140.644: Statistical Machine Learning **Name:** _____

Instructor: Yiqun Chen TA: Junwei (Ivy) Sun

Quiz 4 — Practice — Model Selection, Regularization, and Evaluation

Not graded. Bring questions to class.

This is a shortened sample so you know what to expect. The real quiz has **4 multiple choice** and **2 questions in Part B**, and is worth 100 points; this one has 3 and 2, and the table compares two fits rather than four. The style and the difficulty are the same. No calculator is needed.

Part A: Multiple Choice [30 points]

Circle the single best answer. **(10 points)** each.

1. Which of the following is true of ridge regression but *not* of the lasso?

   (A) it can set coefficients exactly to zero

   (B) it shrinks every coefficient toward zero but never sets one exactly to zero

   (C) it needs no tuning parameter

   (D) its predictors must be standardized, whereas the lasso's need not be

2. The `NCI60` data have gene expression for 6,830 genes on 64 cancer cell lines. A student wants to build a KNN classifier of cell-line subtype using only the strongest genes, and uses `make_folds()` and `cv_error()` (trusted helpers for 5-fold cross-validation *within the training set*) as follows:

   R

   ```
   assoc <- apply(X, 2,
     function(g) cor(g, y))
   top20 <- order(abs(assoc),
     decreasing = TRUE)[1:20]
   X20 <- X[, top20]
   folds <- make_folds(
     X20, k = 5)
   cv_error(X20, y, folds,
            knn_fit)
   ```

   Python

   ```
   assoc = [np.corrcoef(
       X[:, j], y)[0, 1]
       for j in range(p)]
   top20 = np.argsort(
       -np.abs(assoc))[:20]
   X20 = X[:, top20]
   folds = make_folds(
     X20, k=5)
   cv_error(X20, y, folds,
            knn_fit)
   ```

   What is the problem with this code?

   (A) the 20 genes are selected by their association with subtype using the entire training set, so each fold's held-out rows already helped choose the genes it is later scored on

   (B) correlation is not a valid way to screen predictors before classification; a t-test must be used instead

   (C) 20 genes is too few for KNN to compute a meaningful distance across 64 cell lines

   (D) 5-fold cross-validation is invalid once features have been selected; only a single train/test split can be used afterward

3. A team has 300 training observations and uses 5-fold cross-validation to choose k for a k-nearest-neighbors classifier. In each of the five rounds, the model is fitted on:

   (A) 240 observations, and its error is measured on the other 60

(B) all 300 observations, and its error is measured on the same 300

(C) 60 observations, and its error is measured on the other 240

(D) 240 observations, and its error is measured on the test set

Part B: Ridge and Lasso on the Same Data **[40 points]**

The two columns below are a ridge fit and a lasso fit to the same 300 training cards from the `Credit` data seen in lecture, predicting card `Balance` from seven variables. One is ridge with $\lambda = 20$ and the other is lasso with $\lambda = 40$.

	Model A	**Model B**
(Intercept)	−304	−501
`Income`	−2.03	−6.33
`Limit`	0.121	0.129
`Rating`	0.91	1.75
`Cards`	.	15.0
`OwnYes`	.	−14.6
`MarriedYes`	.	−3.9
`StudentYes`	253	389

A dot means the coefficient is exactly zero. `glmnet` *standardized the predictors internally; coefficients are reported on the original scale.*

1. **(20 points)** Which column is the ridge fit and which is the lasso fit? Which feature of the table decides it? Cite specific numbers.

2. **(20 points)** The lasso is refitted with a much larger penalty, $\lambda = 100$. Describe two things you would expect to change in its column, and say why.