NOTES. There are two sections of the homework. Section 1 is required for all students. While Section 2 is only required for Ph.D. students in Statistics Department, you are all encouraged to try and earn bonus points. All problems are to be done individually. Keep the code you develop as you and your group may be asked to present your work later. Problems in Section 2 are to be done individually.

1. Data Analysis

1. *Prostate Cancer Data.* For the `prostate` data available on the book website (see § 3.2.1 for a description), carry out a best-subset linear regression analysis, as in Table 3.3 (third column from left). Compute the AIC, BIC, five- and tenfold cross-validation, and bootstrap .632 estimates of prediction error. Discuss the results. [Hint. (i) Use the subset size as the effective number of parameters. (ii) Use (7.30) to compute AIC, and (7.36) to compute BIC.] Here are some details for cross-validation and bootstrap methods. For each of them, I list two approaches. You are free to use any one of them, and encouraged to try both.

(a) *Cross Validation: Approach I.* Since there are 8 inputs, there are in total $2^8 = 256$ models corresponding to 256 subsets. Use cross-validation to estimate the prediction error for each of these 256 models, and then pick the best one. Finally, fit the chosen model using the whole training set and report the results.

(b) *Cross Validation: Approach II.* For each subset size $\alpha \in \{0, 1, \ldots, 8\}$, do cross-validation as follows.
   (1) Split the training set into K folds. You can use for example $K = 5$.
   (2) Use the k-th fold as the validation set, and fit all $\binom{8}{\alpha}$ models with subset size α using all training data that are not in the k-th fold. Among these $\binom{8}{\alpha}$ models, pick the one with the smallest training error, and then compute the prediction error of this model on the validation set which is the k-th fold.
   (3) Repeat Step (2) for each $1 \le k \le K$, and then take the average of the K prediction errors. This average would be your estimated prediction error for best subset selection with size α.

   Once you obtain the estimated prediction error for each size α, you can pick the size $\hat{\alpha} \in \{0, 1, \ldots, 8\}$ with the smallest prediction error. Finally, fit all $\binom{8}{\hat{\alpha}}$ models of size $\hat{\alpha}$ using the whole training set, pick the one with the smallest training error and report the fitting results of this model.

   Here are some implementation issues of cross-validation.
   - When you split the training set into, say 5, folds, you should do a random splitting. This can be done assigning a random label uniformly chosen from $\{1, 2, 3, 4, 5\}$ to each training point. Then group the training data according to the labels.
   - Note that the cross-validation is used only on the training set.
   - When you pick the best subset using Approach I, or best subset size using Approach II, you can also applying the *one-standard-error rule* we talked about in the lecture.
   - If you find difficult to list all subsets, try the `R` function `combn` that is available from the package `combinat`.

(c) *Bootstrap: Approach I.* Use bootstrap .632 estimates for each of the 256 models, and pick the best one. Fit this model using the whole training set and report the results.

(d) *Bootstrap: Approach II.* For each subset size $\alpha \in \{0, 1, \ldots, 8\}$, compute the bootstrap .632 estimate as follows.
   (1) Sample with replacement from the training set to get a bootstrap sample of the same size as the training set.
   (2) Fit all $\binom{8}{\alpha}$ models with subset size α using the bootstrap sample, pick the model with the smallest training error.
   (3) Repeat Steps (1) and (2) B times. You can take for example $B = 1000$ or $B = 5000$.
   (4) Use the model you pick for each bootstrap sample to compute the *leave-one-out* bootstrap estimate of the prediction error.
   (5) Fit all $\binom{8}{\alpha}$ models using the whole training set, and find the smallest training error.
   (6) Use a linear combination of the results from Step (4) and (5) to compute the bootstrap .632 estimate.

Once you obtain the estimated prediction error for each size α, you can pick the size $\hat{\alpha} \in \{0, 1, \ldots, 8\}$ with the smallest prediction error. Finally, fit all $\binom{8}{\hat{\alpha}}$ models of size $\hat{\alpha}$ using the whole training set, pick the one with the smallest training error and report the fitting results of this model.

2. *Neural Networks.* Consider the following two models:

$$Y = \sigma(a_1^T X) + \sigma(a_2^T X) + 0.30 \cdot Z \tag{1.1}$$

$$Y = \sigma(a_1^T X) + (a_2^T X)^2 + 0.30 \cdot Z \tag{1.2}$$

where σ is the sigmoid function, Z is standard normal, $X^T = (X_1, X_2)$ with X_1 and X_2 being independent standard normal, and $a_1^T = (3, 3)$, $a_2^T = (3, -3)$. For each model, carry out the following analysis.

(a) Generate a training sample of size 100, and a test sample of size 1000. Apply a single layer neural network with 10 hidden units. Plot the training and test error curves as a function of the number of training epochs, for different values of the weight decay parameter. Discuss the over-fitting behavior in each case.

(b) Vary the number of hidden units in the network, from 1 up to 10, and determine the minimum number needed to perform well. Do this step with no weight decay and with weight decay that you choose from part (a).

Some issues.

- You may want to try multiple (say, 10) starting weights.
- You may want to rescale the inputs.

2. Theoretical Problems for Ph.D. Students

3. *Expected Optimism.* Consider a binary classification problem with 0-1 loss. Show that the following formula for the expected optimism holds

$$\omega = \frac{2}{N} \sum_{i=1}^{N} \mathrm{Cov}(\hat{y}_i, y_i).$$

4. *Generalized Cross-Validation.* Let $\hat{\boldsymbol{f}} = \boldsymbol{S}\boldsymbol{y}$ be a linear smoothing of $\boldsymbol{y}$. If S_{ii} is the i-th diagonal entry of $\boldsymbol{S}$, show that for $\boldsymbol{S}$ arising from least squares and cubic smoothing splines, the cross-validated residual can be written as

$$y_i - \hat{f}^{-i}(x_i) = \frac{y_i - \hat{f}(x_i)}{1 - S_{ii}}. \tag{2.1}$$

Find general conditions on any smoothing matrix $\boldsymbol{S}$ to make (2.1) holds.

5. *Cross-entropy for SNN.* Derive the forward and backward propagation equations for the cross-entropy loss function.