PROJECT II: LOGISTIC REGRESSION
ECE565: Estimation, Filtering, and Detection

Nam Nguyen

December 12, 2023

Oregon State University | Electrical Engineering and Computer Science

Motivation
Classificaiton problem

- Give input $\mathbf{x}$, predict the label y.
  - $\mathbf{x}$: feature vector
  - y: class label
- Example:
  - $\mathbf{x}$: monthly income and bank saving account

    y: whether a person will buy a house or not (binary class)
  - Review text for a product

    y: sentiment positive, negative, or neutral (multi-class)

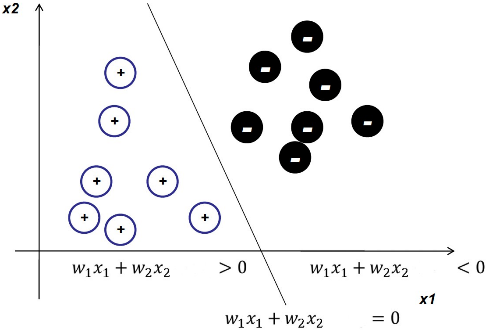


Figure: Binary linear classifier [1].

- w_1 and w_2 are the parameters of the linear classifier.
- The decision boundary: $w_1x_1 + w_2x_2 = 0$.

- Logistic regression is a linear classification model that is widely used in machine learning and statistics.
- Dataset: $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ are assumed to be independent and identically distributed.
- The observation vector is $\mathbf{x}_i = \begin{bmatrix} x_1 & \dots & x_d \end{bmatrix}^T \in \mathbb{R}^d$. $y_i \in \{0, 1\}$ is its label.
- The parameter vector is $\mathbf{w} = \begin{bmatrix} w_1 & \dots & w_d \end{bmatrix}^T \in \mathbb{R}^d$.

- The logistic function models the conditional probability:

$$p(y|\mathbf{x};\mathbf{w}) = \frac{e^{(\mathbf{w}^T\mathbf{x})y}}{1+e^{\mathbf{w}^T\mathbf{x}}}, \tag{1}$$

- The joint probability model of the observations:

$$p\left(\mathbf{X},\mathbf{y}|\mathbf{w}\right) = \prod_{i=1}^{n} \frac{e^{(\mathbf{w}^T\mathbf{x}_i)y_i}}{1+e^{\mathbf{w}^T\mathbf{x}_i}}, \tag{2}$$

where $\mathbf{X} = [\mathbf{x}_1^T, \mathbf{x}_2^T, ..., \mathbf{x}_n^T]$ and $\mathbf{y} = [y_1, y_2, ..., y_n]^T$ [2].

- The log-likelihood function:

$$l(\mathbf{w}) = \sum_{i=1}^{n} \left[y_i(\mathbf{w}^T\mathbf{x}_i) - \log(1+e^{\mathbf{w}^T\mathbf{x}_i}) \right], \tag{3}$$

CRLB Analysis

- Fisher information matrix:

$$\text{FIM} = -\mathbb{E}\left[\frac{d^2 l(\mathbf{w})}{d\mathbf{w}\mathbf{w}^T}\right] = n\mathbb{E}\left[\frac{e^{\mathbf{w}^T\mathbf{x}}}{(1+e^{\mathbf{w}^T\mathbf{x}})^2}\mathbf{x}\mathbf{x}^T\right], \quad (4)$$
$$= n\sigma^2\left[(\alpha_2-\alpha_0)\mathbf{u}_1\mathbf{u}_1^T + \alpha_0\mathbf{I}\right], \quad \mathbf{x} \sim N(\mathbf{0}, \sigma^2\mathbf{I}).$$

where $\mathbf{u}_1 = \frac{\mathbf{w}}{\|\mathbf{w}\|}$, $\alpha_k = \mathbb{E}\left[\frac{e^{\sigma\|\mathbf{w}\|z}}{(1+e^{\sigma\|\mathbf{w}\|z})^2}z^k\right]$ and $z \sim \mathcal{N}(0,1)$.

- CRLB:

$$\text{CRLB} = \text{FIM}^{-1} = \frac{1}{n\sigma^2\alpha_0}\left[\mathbf{I} - \frac{\alpha_2-\alpha_0}{\alpha_2}\mathbf{u}_1\mathbf{u}_1^T\right], \quad (5)$$

- MSE:

$$\mathbb{E}(\|\hat{\mathbf{w}} - \mathbf{w}\|^2) \geq \text{tr}(\text{CRLB}) = \sum_{i=1}^{d}\text{CRLB}_{ii}, \quad (6)$$

Likelihood of ML Estimation

- The log-likelihood function:

$$l(\mathbf{w}) = \sum_{i=1}^{n} \left(y_i(\mathbf{w}^T\mathbf{x}_i) - \log(1 + e^{\mathbf{w}^T\mathbf{x}_i}) \right), \tag{7}$$

- Optimization problem:

$$\hat{\mathbf{w}}_{MLE} = \arg\max_{\mathbf{w}} \sum_{i=1}^{n} \left(y_i(\mathbf{w}^T\mathbf{x}_i) - \log(1 + e^{\mathbf{w}^T\mathbf{x}_i}) \right), \tag{8}$$

- The gradient of the log-likelihood function:

$$\nabla_{\mathbf{w}} l(\mathbf{w}) = \sum_{i=1}^{n} \left(y_i - \frac{e^{\mathbf{w}^T\mathbf{x}_i}}{1 + e^{\mathbf{w}^T\mathbf{x}_i}} \right) \mathbf{x}_i, \tag{9}$$

Iterative Scaling Approach

Version 1

- Iterative scaling iterations [3]:

$$\hat{w}_k^{new} = \hat{w}_k^{old} + \frac{1}{2s} \log \frac{\sum_{i|y_i x_{ik}>0} (1 - \sigma(y_i \left(\hat{\mathbf{w}}^{old}\right)^T \mathbf{x}_i))|x_{ik}|}{\sum_{i|y_i x_{ik}<0} (1 - \sigma(y_i \left(\hat{\mathbf{w}}^{old}\right)^T \mathbf{x}_i))|x_{ik}|}, \tag{10}$$

where $s = \max_i \sum_k |x_{ik}|$ and $\sigma(z) = \frac{1}{1+e^{-z}}$.

Algorithm 1 Iterative Scaling Approach Version 1.

1: **Initialization:** Set $t = 0$, $\hat{\mathbf{w}}^{(0)}$, $\mathbf{X}$, $\mathbf{y}$
2: **Repeat**
3: $s = \max_i \sum_k |x_{ik}|$;
4: $\hat{w}_k^{(t+1)} = \hat{w}_k^{(t)} + \frac{1}{2s} \log \frac{\sum_{i|y_i x_{ik}>0} (1-\sigma(y_i \left(\hat{\mathbf{w}}^{(t)}\right)^T \mathbf{x}_i))|x_{ik}|}{\sum_{i|y_i x_{ik}<0} (1-\sigma(y_i \left(\hat{\mathbf{w}}^{(t)}\right)^T \mathbf{x}_i))|x_{ik}|}$;
5: $t = t + 1$;
6: **Until** convergence

Iterative Scaling Approach
Version 2

- Iterative scaling iterations [4]:

$$\hat{w}_k^{new} = \hat{w}_k^{old} + \frac{1}{S}\log\left(\frac{B_k + \sqrt{B_k^2 + 4A_{1k}A_{2k}}}{2A_{1k}}\right), \qquad (11)$$

where

$$\begin{cases} A_{1k} = \frac{1}{2}\sum_i(|x_{ik}| + x_{ik})p_i(\hat{\mathbf{w}}^{old}), \\ A_{2k} = \frac{1}{2}\sum_i(|x_{ik}| - x_{ik})p_i(\hat{\mathbf{w}}^{old}), \\ B_j = \sum_i x_{ik}y_i, \end{cases} \qquad (12)$$

and $p_i(\hat{\mathbf{w}}^{old}) = \frac{e^{{x_i}^T\hat{\mathbf{w}}^{old}}}{1+e^{{x_i}^T\hat{\mathbf{w}}^{old}}}$, $S = \max_i \sum_k |x_{ik}|$.

Iterative Scaling Approach

Version 2

Algorithm 2 Iterative Scaling Approach Version 2.

Initialization: Set $t = 0$, $\hat{\mathbf{w}}^{(0)}$, $\mathbf{X}$, $\mathbf{y}$

Repeat

$S = \max_i \sum_k |x_{ik}|$;

$p_i(\hat{\mathbf{w}}^{(t)}) = \frac{e^{\mathbf{x}_i{}^T \hat{\mathbf{w}}^{(t)}}}{1+e^{\mathbf{x}_i{}^T \hat{\mathbf{w}}^{(t)}}}$;

$A_{1k} = \frac{1}{2} \sum_i (|x_{ik}| + x_{ik}) p_i(\hat{\mathbf{w}}^{(t)})$;

$A_{2k} = \frac{1}{2} \sum_i (|x_{ik}| - x_{ik}) p_i(\hat{\mathbf{w}}^{(t)})$;

$B_j = \sum_i x_{ik} y_i$;

$\hat{w}_k^{(t+1)} = \hat{w}_k^{(t)} + \frac{1}{S} \log\left(\frac{B_k + \sqrt{B_k^2 + 4A_{1k}A_{2k}}}{2A_{1k}}\right)$;

$t = t + 1$;

Until convergence

Gradient Descent Approach

- Gradient descent iterations:

$$\begin{aligned}\hat{\mathbf{w}}^{new} &= \hat{\mathbf{w}}^{old} + \eta \nabla_{\mathbf{w}} l(\hat{\mathbf{w}}^{old}) \\ &= \hat{\mathbf{w}}^{old} + \eta \sum_{i=1}^{n} \left(y_i - \frac{e^{\hat{\mathbf{w}}^{old^T} \mathbf{x}_i}}{1 + e^{\hat{\mathbf{w}}^{old^T} \mathbf{x}_i}} \right) \mathbf{x}_i, \end{aligned} \qquad (13)$$

where η is the step size.

Algorithm 3 Gradient Descent Approach.

1: **Initialization:** Set $t = 0$, $\hat{\mathbf{w}}^{(0)}$, $\mathbf{X}$, $\mathbf{y}$
2: **Repeat**
3: $\hat{\mathbf{w}}^{(t+1)} = \hat{\mathbf{w}}^{(t)} + \eta \nabla_{\mathbf{w}} l(\hat{\mathbf{w}}^{(t)})$,
$= \hat{\mathbf{w}}^{(t)} + \eta \sum_{i=1}^{n} \left(y_i - \frac{e^{\left(\hat{\mathbf{w}}^{(t)}\right)^T \mathbf{x}_i}}{1+e^{\left(\hat{\mathbf{w}}^{(t)}\right)^T \mathbf{x}_i}} \right) \mathbf{x}_i$;
4: $t = t + 1$;
5: **Until** convergence

CRLB and MSE as a function of the data size n

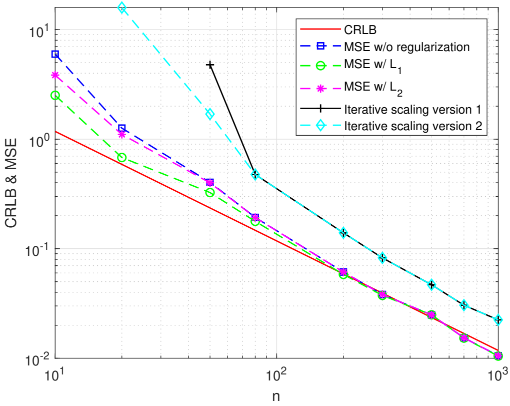


Figure: **CRLB** and **MSE** of **ML** estimations as a function of n for $\mathbf{w} = [1, 1]^T\sqrt{2}$.

Experimental Results
CRLB and MSE as a function of the data size n

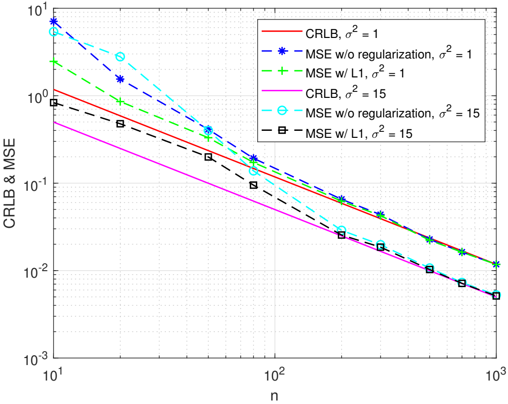


Figure: **CRLB** and **MSE** of **ML** estimations (no regularization, l_1-regularization) as a function of n with $d = 2$ for different values of σ^2.

Experimental Results
CRLB and MSE as a function of the norm of $\| \mathbf{w} \|$

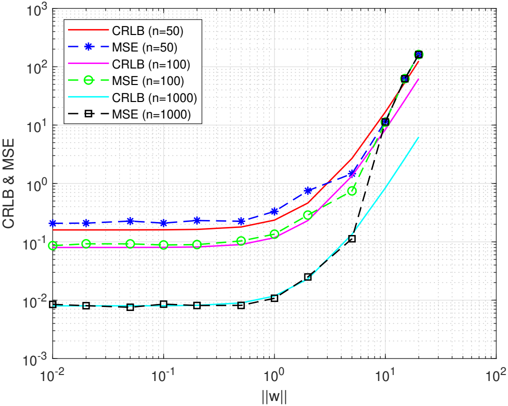
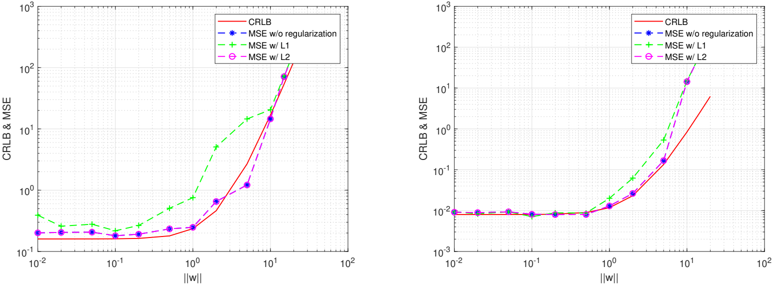


Figure: **CRLB** and **MSE** as a function of $\| \mathbf{w} \|$ for $n \in \{50, 100, 1000\}$.

Experimental Results

CRLB and MSE of ML estimation



(a) $n = 50$ (b) $n = 1000$

Figure: **CRLB** and **MSE** of **ML** estimations (no regularization, l_1-regularization) as a function of $\| \mathbf{w} \|$ with $d = 2$ for (a) $n = 50$ and (b) $n = 1000$.

Conclusion

- The CRLB and MSE of ML estimation are derived.
- The iterative scaling and gradient descent approaches are used to estimate the parameter vector.
- The CRLB and MSE of ML estimation are compared with the iterative scaling and gradient descent approaches.

References

- Xiaoli Z. Fern, "Logistic Regression," *AI534 Machine Learning,* Oregon State University, 2023.
- T. Nguyen, R. Raich and P. Lai, "Jeffreys prior regularization for logistic regression," *2016 IEEE Statistical Signal Processing Workshop (SSP),* Palma de Mallorca, Spain, 2016, pp. 1-5.
- Thomas P Minka, "Algorithms for maximum-likelihood logistic regression," 2003.
- R. Raich, "Iterative scaling note," Oregon State University, 2023.