Homework 3

Deadline: November 28

Exercise 1 (20%)

Suppose we collect data from a group of students in a Machine learning class with variables x_1 = hours studied, x_2 = grade point average, and $y = a$ binary output if that student received grade 5 ($y = 1$) or not ($y = 0$). We learn a logistic regression model

$$p(y = 1 \mid \boldsymbol{x}) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}$$

with parameters $\hat{\beta}_0 = -6, \hat{\beta}_1 = 0.05, \hat{\beta}_2 = 1$.

Q1 Estimate the probability according to the logistic regression model that a student who studies for 40 h and has the grade point average of 3.5 gets a 5 in the Machine learning class.

Exercise 2 (20%)

Q1 Let $\sigma(a) = \frac{1}{1+e^{-a}}$ be the sigmoid function. Show that

$$\frac{d\sigma(a)}{da} = \sigma(a)(1 - \sigma(a))$$

Q2 Using the previous result and the chain rule of calculus, derive an expression for the gradient of the log likelihood for logistic regression.

Exercise 3 (20%)

In this assignment, you will explore generative classification using Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). You will derive theoretical results, implement these models in Python, and analyze their performance.

Q1 Maximum Likelihood for Class Probabilities

Show that the maximum likelihood estimate for the class probabilities $\boldsymbol{\pi}$ is given by:

$$\pi_c = \frac{N_c}{N},$$

where N_c is the number of data points assigned to class c, and N is the total number of data points.

Q2 Class-Conditional Densities with Shared Covariance Matrix (LDA)
Assuming the class-conditional densities follow a Gaussian distribution with a shared covariance matrix:

$$p(\boldsymbol{x} \mid y = c, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_c, \boldsymbol{\Sigma}),$$

derive the following maximum likelihood estimates:

a) The mean for each class c:

$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{\substack{n=1 \\ y^{(n)}=c}}^{N} \boldsymbol{x}^{(n)}.$$

b) The shared covariance matrix:

$$\boldsymbol{\Sigma} = \sum_{c=1}^{C} \frac{N_c}{N} \boldsymbol{S}_c, \quad \text{where} \quad \boldsymbol{S}_c = \frac{1}{N_c} \sum_{\substack{n=1 \\ y^{(n)}=c}}^{N} (\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)^T.$$

Q3 Class-Conditional Densities with Separate Covariance Matrices (QDA)
Assuming the class-conditional densities follow Gaussian distributions with separate covariance matrices:

$$p(\boldsymbol{x} \mid y = c, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c),$$

derive the maximum likelihood estimate for the class-specific covariance matrix:

$$\boldsymbol{\Sigma}_c = \frac{1}{N_c} \sum_{\substack{n=1 \\ y^{(n)}=c}}^{N} (\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)^T.$$

Q4 Model Implementation
Implement LDA and QDA fit function to learn $\mu_0\,, \mu_1\,, \Sigma_0\,, \Sigma_1\,, \pi, w_0\,, w_1$

Q5 Performance Comparison
Evaluate the performance of LDA and QDA by calculating accuracy on the test set. Discuss the differences between the decision boundaries of LDA and QDA and their behavior under different class distributions.

Exercise 4 (20%)

This assignment focuses on implementing and analyzing logistic regression for binary classification. You will derive key mathematical components (negative log-likelihood, gradient, Hessian) and use the Newton-Raphson algorithm for optimization. The practical tasks involve preprocessing data, implementing logistic regression, and evaluating it on the Breast Cancer Survival dataset.

Q1 Negative Log-Likelihood Derivation Derive the negative log-likelihood (NLL) for the logistic regression model.

Q2 Gradient and Hessian Analysis

a) Derive the gradient of the NLL with respect to the parameter vector β.

b) Compute the Hessian matrix for the logistic regression model.

c) How we can use the Hessian for optimization using Newton-Raphson algorithm.

Q3 Data Exploration and Preprocessing

a) Load the Breast Cancer Survival dataset and explore its characteristics (e.g., summary statistics, missing values, distribution of classes).

The dataset contains cases from a study conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer. The dataset has the following attributes:

- **Age:** Age of patient at the time of operation (numerical).
- **Year:** Patient's year of operation (year - 1900, numerical).
- **Positive Axillary Nodes:** Number of positive axillary nodes detected (numerical).
- **Survival Status:**
  - 1 = The patient survived 5 years or longer.
  - 2 = The patient died within 5 years.

See more info at the UCI Repository. The dataset can be downloaded from this link.

b) Normalize or standardize the numerical features to prepare them for logistic regression.

Q4 Logistic Regression Implementation

a) Implement the logistic function $\sigma(x)$ in Python.

b) Write a function to compute the NLL for a given dataset and parameter vector β.

c) Implement gradient computation and verify it numerically (e.g., using finite differences).

Q5 Newton-Raphson Algorithm

a) Write a Python function to optimize the NLL using the Newton-Raphson method.

b) Apply this function to the Breast Cancer Survival dataset to find the MLE estimates for β.

c) **Hint:**

- In the Newton-Raphson method, update β iteratively as:

$$\beta^{(t+1)} = \beta^{(t)} - H^{-1}\nabla\text{NLL}$$

  where H is the Hessian matrix and ∇NLL is the gradient of the negative log-likelihood.

- Use numerical stability techniques like adding a small constant ϵ to the diagonal of H if it is singular or poorly conditioned.

Q6 Model Evaluation

a) Compute classification accuracy, precision, recall, and F1-score for the logistic regression model. Use a train-test split or cross-validation for evaluation.

b) Plot the ROC curve and compute the AUC for the model.

Q7 Feature Importance After fitting the model, analyze the learned coefficients β. Discuss which features contribute most to predicting survival status and why.

Q8 Comparison with Library Implementation

a) Use a standard library (e.g., `scikit-learn`) to fit a logistic regression model to the dataset.

b) Compare the coefficients, accuracy, and other metrics with your implementation.

Extra Credit (Optional)

- Implement L_2-regularized logistic regression using the Newton-Raphson method. Compare the results with the unregularized version.
- Visualize the decision boundary for logistic regression (if feasible in a reduced feature space). Use plots to show how the model classifies different regions of the feature space.

Exercise 5 (20%)

This assignment explores Bayesian logistic regression using the Laplace approximation. You will derive the posterior distribution, compute the maximum a posteriori (MAP) estimate, and approximate the posterior as a Gaussian using the Hessian of the negative log-posterior. The programming tasks involve implementing the gradient and Hessian, extending them to compute evaluation metrics on the Breast Cancer Survival dataset.

Q1 Derive the Laplace approximation for the posterior distribution in a Bayesian logistic regression setting. Assume a binary classification problem where the likelihood is given by the logistic regression model:

$$p(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = \sigma(\mathbf{x}_i^\top \boldsymbol{\beta})^{y_i} \left[1 - \sigma(\mathbf{x}_i^\top \boldsymbol{\beta})\right]^{1-y_i},$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$. The prior on $\boldsymbol{\beta}$ is Gaussian: $p(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta}|\mathbf{0}, \mathbf{I})$. Derive the expression for the mode of the posterior (i.e., maximum a posteriori, or MAP estimate), and explain how the Hessian matrix of the negative log-posterior is used to approximate the posterior as a Gaussian.

Q2 Implement the Laplace approximation for Bayesian logistic regression. Given the Breast Cancer Survival dataset, write a Python program to:

Q1 implement Gradient and Hessian of negative log-posterior.

Q2 overwrite the gradient and hessian function in the previous exercise, to compute classification accuracy, precision, recall, and F1-score for the logistic regression model implemented in Exercise 4.