hat{\mathbf{w}}^T \mathbf{w}^{(k-1)} + y_i \hat{\mathbf{w}}^T \mathbf{x}_i \\
&\geq \hat{\mathbf{w}}^T \mathbf{w}^{(k-1)} + \gamma && \text{definition of margin} \\
&\geq k\gamma && \text{induction}
\end{aligned}$$

Based on Cauchy-Schwarz inequality we have:

$$\hat{\mathbf{w}}^T \mathbf{w}^{(k)} \leq ||\hat{\mathbf{w}}|| \times ||\mathbf{w}^{(k)}|| = 1 \times ||\mathbf{w}^{(k)}|| \Longrightarrow ||\mathbf{w}^{(k)}|| \geq k\gamma$$

Since each update of $\mathbf{w}^{(k)}$ only occurs when $y_i \neq \text{sgn}(\mathbf{x}_i^T \mathbf{w}^{(k)})$, we can say that

$$\begin{aligned}
||\mathbf{w}^{(k)}||^2 &= ||\mathbf{w}^{(k-1)} + y_i \mathbf{x}_i||^2 && \text{update rule} \\
&= ||\mathbf{w}^{(k-1)}||^2 + |y_i|^2 \times ||\mathbf{x}_i||^2 + 2 y_i \mathbf{x}_i^T \mathbf{w}^{(k-1)} \\
&\leq ||\mathbf{w}^{(k-1)}||^2 + |y_i|^2 \times ||\mathbf{x}_i||^2 && y_i \neq \text{sgn}(\mathbf{x}_i^T \mathbf{w}^{(k)}) \\
&\leq ||\mathbf{w}^{(k-1)}||^2 + 1 \times R^2 && R = \max_i ||\mathbf{x}_i|| \\
&\leq kR^2 && \text{induction}
\end{aligned}$$

Thus we have:

$$k\gamma \leq ||\mathbf{w}^{(k)}|| \leq \sqrt{k} R \Longrightarrow \sqrt{k} \leq \frac{R}{\gamma} \Longrightarrow k \leq (\frac{R}{\gamma})^2$$

4. [20] In a peaceful village, two farmers—Farmer Green and Farmer Brown—each grow apples with unique characteristics. A feature vector $\mathbf{X} \in \mathbb{R}^d$ captures traits such as color, size, and sweetness. Observations suggest that the apples from each farm follow a Gaussian distribution.

   Your goal is to determine the Bayes decision boundary separating Farmer Green's apples (class $k = 1$) from Farmer Brown's apples (class $k = 2$).

   1. Linear Decision Boundary (Shared Covariance)

   Assume the apples from both farms each have a Gaussian class-conditional distribution:

$$p(\mathbf{X} \mid Y = k) = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}),$$

   where both classes share the same covariance matrix:

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}.$$

   (a) Show that the Bayes decision boundary is linear.

   (b) Derive its explicit form, illustrating why the boundary can be represented by a hyperplane in $\mathbb{R}^d$.

   2. Quadratic Decision Boundary (Different Covariances)

   Next, suppose each farm's apples still follow Gaussian distributions, but with different covariance matrices:

$$p(\mathbf{X} \mid Y = k) = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

   where $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$.

   (a) Show that the Bayes optimal decision boundary is quadratic.

   (b) Derive its explicit form and discuss how different covariance structures lead to this more complex boundary.

Solution:

Step 1. Optimal Bayes Decision Rule.

The optimal rule tells us that we assign the label 1 if

$$\mathbb{P}(y = 1 \mid x) > \mathbb{P}(y = 0 \mid x).$$

A common way to express this is to use the sign function:

$$\hat{y} = \operatorname{sign}\left(\mathbb{P}(y = 1 \mid x) - \mathbb{P}(y = 0 \mid x)\right).$$

Using Bayes' rule, we have:

$$\mathbb{P}(y = i \mid x) = \frac{\mathbb{P}(x \mid y = i)\pi_i}{\mathbb{P}(x)} \quad \text{for } i = 0, 1.$$

Thus, the decision criterion becomes:

$$\frac{\mathbb{P}(x \mid y = 1)\pi_1}{\mathbb{P}(x)} > \frac{\mathbb{P}(x \mid y = 0)\pi_0}{\mathbb{P}(x)}.$$

Cancelling the common denominator $\mathbb{P}(x)$ and taking the natural logarithm of both sides gives:

$$\log \mathbb{P}(x \mid y=1) - \log \mathbb{P}(x \mid y=0) + \log \frac{\pi_1}{\pi_0} > 0.$$

Step 2. The Gaussian Likelihoods.

For a d-variate Gaussian,

$$\mathbb{P}(x \mid y=i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right),$$

so that

$$\log \mathbb{P}(x \mid y=i) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i).$$

(a) Linear Discriminant Analysis (LDA): $\Sigma_0 = \Sigma_1 = \Sigma$.

Since the covariance matrices are equal, the logarithms of the densities differ only in the quadratic forms and a constant term related to μ_i. In particular, we have:

$$\log \mathbb{P}(x \mid y=1) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1),$$
$$\log \mathbb{P}(x \mid y=0) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x-\mu_0)^T \Sigma^{-1}(x-\mu_0).$$

Subtracting yields:

$$\begin{aligned}\log \mathbb{P}(x \mid y=1) - \log \mathbb{P}(x \mid y=0) &= -\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^T \Sigma^{-1}(x-\mu_0) \\ &= \frac{1}{2}\Big[(x-\mu_0)^T \Sigma^{-1}(x-\mu_0) - (x-\mu_1)^T \Sigma^{-1}(x-\mu_1)\Big].\end{aligned}$$

Expanding the quadratic forms, notice that:

$$(x-\mu_i)^T \Sigma^{-1}(x-\mu_i) = x^T\Sigma^{-1}x - 2\mu_i^T\Sigma^{-1}x + \mu_i^T\Sigma^{-1}\mu_i, \quad i=0,1.$$

The $x^T\Sigma^{-1}x$ terms cancel and we obtain:

$$\begin{aligned}\log \mathbb{P}(x \mid y=1) - \log \mathbb{P}(x \mid y=0) &= \frac{1}{2}\left[-2\mu_0^T\Sigma^{-1}x + \mu_0^T\Sigma^{-1}\mu_0 + 2\mu_1^T\Sigma^{-1}x - \mu_1^T\Sigma^{-1}\mu_1\right] \\ &= (\mu_1-\mu_0)^T\Sigma^{-1}x - \frac{1}{2}\left(\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0\right).\end{aligned}$$

Thus, including the prior term, the inequality becomes:

$$(\mu_1-\mu_0)^T\Sigma^{-1}x - \frac{1}{2}\left(\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0\right) + \log\frac{\pi_1}{\pi_0} > 0.$$

In other words, the decision function is linear:

$$\hat{y} = \operatorname{sign}\left(\underbrace{(\mu_1-\mu_0)^T\Sigma^{-1}}_{\mathbf{w}^T}x + \underbrace{\left(-\frac{1}{2}\left(\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0\right) + \log\frac{\pi_1}{\pi_0}\right)}_{b}\right).$$

(b) Quadratic Discriminant Analysis (QDA): $\Sigma_0 \neq \Sigma_1$.

When the covariance matrices differ, we write the log-likelihoods as:

$$\log \mathbb{P}(x \mid y=1) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_1| - \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1),$$
$$\log \mathbb{P}(x \mid y=0) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_0| - \frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0).$$

Taking the difference gives:

$$\log \mathbb{P}(x \mid y=1) - \log \mathbb{P}(x \mid y=0) = -\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) - \frac{1}{2}\log\frac{|\Sigma_1|}{|\Sigma_0|}.$$

After including the log-prior ratio, the decision rule becomes:

$$-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) - \frac{1}{2}\log\frac{|\Sigma_1|}{|\Sigma_0|} + \log\frac{\pi_1}{\pi_0} > 0.$$

Let us expand the quadratic terms:

$$(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) = x^T\Sigma_i^{-1}x - 2\mu_i^T\Sigma_i^{-1}x + \mu_i^T\Sigma_i^{-1}\mu_i, \quad i=0,1.$$

Thus, the difference becomes:

$$\begin{aligned}
&-\frac{1}{2}\left[x^T\Sigma_1^{-1}x - 2\mu_1^T\Sigma_1^{-1}x + \mu_1^T\Sigma_1^{-1}\mu_1\right] \\
&\quad + \frac{1}{2}\left[x^T\Sigma_0^{-1}x - 2\mu_0^T\Sigma_0^{-1}x + \mu_0^T\Sigma_0^{-1}\mu_0\right] - \frac{1}{2}\log\frac{|\Sigma_1|}{|\Sigma_0|} + \log\frac{\pi_1}{\pi_0} \\
&= \frac{1}{2}\,x^T\left(\Sigma_0^{-1} - \Sigma_1^{-1}\right)x \\
&\quad + \left(\mu_1^T\Sigma_1^{-1} - \mu_0^T\Sigma_0^{-1}\right)x \\
&\quad + \frac{1}{2}\left(\mu_0^T\Sigma_0^{-1}\mu_0 - \mu_1^T\Sigma_1^{-1}\mu_1\right) - \frac{1}{2}\log\frac{|\Sigma_1|}{|\Sigma_0|} + \log\frac{\pi_1}{\pi_0}\,.
\end{aligned}$$

This quadratic expression in x defines a *quadratic* decision boundary. We can write the decision function in the form:

$$\hat{y} = \operatorname{sign}\Big(x^T\mathbf{W}x + x^T\mathbf{w} + b\Big),$$

with:

$$\mathbf{W} = \frac{1}{2}\left(\Sigma_0^{-1} - \Sigma_1^{-1}\right),$$
$$\mathbf{w} = \mu_1^T\Sigma_1^{-1} - \mu_0^T\Sigma_0^{-1},$$
$$b = \frac{1}{2}\left(\mu_0^T\Sigma_0^{-1}\mu_0 - \mu_1^T\Sigma_1^{-1}\mu_1\right) - \frac{1}{2}\log\frac{|\Sigma_1|}{|\Sigma_0|} + \log\frac{\pi_1}{\pi_0}\,.$$

5. [15 + 15] Complete the attached notebooks. You are permitted to use chatbots or other resources for assistance, but you must ensure that you fully understand the code and implementations. During the online sessions, you may be asked to explain specific functions, code lines, and your overall approach. Be prepared to demonstrate your understanding in detail.