ECE 792-41 Homework 4
Material Covered: Regularization, Classification, Generalized Linear Models

Problem 1 (Regularization with Centered Response and Predictors) In Homework 2.2, you have proved that for simple linear regression with a centered predictor, the estimate for the intercept is $\bar{y}$. In this problem, you will show that using centered predictors for multiple regression with some regularizer on the weights, the same result still holds.

a) Show that the following two optimization problems are equivalent. Give the correspondence between $\underset{\sim}{\beta}^c$ and the original $\underset{\sim}{\beta}$.

$$\hat{\underset{\sim}{\beta}}^{\text{ridge}} = \underset{\underset{\sim}{\beta}}{\text{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}. \tag{1}$$

$$\hat{\underset{\sim}{\beta}}^{c} = \underset{\underset{\sim}{\beta}^c}{\text{argmin}} \left\{ \sum_{i=1}^{N} \left[y_i - \beta_0^c - \sum_{j=1}^{p} (x_{ij} - \bar{x}_j)\, \beta_j^c \right]^2 + \lambda \sum_{j=1}^{p} {\beta_j^c}^2 \right\}. \tag{2}$$

b) For (2), show that $\hat{\beta}_0^c = \bar{y}$.

c) Use the results from (a) and (b), explain why the following steps for obtaining $\hat{\underset{\sim}{\beta}}^{\text{ridge}}$ are correct.

1. Regress $(y_i - \bar{y})$ on centered predictors $\{x_{ij} - \bar{x}_j\}_{j=1}^p$ with the squared penalty term to obtain $\{\hat{\beta}_j^c\}_{j=1}^p$. (Hint: Is the value of $\hat{\beta}_0^c$ dependent on the value of $\{\hat{\beta}_j^c\}_{j=1}^p$ and/or the regularization term?)
2. Obtain the ridge solution by setting $\hat{\beta}_j^{\text{ridge}} = \hat{\beta}_j^c$ for $j = 1, \ldots, p$ and $\hat{\beta}_0^{\text{ridge}} = \bar{y} - \sum_{j=1}^p \bar{x}_j \hat{\beta}_j^c$.

Problem 2 (Error Metrics for Binary Classification) Assume data points are generated from two distributions, namely, $\mathcal{N}(\mu_0, \sigma_0^2)$ for Class 0 and $\mathcal{N}(\mu_1, \sigma_1^2)$ for Class 1, where $\mu_0 < \mu_1$ and $\sigma_0 > \sigma_1$.

a) Draw by hand an illustration that contains the PDFs for both classes. Your illustration must reflect the relative relations of the means and standard deviations. Pick an arbitrary decision threshold on the horizontal axis and denote it as η. Use shades to label the areas under curves corresponding to the false positive rate (FPR) and the false negative rate (FNR), respectively. Show that

$$\text{FPR}(\eta) = 1 - \Phi\left(\frac{\eta - \mu_0}{\sigma_0}\right) \ \text{ and } \ \text{FNR}(\eta) = \Phi\left(\frac{\eta - \mu_1}{\sigma_1}\right), \tag{3}$$

where $\Phi(\cdot)$ is the CDF for the standard Gaussian random variable.

b) Prove that a decision threshold leading to the equal error rate (EER) is of the following form:

$$\eta^* = \frac{\sigma_1\mu_0 + \sigma_0\mu_1}{\sigma_0 + \sigma_1}. \tag{4}$$

(Hints: Set $\text{FPR}(\eta^*) = \text{FNR}(\eta^*)$. Exploit this property: The left and right tails of a *standard* Gaussian distribution have the same probability.) Prove that

$$\text{EER} = \text{FPR}(\eta^*) = \text{FNR}(\eta^*) = \Phi\left(\frac{\mu_0 - \mu_1}{\sigma_0 + \sigma_1}\right). \tag{5}$$

c) Let $\mu_0 = 0$, $\mu_1 = 5$, $\sigma_0 = 2$, and $\sigma_1 = 1$. Simulate $N = 5000$ data points for each class. Generate an empirical ROC curve by varying threshold from the smallest value to the largest value of the overall dataset.

d) Draw the theoretical ROC curve using the results in part (a). Is the empirical ROC curve consistent with theoretical one?

Problem 3 (Generalized Linear Model) Response $Y_i \sim \text{B}(n, p_i)$ is a binomial random variable in which n is known. The (conditional) PDF is shown as follows:

$$\mathbb{P}[Y_i = k | \underset{\sim}{X}_i = \underset{\sim}{x}_i] = \binom{n}{k} p_i^k (1 - p_i)^{n-k}, \quad k \in \{0, 1, \ldots, n\}. \tag{6}$$

a) Explain why the linear regression may not the best fit to find the relation between Y_i and a set of predictors $X_{i,1}, \ldots, X_{i,q}$.

b) One proposes to link the conditional mean μ_i and the predictors $\underset{\sim}{x}_i$ using a generalized linear model shown as follows:

$$g(\mu_i) = \underset{\sim}{\beta}^T \underset{\sim}{x}_i \tag{7}$$

where $g(u) = \log(\frac{u}{n-u})$ and $\mu_i = \mathbb{E}[Y_i | \underset{\sim}{X}_i = \underset{\sim}{x}_i] = np_i$. From the variable transformation viewpoint, show that $g(\cdot)$ matches the ranges for the two sides of Eq. (7).

c) Rewrite the PDF into an exponential family form shown as follows:

$$f_Y(y; \theta) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right), \tag{8}$$

where θ is the natural parameter. Show that $g(\cdot)$ in (b) is the canonical link function when taking μ_i as the input.

(This homework has only three problems. Please start your review for the midterm early.)