L_1-Penalized Logistic Regression

Setia Pramana

Medical Epidemiology and Biostatistics Department,
Karolinska Institute, Stockholm

Introduction

- Logistic regression is a supervised method for binary or multi-class classification.
- Logistic regression models the probability of membership of a class with transforms of linear combinations of explanatory variables.

Introduction

- In high-dimensional data (e.g., microarray): More variables than the observations → Classical logistic regression does not work.
- Other problems: Variables are correlated (multicolinierity) and overfitting.
- Solution: Introduce a penalty for complexity in the model.

Logistic Regression

- Let y_j is an array of binary (e.g., malaria types) and x_j is expression level of the *j*-th protein. We can fit the following logistic model:

$$logit(p(x_i)) = log\left\{\frac{p(x_i)}{1-p(x_i)}\right\} = \beta_0 + \sum_{j=1}^{m}\beta_j x_j,$$

- $p(x_i)$: probability that an array with measured expression x represents a class of type $y = 1$.
- Maximize the log-likelihood:

$$\ell(\beta) = \sum_{i=1}^{n}\{y_i log(p(x_i)) + (1-y_i)log(1-p(x_i))\}.$$

Penalized Logistic Regression

- Penalized log-likelihood:

$$\ell^*(\beta) = \ell(\beta) - \lambda J(\theta),$$

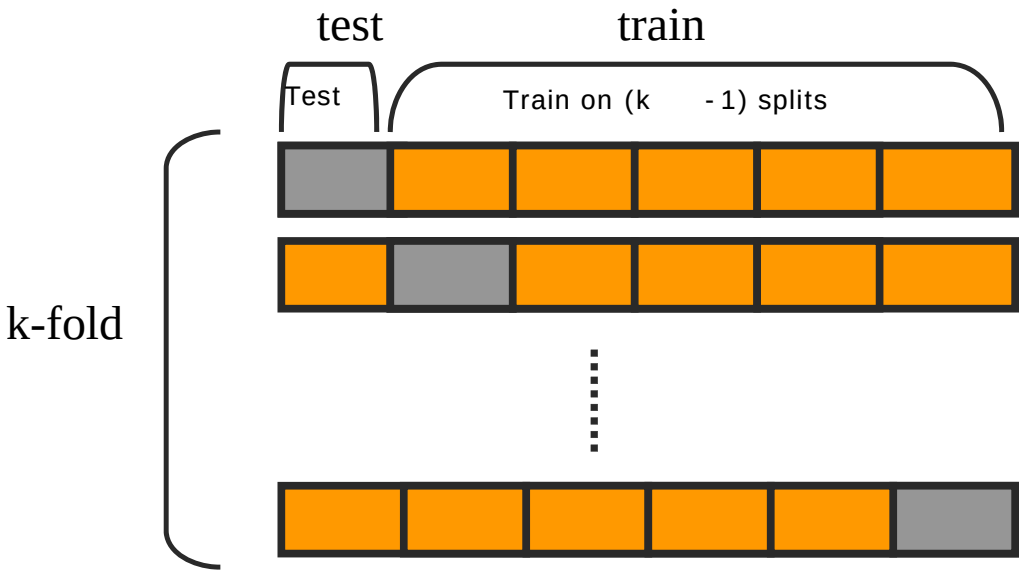
- λ: Tuning parameter
- $J(\theta)$: a penalty function
- L_1-penalization (Lasso):

$$J(\theta) = \sum_{i=1}^{n} \|\beta_i\|$$

L_1-Penalized Logistic Regression

- Shrinks all regression coefficients (β) toward zero and set some of them to zero.
- Performs parameter estimation and variable selection at the same time.
- L_1-penalized log-likelihood is maximized using *full gradient* algorithm (Goeman, 2010).
- The choice of λ is crucial and chosen via cross-validation procedure.
- The procedure is implemented in an R package called `penalized`.

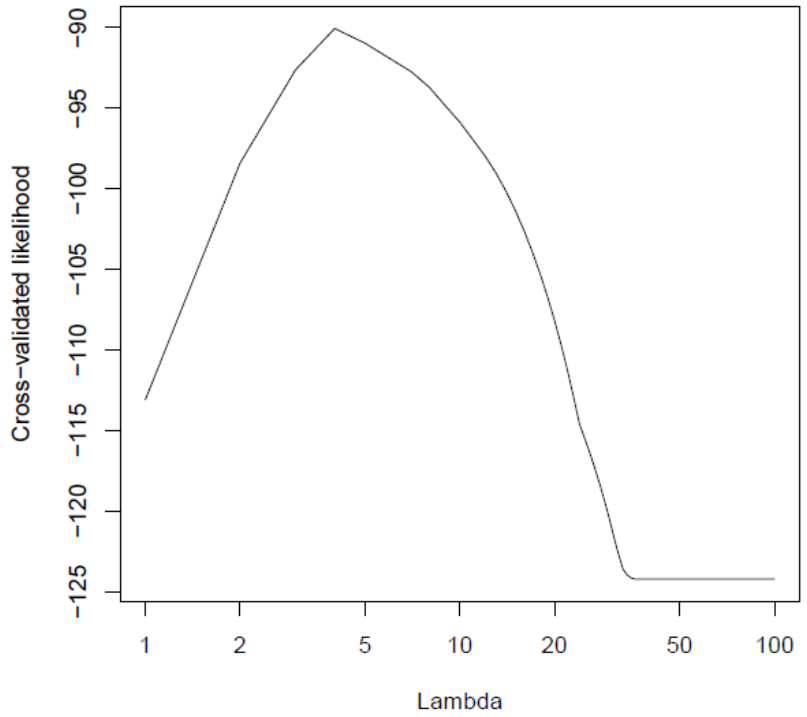
K-fold Cross Validation



- Randomly divide your data into K pieces/folds
- Treat 1st fold as the test dataset. Fit the model to the other folds (training data).
- Apply the model to the test data and repeat k times.
- Calculate statistics of model accuracy and fit from the test data only.

Obtaining The Optimal λ

- Plot of cross-validated likelihood against λ:



- In this example, the optimal λ= 4.11 gives the maximum log-likelihood.

Obtaining The Optimal λ

- Note that in *k*-fold CV method, allocation of the sample into *k* folds are perform randomly.
- Different randomization may lead to slightly different results (λ).
- To obtain more robust λ value, a hundred different cross validations were performed.
- Take the average of $\lambda_1, \lambda_2, \ldots, \lambda_{100}$ ($\bar{\lambda}$).
- Use $\bar{\lambda}$ for fitting the final L_1-penalized logistic model.

Validation?

- Use another experiment for validation

Thank you for your attention !