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.*

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: $P(Y=1|X, w_1, w_2) = g(w_1 X_1 + w_2 X_2)$
- Model 2: $P(Y=1|X, w_0, w_1, w_2) = g(w_0 + w_1 X_1 + w_2 X_2)$

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 $\lambda$ (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 $\lambda$ and training data $\{x^{(i)}, y^{(i)}\}$. Based on this, explain

how $\mathbf{w}$ behaves as $\lambda$ 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?

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).

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.

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.