Statistical Learning: Homework Assignment 1

(Not to turn in)

1. (Ex.2.3, 5 points) Consider N data points uniformly distributed in a p-dimensional unit ball centered at the origin. Suppose we consider a nearest-neighbor estimate at the origin. Show that the median distance from the origin to the closest data point is given by the expression

$$d(p, N) = \left(1 - \left(\frac{1}{2}\right)^{1/N}\right)^{1/p}.$$

2. (Ex.2.9, 10 points) Consider a linear regression model with p parameters, fit by least squares to a set of training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ drawn at random from a population. Let $\hat{\boldsymbol{\beta}}$ be the least squares estimate. Suppose we have some test data $(\tilde{\mathbf{x}}_1, \tilde{y}_1), \ldots, (\tilde{\mathbf{x}}_M, \tilde{y}_M)$ drawn at random from the same population as the training data. If $R_{tr}(\boldsymbol{\beta}) = \frac{1}{N}\sum_1^N (y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2$ and $R_{te}(\boldsymbol{\beta}) = \frac{1}{M}\sum_1^M (\tilde{y}_i - \tilde{\mathbf{x}}_i^T\boldsymbol{\beta})^2$, prove that

$$E[R_{tr}(\hat{\boldsymbol{\beta}})] \leq E[R_{te}(\hat{\boldsymbol{\beta}})]$$

where the expectations are over all that is random in each expression.

3. (Ex.3.3(b), 5 points) The matrix inequality $\mathbf{B} \preceq \mathbf{A}$ holds if $\mathbf{A} - \mathbf{B}$ is positive semidefinite. Show that if $\hat{\mathbf{V}}$ is the variance-covariance matrix of the least squares estimator of $\boldsymbol{\beta}$ and $\tilde{\mathbf{V}}$ is the variance-covariance matrix of any other linear unbiased estimator, then $\hat{\mathbf{V}} \preceq \tilde{\mathbf{V}}$.

4. (Ex.3.6, 5 points) Show that the ridge regression estimate is the mean (and mode) of the posterior distribution, under a Gaussian prior $\boldsymbol{\beta} \sim N_p(0, \tau^2 \mathbf{I}_p)$ and Gaussian sampling model $\mathbf{y} \sim N_N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_N)$ with $\lambda = \sigma^2/\tau^2$.

5. (Ex.3.12, 5 points) Show that the ridge regression estimates can be obtained by ordinary least squares regression on an augmented data set. We augment the centered matrix $\mathbf{X}$ with p additional rows $\sqrt{\lambda}\mathbf{I}$, and augment $\mathbf{y}$ with p zeros. By introducing artificial data having response value zero, the fitting procedure is forced to shrink the coefficients toward zero.

6. (10 points) Let $\mathbf{X}_{N\times p}$ and $\mathbf{y}$ be the design matrix and response vector for the standardized data. Assume the linear model is correct, i.e. $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$. Derive an explicit expression for the MSE of the ridge estimator $\hat{\boldsymbol{\beta}}^{Ridge} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$, i.e.

$$MSE(\lambda) = E[(\hat{\boldsymbol{\beta}}^{Ridge} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}^{Ridge} - \boldsymbol{\beta})],$$

and show that

$$\frac{d}{d\lambda} MSE(\lambda)\Big|_{\lambda=0+} < 0.$$

7. (10 points) Suppose $z_i \sim ind\ N(\mu_i, 1),\ i = 1, \ldots, p\ (p \geq 3)$. For the James-Stein estimator

$$\hat{\boldsymbol{\mu}}^{JS} = \left(1 - \frac{p-2}{S}\right)\mathbf{z}$$

where $S = \sum_{i=1}^{p} z_i^2$, show that

$$E_\mu\left\{||\hat{\boldsymbol{\mu}}^{JS} - \boldsymbol{\mu}||^2\right\} < E_\mu\left\{||\hat{\boldsymbol{\mu}}^{ML} - \boldsymbol{\mu}||^2\right\} = p$$

for every choice of $\boldsymbol{\mu}$.

8. (10 points) Show that $\hat{\theta}^{Lasso}$ which minimizes

$$F(\theta) = (z-\theta)^2 + \lambda|\theta|$$

is given by

$$\hat{\theta}^{Lasso} = \text{sign}(z)(|z| - \lambda/2)_+.$$

9. (30 points) Consider a regression problem with all variables and response having mean zero and standard deviation one. Suppose also that each variable has identical absolute correlation with the response:

$$\frac{1}{N}|\langle \mathbf{x}_j, \mathbf{y}\rangle| = \lambda, \ j = 1, \ldots, p.$$

Let $\hat{\boldsymbol{\beta}}$ be the least-squares coefficient of $\mathbf{y}$ on $\mathbf{X}$, and let $\mathbf{u}(\alpha) = \alpha\mathbf{X}\hat{\boldsymbol{\beta}}$ for $\alpha \in [0, 1]$ be the vector that moves a fraction α toward the least squares fit $\mathbf{u}$. Let RSS be the residual sum-of squares from the full least squares fit.

(a) (10 pts) Show that

$$\frac{1}{N}|\langle \mathbf{x}_j, \mathbf{y} - \mathbf{u}(\alpha)\rangle| = (1-\alpha)\lambda, \ j = 1, \ldots, p,$$

and hence the correlations of each $\mathbf{x}_j$ with the residuals remain equal in magnitude as we progress toward $\mathbf{u}$.

(b) (20 pts) Show that these absolute correlations are all equal to

$$\lambda(\alpha) = \frac{(1-\alpha)}{\sqrt{(1-\alpha)^2 + \frac{\alpha(2-\alpha)}{N}RSS}}\lambda,$$

and they decrease monotonically to zero.

10. (30 points) In the k-th LAR step, let $\mathcal{A}_k$ be the active set, $\mathbf{u}_k$ be the LAR direction, and $\mathbf{r}_k$ be the residual vector at the beginning of the k-th step. Show that the absolute correlation between $\mathbf{r}_k(\alpha) = \mathbf{r}_k - \alpha\mathbf{u}_k$ and each of the predictors in $\mathcal{A}_k$ are all equal for any $0 \leq \alpha \leq 1$, and the absolute correlation is decreasing as $\alpha \to 1$.

11. (20 points) Derive expressions to identify the next variable to enter the active set at step $k+1$, and the value of α at which this occurs.

12. (30 points) For given $\lambda > 0$, let $\hat{\boldsymbol{\beta}}(\lambda)$ be the lasso solution that minimizes
$$\frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1$$
and $\mathcal{B}_\lambda$ be the active set of variables in the solution.

    (a) (10 pts) Show that for any active variable $\mathbf{x}_j$ ($j \in \mathcal{B}_\lambda$) we have
    $$\mathbf{x}_j'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\lambda)) = \lambda \cdot \text{sign}(\hat{\beta}_j(\lambda)).$$

    (b) (10 pts) Show that for any non-active variable $\mathbf{x}_k$ ($k \notin \mathcal{B}_\lambda$) we have
    $$|\mathbf{x}_k'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\lambda))| \leq \lambda.$$

    (c) (10 pts) Suppose that the set of active predictors is unchanged for $\lambda_0 \geq \lambda \geq \lambda_1$. Show that there is a vector $\boldsymbol{\gamma}_0$ such that
    $$\hat{\boldsymbol{\beta}}(\lambda) = \hat{\boldsymbol{\beta}}(\lambda_0) - (\lambda - \lambda_0)\boldsymbol{\gamma}_0$$
    Thus the lasso solution path is linear as λ ranges from λ_0 to λ_1.