Homework 1

Ryan Koo
CSCI 5525

February 1, 2023

Problem 1. (10 points) A generalization of the least squares problem adds an affine function to the objective,

$$\min_w ||Xw - y||_2^2 + a^T w + b$$

where $X \in \mathbb{R}^{n \times d}, w \in \mathbb{R}^d, y \in \mathbb{R}^n, a \in \mathbb{R}^d, b \in \mathbb{R}$. Assume the columns of X are linearly independent. This generalized problem can be solved by reducing it to a standard least squares problem, using a trick called completing the square. Show that the objective of the problem above can be expressed in the form

$$||Xw - y||_2^2 + a^T w + b = ||Xw - y + f||_2^2 + g$$

where $f \in \mathbb{R}^n, g \in \mathbb{R}$. Then solve the generalized least squares problem

$$\arg\min_w ||Xw - y + f||_2^2 + g$$

Proof. In order to complete the square for the generalized least-squares problem $||Xw - y||^2 + a^T w + b$ we first expand the L2 norm to be in the form $(Xw-y)^T(Xw-y)+a^T w+b$ and similarly for the RHS $(Xw-y+f)^T(Xw-y+f)+g$. We show that the two equations are equal to each other

$$= (Xw - y)^T(Xw - y) = a^T w + b = (Xw - y + f)^T(Xw - y + f) + g$$
$$= w^T X^T Xw - 2y^T Xw + y^T y + a^T w + b = w^T X^T Xw - 2(y - f)^T Xw + (y - f)^T(y - f) + g$$
$$= w^T X^T Xw - 2y^T Xw + 2f^T Xw + y^T y - 2y^T f + f^T f + g$$
$$\text{Then, let } a^T w + b = 2f^T Xw - 2y^T f + f^T f + g$$
$$\therefore a = 2X^T f \text{ and } b = -2y^T f + f^T f + g$$

We can solve the least-squares problem by simplifying taking the gradient w.r.t w and then setting the equation to 0 and solving. Thus, similarly above we have

$$\frac{\partial}{\partial w}||Xw - y + f||_2^2 + g =$$
$$\frac{\partial}{\partial w}(Xw - y + f)^T(Xw - y + f). \text{ Let } c = y - f$$
$$\frac{\partial}{\partial w}(Xw - c)^T(Xw - c) = 0 \rightarrow w = (X^T X)^{-1} X^T (y - f)$$

□

Problem 3. We find that the optimal regularization parameter λ for Ridge Regression and Lasso both is $\lambda = 0.01$. For my implementation of Ridge Regression, since the minimizer is a stationary point I found the w vector as

$$w^* = (X^T X + \lambda I)^{-1} X^T y$$

And using my implementation of cross-validation found the mean-squared error over $k = 10$ folds.

Results:
Thus, using a $\lambda = 0.01$ we retrieve the MSE one the test data set as 0.527 for Ridge Regression and 0.535 for Lasso.

λ	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Mean	STD
0.01	0.512	0.501	0.54	0.536	0.56	0.485	0.543	0.531	0.559	0.510	**0.528**	0.023
0.1	0.515	0.504	0.541	0.541	0.559	0.487	0.544	0.532	0.556	0.512	0.529	0.023
1	0.545	0.526	0.562	0.576	0.577	0.505	0.569	0.556	0.565	0.536	0.552	0.022
10	0.597	0.571	0.607	0.635	0.619	0.548	0.614	0.602	0.595	0.583	0.597	0.024
100	0.610	0.585	0.624	0.646	0.632	0.559	0.626	0.613	0.593	0.595	0.608	0.024

Table 1: MSE for Ridge Regression. The best lambda is bolded

λ	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Mean	STD
0.01	0.516	0.517	0.556	0.535	0.567	0.49	0.545	0.535	0.528	0.514	**0.531**	0.215
0.1	0.604	0.604	0.644	0.608	0.632	0.563	0.622	0.611	0.573	0.583	0.605	0.024
1	0.925	0.977	0.986	0.941	0.958	0.909	0.956	1.00	0.941	0.903	0.949	0.031
10	1.318	1.371	1.361	1.308	1.337	1.275	1.342	1.406	1.334	1.272	1.332	0.039
100	1.319	1.371	1.362	1.309	1.337	1.276	1.343	1.408	1.335	1.273	1.333	0.040

Table 2: MSE for Lasso. The best lambda is bolded

Problem 4. For my implementation of Fisher's LDA I calculated the within-class covariance S_w by adding the covariances for each class of the training data. Then using the means of the two classes I took the difference of them and multiplied it with the calculated within-class covariance to get the projection matrix:

$$w \propto S_w^{-1}(\mu_2 - \mu_1)$$

Since we are only doing a binary classification it wasn't necessary to find the between-class covariance for the projection matrix. To find the optimal threshold value, I initially started with the lambda values we originally tested for Ridge Regression and Lasso and mirrored them within for the negative values as well. Then from -10 and -1, I kept taking the middle between the two values until performance stopped improving.

Results: Thus, using a $\lambda = -1.25$ we retrieved the best Err. Rate on the test data set as 0.005. **Note:** not every single λ values was shown here for the sake of concision: $\{-100, -10, -5, -2.5, -1.5, -1.25, -1, -0.1, 0, 0.01, 0.1, 1, 10, 100\}$.

λ	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Mean	STD
-1.5	0.025	0.0125	0.0125	0	0.0187	0.0125	0.0125	0	0.0125	0.0375	0.0144	0.0104
-1.25	0.0187	0.0125	0.0125	0.0062	0.0187	0.0062	0	0	0.00625	0.03125	**0.0112**	0.009
-1	0.0125	0.0062	0.0125	0.0062	0.0187	0.0125	0.0062	0	0.01875	0.03125	0.0125	0.008
-0.1	0.0125	0.0062	0.0062	0.0125	0.0375	0.025	0.025	0.01875	0.025	0.04375	0.0212	0.0119
0.01	0.0125	0.0125	0.0062	0.0125	0.0375	0.0312	0.025	0.01875	0.03125	0.04375	0.0231	0.0118
0.1	0.0125	0.0125	0.0062	0.0125	0.0312	0.0312	0.025	0.03125	0.03125	0.04375	0.0237	0.0114
1	0.05	0.025	0.0375	0.0187	0.0625	0.0687	0.056	0.0375	0.06875	0.0625	0.0487	0.017

Table 3: Err. Rates for LDA. The best lambda is bolded