Summer 2025 MFE 230P Problem Set 1

Problem 1 Credit Default Prediction and Statistical Learning

The bank must balance these tradeoffs while considering the asymmetric costs of different types of errors in financial decision-making.

Part A Model Setup and True Data Generating Process (3 pts)

Consider a portfolio of N corporate bonds. For each bond i, let $Y_i \in \{0, 1\}$ indicate default status (1 = default, 0 = no default). The true default probability depends on a risk score X_i:

$$P(Y_i = 1|X_i = x) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

where:

- $\Phi(\cdot)$ is the standard normal CDF (probit model)
- $X_i \sim N(\mu_X, \sigma_X^2)$ are observable risk scores
- $\mu = 2.0$ and $\sigma = 0.5$ are unknown parameters to be estimated
- $\mu_X = 1.5$ and $\sigma_X = 1.0$ characterize the risk score distribution

Task 1: Calculate the marginal default probability $P(Y_i = 1)$ by integrating over the distribution of X_i.
Task 2: The bank will classify a bond as "high risk" (predict default) if $\hat{p}_i > \tau$ where $\hat{p}_i$ is the estimated default probability and τ is a threshold. Express the theoretical classification accuracy as a function of τ, μ, and σ.
Task 3: Define the confusion matrix in terms of the model parameters:

Part B Classification Metrics and Financial Interpretation (4 pts)

The bank estimates the default probability using a simple logistic regression: $\hat{p}_i = \frac{1}{1+e^{-(\hat{\alpha}+\hat{\beta}X_i)}}$.
Due to limited data, the parameter estimates have known distributions:

$$\hat{\alpha} \sim N(\alpha_0, \sigma_\alpha^2), \tag{1}$$
$$\hat{\beta} \sim N(\beta_0, \sigma_\beta^2), \tag{2}$$

where α_0, β_0, σ_α^2, σ_β^2 are known constants.
Task 1: For a given threshold τ, derive analytical expressions for:
a) Precision: Precision $= \frac{\text{TP}}{\text{TP+FP}}$
b) Recall: Recall $= \frac{\text{TP}}{\text{TP+FN}}$
c) F1 Score: F1 $= \frac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision+Recall}}$
Express these in terms of the underlying parameters and threshold τ.
Task 2: The bank faces asymmetric costs:

- Cost of false positive (Type I): $C_1 = \$50,000$ per bond (lost lending profit)

- Cost of false negative (Type II): $C_2 = \$500,000$ per bond (credit loss)

Derive the expected total cost per bond as a function of the threshold τ. Find the optimal threshold τ^* that minimizes expected cost.

Problem 2 Simple Linear Regression via Maximum Likelihood

Part A Derivation of Log-Likelihood Function (1 pts)

Suppose we observe n independent and identically distributed data points $(x_i, y_i)_{i=1}^n$, where $x_i \in \mathbb{R}$ are fixed design points and $y_i \in \mathbb{R}$ are random responses. Assume the data follows the linear regression model:

$$y_i = \beta_0 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

where $\varepsilon_i \sim N(0, \sigma^2)$ are independent error terms, and $\beta_0 \in \mathbb{R}$ is the unknown parameter of interest.
Task: Derive the log-likelihood function $\ell(\beta)$ for the parameter $\beta \in \mathbb{R}$ given the observed data $\{(x_i, y_i)\}_{i=1}^n$.

Part B Maximum Likelihood Estimation (1 pts)

Task: Find the maximum likelihood estimator (MLE) $\hat{\beta}_{MLE}$ of the true parameter β_0. The MLE is defined as:

$$\hat{\beta}_{MLE} = \arg\max_{\beta \in \mathbb{R}} \ell(\beta).$$

Compare your result with the ordinary least squares (OLS) estimator obtained by minimizing the mean squared error:

$$\hat{\beta}_{OLS} = \arg\min_{\beta \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \beta x_i)^2.$$

Discussion: Explain why the two estimators are identical in this case.

Part C Variance of the MLE (2 pts)

Task: Derive an expression for the variance of the maximum likelihood estimator $\hat{\beta}_{MLE}$.
Hint: (You are not required to use the hint) You may use the fact that for a correctly specified model, the asymptotic variance of the MLE is given by the inverse of the Fisher information matrix.
Fisher Information Matrix - Definition and Properties:

For a parameter vector $\theta \in \mathbb{R}^d$ and under regularity conditions, the Fisher information matrix $I(\theta)$ is defined as:

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^T}\right] = \mathbb{E}\left[\left(\frac{\partial \ell(\theta)}{\partial\theta}\right)\left(\frac{\partial \ell(\theta)}{\partial\theta}\right)^T\right],$$

where $\ell(\theta)$ is the log-likelihood function.

Problem 3 Multiple Linear Regression via Maximum Likelihood

Part A Multivariate Log-Likelihood Function (1 pts)

Now consider the multiple linear regression setting. Suppose we observe n independent and identically distributed data points $(x_i, y_i)_{i=1}^n$, where $x_i \in \mathbb{R}^p$ are fixed design vectors and $y_i \in \mathbb{R}$ are random responses. Assume the data follows the linear regression model:

$$y_i = x_i^T \beta_0 + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

where $\varepsilon_i \sim N(0, \sigma^2)$ are independent error terms, and $\beta_0 \in \mathbb{R}^p$ is the unknown parameter vector.

Task: Derive the log-likelihood function $\ell(\beta)$ for the parameter vector $\beta \in \mathbb{R}^p$ given the observed data $\{(x_i, y_i)\}_{i=1}^n$. Express your answer in both summation form and matrix form using the design matrix $X \in \mathbb{R}^{n \times p}$ with rows x_i^T and response vector $y \in \mathbb{R}^n$ with entries y_i.

Part B Multivariate Maximum Likelihood Estimation (1 pts)

Task: Find the maximum likelihood estimator $\hat{\beta}_{MLE} \in \mathbb{R}^p$ of the true parameter β_0. The MLE is defined as:

$$\hat{\beta}_{MLE} = \arg\max_{\beta \in \mathbb{R}^p} \ell(\beta).$$

Compare your result with the closed-form solution of the OLS estimator derived from minimizing the mean squared error:

$$\hat{\beta}_{OLS} = \arg\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2.$$

Problem 4 Total Least Squares and Errors-in-Variables Models

Problem Context and Motivation:

In ordinary least squares (OLS), we assume that the independent variable x is measured without error, and only the dependent variable y contains random noise. However, in many real-world applications, both variables are subject to measurement errors. Examples include:

- Measuring the relationship between height and weight when both measurements have instrument errors
- Economic data where both GDP and unemployment rates have reporting errors
- Scientific experiments where both input and output variables have measurement uncertainty

Total Least Squares (TLS) addresses this by allowing errors in both variables, leading to a fundamentally different estimation problem.

Part A Model Setup and Problem Definition (0 pt)

The True Relationship: Suppose there exist unobserved "true" values $(\xi_i, \eta_i)_{i=1}^n$ that follow a perfect linear relationship:

$$\eta_i = \beta_0 \xi_i, \quad i = 1, 2, \ldots, n$$

where $\beta_0 \in \mathbb{R}$ is the unknown slope parameter (note: we assume no intercept for simplicity).

The Observation Model: However, we don't observe (ξ_i, η_i) directly. Instead, we observe (x_i, y_i) where:

$$x_i = \xi_i + \varepsilon_i \tag{3}$$
$$y_i = \eta_i + \delta_i \tag{4}$$

for $i = 1, 2, \ldots, n$, where:

- $\varepsilon_i \sim N(0, \sigma_x^2)$ represents measurement error in the x-variable
- $\delta_i \sim N(0, \sigma_y^2)$ represents measurement error in the y-variable
- All error terms $\{\varepsilon_i, \delta_i\}_{i=1}^n$ are mutually independent
- The true values ξ_i are unknown parameters to be estimated alongside β_0

Part B Likelihood Function Derivation (2 pts)

Task: Derive the joint likelihood function for the observed data $(x_i, y_i)_{i=1}^n$ given the parameters β_0, σ_x^2, σ_y^2, and the latent variables ξ_i.

Step-by-Step Guidance:

Step 1: Write down the joint probability density function for a single observation (x_i, y_i) given ξ_i, β_0, σ_x^2, and σ_y^2.

Hint: Use the fact that $x_i|\xi_i \sim N(\xi_i, \sigma_x^2)$ and $y_i|\xi_i \sim N(\beta_0\xi_i, \sigma_y^2)$, and that the errors are independent.

Step 2: Write the joint likelihood for all n observations:

$$L(\beta_0, \sigma_x^2, \sigma_y^2, \xi_1, \ldots, \xi_n) = \prod_{i=1}^{n} f(x_i, y_i|\xi_i, \beta_0, \sigma_x^2, \sigma_y^2)$$

Step 3: Take the logarithm to obtain the log-likelihood function:

$$\ell(\beta_0, \sigma_x^2, \sigma_y^2, \xi_1, \ldots, \xi_n) = \log L(\beta_0, \sigma_x^2, \sigma_y^2, \xi_1, \ldots, \xi_n)$$

Express your final answer in a simplified form, clearly showing the dependence on all parameters.

Part C Maximum Likelihood Estimation (2 pts)

Task: Find the maximum likelihood estimators for β_0 and $\{\xi_i\}_{i=1}^n$. Assume that σ_x^2 and σ_y^2 are known constants.

For fixed β_0, find the MLE of each ξ_i by maximizing the log-likelihood with respect to ξ_i. Show that:

$$\hat{\xi}_i(\beta_0) = \frac{\sigma_y^2 x_i + \beta_0\sigma_x^2 y_i}{\sigma_y^2 + \beta_0^2\sigma_x^2}$$

Explain why this estimator represents a weighted average of information from both x_i and y_i.

Substitute $\hat{\xi}_i(\beta_0)$ back into the log-likelihood to obtain a concentrated log-likelihood function $\ell_c(\beta_0)$ that depends only on β_0. Then find the MLE $\hat{\beta}_0$ by solving:

$$\frac{d\ell_c(\beta_0)}{d\beta_0} = 0$$

Part D Comparison with OLS and Geometric Interpretation (1 pts)

Task: Explain the geometric interpretation: while OLS minimizes vertical distances from points to the line, what does TLS minimize?

Problem 5 Bootstrap Bias Problem: Discontinuous CDF and Quantile Estimation

Problem Context and Motivation:

The bootstrap is a powerful resampling method for statistical inference, particularly useful when the theoretical distribution of a statistic is unknown or complex. However, when the underlying distribution has a discontinuous cumulative distribution function (CDF), full-sample bootstrapping can exhibit bias that does not vanish as the sample size increases. This problem explores this phenomenon.

Part A Model Setup and Distribution Definition (3 pts)

Consider a random variable X with the following cumulative distribution function:

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \alpha & \text{if } 0 \le x < c \\ \alpha + (1-\alpha) \cdot \frac{x-c}{d-c} & \text{if } c \le x \le d \\ 1 & \text{if } x > d, \end{cases}$$

where $0 < \alpha < 1$, $0 < c < d$, and all parameters are known constants.

Task 1: Describe this distribution in words. What are the components of this mixture distribution?

Task 2: Find the probability density function (PDF) $f(x)$ corresponding to this CDF. Be careful to account for the point mass.

Task 3: Calculate the true p-quantile ξ_p for this distribution, where $0 < p < 1$. Consider the cases where $p < \alpha$ and $p \ge \alpha$ separately.

Part B Sample Generation and Empirical Distribution (2 pts)

Suppose we observe an i.i.d. sample $X_1, X_2, \dots, X_n$ from the distribution $F(x)$.

Task: Let $N_0 = \sum_{i=1}^n \mathbf{1}_{\{X_i = 0\}}$ be the number of observations equal to zero, and let $Y_1, Y_2, \dots, Y_{n-N_0}$ be the observations that fall in the interval $[c, d]$.

Show that:

- $N_0 \sim \text{Binomial}(n, \alpha)$
- $Y_j \sim \text{Uniform}(c, d)$ for $j = 1, \dots, n - N_0$

Part C Full-Sample Bootstrap Bias Analysis (3 pts)

Consider the estimation of the p-quantile ξ_p where $p \ge \alpha$.

Task 1: For the full-sample bootstrap, we resample n observations with replacement from $\{X_1, \dots, X_n\}$ to create a bootstrap sample $\{X_1^*, \dots, X_n^*\}$.

Let $\hat{\xi}_p^*$ be the p-quantile of the bootstrap sample. Show that the bootstrap estimate can be written as:

$$\hat{\xi}_p^* = \begin{cases} 0 & \text{if } p \leq \frac{N_0}{n} \\ c + (d-c) \cdot \frac{np - N_0}{n - N_0} & \text{if } \frac{N_0}{n} < p \leq 1 \end{cases}$$

Task 2: Calculate the expected value of the bootstrap quantile estimator:

$$E[\hat{\xi}_p^*] = E[E[\hat{\xi}_p^* | N_0]]$$

Hint: Use the law of total expectation, conditioning on N_0.

Task 3: Compare $E[\hat{\xi}_p^*]$ with the true quantile ξ_p. Show that:

$$\lim_{n \to \infty} E[\hat{\xi}_p^*] \neq \xi_p$$

This demonstrates that the full-sample bootstrap is **asymptotically biased** for discontinuous CDFs.

Task 4: Explain intuitively why this bias occurs. What is the fundamental issue with resampling from the empirical distribution when the true distribution has a discontinuous CDF?