Machine Learning & Big Data
ECON590
Problem Set 4

1. (10 points) Consider the logistic regression with predictors $x_{i1}, \ldots, x_{ik}$ such that the success probability is given by

$$p(\mathbf{x}_i; \boldsymbol{\beta}) = \mathbb{P}(y_i = 1 \,|\, \mathbf{x}_i, \boldsymbol{\beta}) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})},$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$. Recall that the likelihood function for this model is

$$L(\boldsymbol{\beta} \,|\, \mathbf{y}) = \prod_{i=1}^{n} p(\mathbf{x}_i; \boldsymbol{\beta})^{y_i} (1 - p(\mathbf{x}_i; \boldsymbol{\beta}))^{1-y_i}.$$

Similar to the linear regression, we can also consider a shrinkage estimator of $\boldsymbol{\beta}$. More specifically, consider the penalized log-likelihood of the form

$$f(\boldsymbol{\beta}) \equiv \log L(\boldsymbol{\beta} \,|\, \mathbf{y}) - \lambda \sum_{i=1}^{k} \beta_i^2,$$

where $\lambda > 0$ is a penalty term.

(a) Compute the gradient of the penalized log-likelihood, $\nabla f(\boldsymbol{\beta})$.
[Hint: note that the intercept is not penalized.]

(b) The dataset logit_ridge.csv contains 100 observations with $k = 20$ predictors. The first column contains y, and the second to the last columns contain the predictors. Use the first 10 observations as the test set and the remaining 90 observations as the training set. Set $\lambda = 1$. Apply gradient ascent with tuning parameter $\alpha = 10^{-4}$ to find the penalized maximum likelihood estimate of $\boldsymbol{\beta}$ — i.e., the value of $\boldsymbol{\beta}$ that maximizes $f(\boldsymbol{\beta})$ — using only data in the training set. What are maximum likelihood estimates of β_1 and β_2?

(c) Define the prediction error of the i-th observation in the test data as

$$e_i^2 = (y_i - p(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}))^2,$$

where $\widehat{\boldsymbol{\beta}}$ is the penalized maximum likelihood estimate obtained using the training set. What is the average test error of the first 10 observations?

2. (10 points) In this exercise we derive the Bayesian connection to the Lasso.

(a) Consider the following linear regression without an intercept:

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed as $\mathcal{N}(0, \sigma^2)$ and σ^2 is known. What is the likelihood for the data?

[Hint: the density of the $\mathcal{N}(\mu, \omega^2)$ distribution is

$$p(z; \mu, \omega^2) = (2\pi\omega^2)^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2\omega^2}(z-\mu)^2}.$$

]

(b) Assume the following prior for $\boldsymbol{\beta}$: $\beta_1, \ldots, \beta_p$ are independent and identically distributed as double exponential random variables with density function

$$p(\boldsymbol{\beta}) = \prod_{j=1}^{p} \frac{1}{2\tau} \mathrm{e}^{-|\beta_j|/\tau}.$$

Using the Bayes' Theorem, derive the posterior for $\boldsymbol{\beta}$.

(c) By choosing τ appropriately, show that the mode of the posterior distribution $p(\boldsymbol{\beta} \,|\, \mathbf{y})$ is the same as the Lasso estimate.

[Hint: the value of $\mathbf{z}$ that maximizes $f(\mathbf{z})$ is the same as the value of $\mathbf{z}$ that minimizes $-f(\mathbf{z})$.]