HW3

1 Ridge regression using SVD

Let $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ be the SVD of the design matrix, and let $\mathbf{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$ be the ridge estimate. Show that

$$\mathbf{w} \quad = \quad \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^T\mathbf{y} \tag{1}$$

2 Ridge regression with diagonal prior

Consider the following generalized ridge regression problem

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} ||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2 + \sum_{j=1}^{d} \lambda_j w_j^2 \tag{2}$$

where we have a different penalty term λ_j for each dimension (feature) j. (We are regularizing the offset term, for notational simplicity.)

1. Show how we can compute $\hat{\mathbf{w}}$ using ordinary least squares applied to a modified (expanded) design matrix.
2. Write down some Matlab code that implements your equations from the previous step. (It should only take about 3 lines.) The input is a matrix $\mathbf{X}$, a vector $\mathbf{y}$, and a *vector* $\boldsymbol{\lambda}$. The output should be $\hat{\mathbf{w}}$. Do not worry about standardizing the data or adding a column of 1s.

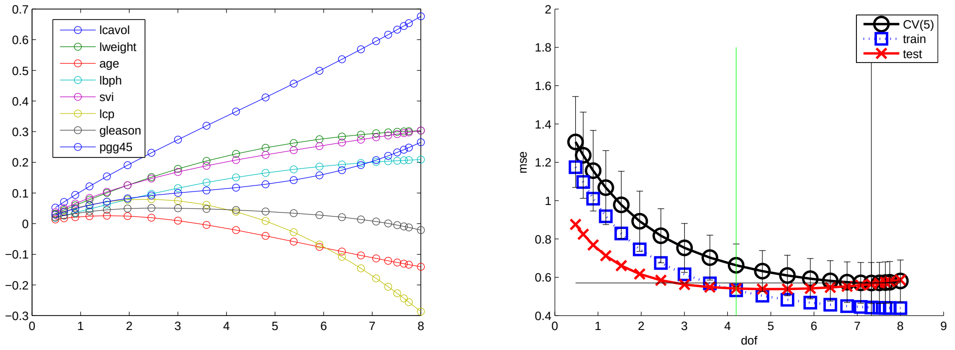
3 Ridge regression on prostate cancer data (Matlab)

Consider the prostate cancer dataset used in HW2. Fit a simple linear model $\hat{y}(x) = w_0 + w_1x_1 + \ldots + w_8x_8$ by ridge regression. Use 5-fold CV to select λ from the range 10^3 to 0. Plot $\mathbf{w}$ vs $df(\lambda)$ (degrees of freedom), as in Figure 1(a). Plot the CV error vs $df(\lambda)$ and indicate the value of λ chosen by the one-standard-error rule, as in Figure 1(b). What coefficients $\mathbf{w}$ do you get? What is the mean squared error and its standard error on the test set? Turn in your numbers, plot and code. (You should get similar results to Table 1, right column.) Hint: download BLT and modify the code for the static method `demoPolyFitRidgeCV` in the `linregDist` class. Just change the data, and use the `addOnesTransformer`. Also, plot vs df rather than vs log(λ).

4 Gradient and Hessian of log-likelihood for logistic regression

1. Let $\sigma(a) = \frac{1}{1+e^{-a}}$ be the sigmoid function. Show that

$$\frac{d\sigma(a)}{da} = \sigma(a)(1 - \sigma(a)) \tag{3}$$



(a)

(b)

Figure 1: (a) Profiles of ridge coefficients vs $df(\lambda)$ on the prostate cancer data. (b) Cross-validation errors. Vertical green line on left is the value of λ chosen by the one standard error rule. Vertical black line on right is the value of λ corresponding to the lowest mean CV error. Produced by Exercise 3

Term	LS	ridge
intercept	2.480	2.472
lcavol	0.676	0.366
lweight	0.303	0.228
age	-0.141	-0.021
lbph	0.209	0.151
svi	0.304	0.207
lcp	-0.287	0.039
gleason	-0.021	0.044
pgg45	0.266	0.117
Test MSE	0.586	0.541
SE	0.184	0.170

Table 1: Coefficients and accuracy of least squares and ridge regression on the prostate cancer data. Based on Table 3.3 of [**?**]. Produced by Exercise 3.

2. Let $J(\mathbf{w})$ be the negative log-likelihood for logistic regression:

$$J(\mathbf{w}) = -\ell(\mathbf{w}) = -\sum_i [y_i \log \mu_i + (1-y_i)\log(1-\mu_i)] \tag{4}$$

where $y_i \in \{0,1\}$ and

$$\mu_i = \sigma(\eta_i), \quad \eta_i = \mathbf{w}^T\mathbf{x}_i \tag{5}$$

Show that

$$\nabla_{\mathbf{w}} J(\mathbf{w}) \;=\; \sum_{i=1}^n (\mu_i - y_i)\mathbf{x}_i = \mathbf{X}^T(\boldsymbol{\mu} - \mathbf{y}) \tag{6}$$

Hint: use the chain rule and the previous result.

3. The Hessian can be written as $\mathbf{H} = \mathbf{X}^T\mathbf{S}\mathbf{X}$, where $\mathbf{S} \overset{\text{def}}{=} \text{diag}(\mu_1(1-\mu_1), \ldots, \mu_d(1-\mu_d))$. Show that $\mathbf{H}$ is positive definite. (You may assume that $0 < \mu_i < 1$, so the elements of $\mathbf{S}$ will be strictly positive.)

5 Regularizing separate terms in 2d logistic regression

(Source: Jaaakkola)

1. Consider the data in Figure 2, where we fit the model $p(y=1|\mathbf{x},\mathbf{w}) = \sigma(w_0 + w_1x_1 + w_2x_2)$. Suppose we fit the model by maximum likelihood, i.e., we minimize

$$J(\mathbf{w}) = -\ell(\mathbf{w}, \mathcal{D}_{\text{train}}) \tag{7}$$

where $\ell(\mathbf{w}, \mathcal{D}_{\text{train}})$ is the log likelihood on the training set. Sketch a possible decision boundary corresponding to $\hat{\mathbf{w}}$. (Copy the figure first (a rough sketch is enough), and then superimpose your answer on your copy, since you will need multiple versions of this figure). Is your answer (decision boundary) unique? How many classification errors does your method make on the training set?

2. Now suppose we regularize only the w_0 parameter, i.e., we minimize

$$J_0(\mathbf{w}) = -\ell(\mathbf{w}, \mathcal{D}_{\text{train}}) + \lambda w_0^2 \tag{8}$$

Suppose λ is a very large number, so we regularize w_0 all the way to 0, but all other parameters are unregularized. Sketch a possible decision boundary. How many classification errors does your method make on the training set? Hint: consider the behavior of simple linear regression, $w_0 + w_1x_1 + w_2x_2$ when $x_1 = x_2 = 0$.

3. Now suppose we heavily regularize only the w_1 parameter, i.e., we minimize

$$J_1(\mathbf{w}) = -\ell(\mathbf{w}, \mathcal{D}_{\text{train}}) + \lambda w_1^2 \tag{9}$$

Sketch a possible decision boundary. How many classification errors does your method make on the training set?

4. Now suppose we heavily regularize only the w_2 parameter. Sketch a possible decision boundary. How many classification errors does your method make on the training set?

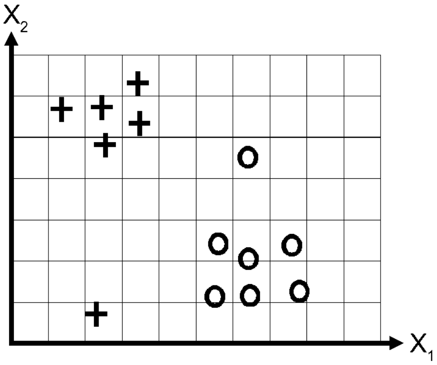


Figure 2: Data for logistic regression question.