CSCI 5525

Homework 1

1. ∵ assume the eigen values of X^TX is
$\lambda_1, \lambda_2, \lambda_3 \cdots \lambda_n$
and the eigenvalue of λI is λ

∴ the eigenvalues of $X^TX + \lambda I$ is
$\lambda_1+\lambda \quad \lambda_2+\lambda \quad \cdots \quad \lambda_n+\lambda$

∵ according to the eigen decomposition $X^TX = V\Lambda V^T$

$$\lambda = V^TX^TXV = \|XV\|^2 \geq 0$$

∴ $\lambda_1+\lambda, \lambda_2+\lambda \cdots \lambda_n+\lambda > 0$
for $\lambda > 0$

∴ $X^TX + \lambda I$ is invertible, if $\lambda > 0$.

2.

a.

$$\mathbb{E}_{(x,y)}\left[\ell(f(x), y)\right]$$

$$= \mathbb{E}\left(\mathbb{E}(\ell(f(x), y)) \mid x\right)$$

$$= \mathbb{E}\left(f(x)^2 - 2f(x)\cdot y + y^2 \mid x\right)$$

$$= \mathbb{E}(f(x)^2 - 2f(x)P(y|x) + P(y^2|x))$$

$$= \underset{f(x)}{\arg\min}\left[f(x)^2 - 2f(x)\mathbb{E}(y|x) + \mathbb{E}(y^2|x)\right]$$

$$f(x)^2 - 2f(x)\,\mathbb{E}(y|x) + \mathbb{E}(y^2|x) + \mathbb{E}^2(y|x) - \mathbb{E}^2(y|x)$$

$$= (f(x) - \mathbb{E}(y|x))^2 + \mathbb{E}(y^2|x) - \mathbb{E}^2(y|x)$$

this one not depends on $f(x)$

$\therefore$ $f(x) = \mathbb{E}(y|x)$ is the optimal value.

b.

To optimalize $\int_x \{\int_y l(f(x), y)\, p(y|x)\, dy\}\, p(x)\, dx$

$\because$ $p(x)$ is fixed ,

$\therefore$ it is only need to optimized $\int_y l(f(x), y)\, p(y|x)\, dy$

$$\frac{\partial \int |f(x)-y|\, p(y|x)\, dy}{\partial f(x)} = 0$$

$$\Downarrow$$

$$\int_{f(x)}^{+\infty} p(y|x)\, dx - \int_{-\infty}^{f(x)} p(y|x)\, dx = 0$$

$\therefore$ when $f(x)$ satisfy either

when can get the optimal for $E_{(x,y)}[l(f(x), y)]$

HW1_Coding

Question3

- result from hw1_q3.py
  - Ridge regression CV MSE values [0.48921793123166984, 0.43335910582310905, 0.8864386636073293, 0.3909161078107007, 0.7479735583197632, 0.5298021908758567, 0.2879844935594736, 0.7732653092966804, 0.6430556228608881, 0.32751024502502246]
  - Logistic Regression CV error rates [0.125, 0.05357142857142857, 0.05357142857142857, 0.07142857142857142, 0.05357142857142857, 0.03571428571428571, 0.03571428571428571, 0.05357142857142857, 0.14285714285714285, 0.03571428571428571]

Question4

- Lasso Regression

Lambda	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	Mean	Std
0.01	0.516951	0.517854	0.556182	0.535897	0.567506	0.490131	0.547233	0.535622	0.526535	0.515397	0.530931	0.0213998
0.1	0.604447	0.603998	0.644897	0.608878	0.632773	0.563871	0.621541	0.6127	0.571342	0.585196	0.604964	0.0243495
1	0.925764	0.976875	0.986779	0.940725	0.958792	0.908691	0.954913	1.00124	0.939356	0.904318	0.949745	0.0306286
10	1.31896	1.3709	1.36091	1.30816	1.33741	1.27513	1.34152	1.4063	1.33418	1.27315	1.33266	0.0392742
100	1.31976	1.3715	1.36136	1.30858	1.33746	1.27589	1.34256	1.40821	1.33488	1.27383	1.3334	0.039485

- Ridge Regression

Lambda	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	Mean	Std
0.01	0.512707	0.503256	0.539607	0.537389	0.560144	0.485528	0.542506	0.532516	0.557331	0.51151	0.52825	0.0229535
0.1	0.515214	0.504062	0.540332	0.540628	0.559718	0.485882	0.543561	0.533591	0.555269	0.512878	0.529114	0.0224973
1	0.545361	0.525577	0.562122	0.576299	0.57725	0.505432	0.566736	0.557084	0.563248	0.537765	0.551687	0.0219374
10	0.597616	0.569903	0.607613	0.635344	0.619906	0.547937	0.612766	0.603523	0.593804	0.585414	0.597383	0.0238399
100	0.61042	0.584774	0.624207	0.644904	0.632242	0.559796	0.624936	0.61461	0.591088	0.596389	0.608337	0.0241482

What do you notice as λ increases? Explain what is happening and why.

- When the λ increase, both Lasso and Ridge error increase. I think the penalty rate of model complex increased, therefore the model will become less generalized.

Which value of λ is optimal for each method?

- Best MSE for **ridge** with lambda 0.01: 0.5272739878977096
- Best MSE for **lasso** with lambda 0.01: 0.5345910890138215
- As shown in result, the ridge regression has better performance.

Question5

- LDA

Lambda	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	Mean	Std
-2	0.025	0.01875	0.0125	0	0.0125	0.01875	0.0125	0.00625	0.0125	0.03125	0.015	0.00847791
-1.5	0.01875	0.0125	0.0125	0.00625	0.01875	0.0125	0	0	0.00625	0.0375	0.0125	0.0104583
-1	0.0125	0.00625	0.0125	0.00625	0.01875	0.0125	0.00625	0	0.0125	0.03125	0.011875	0.008125
-0.5	0.0125	0.00625	0.00625	0.0125	0.01875	0.01875	0.01875	0.01875	0.025	0.0375	0.0175	0.00875
0	0.0125	0.0125	0.00625	0.0125	0.0375	0.03125	0.025	0.03125	0.03125	0.04375	0.024375	0.0120059
0.5	0.05	0.01875	0.0125	0.0125	0.05	0.05	0.03125	0.03125	0.05625	0.05625	0.036875	0.016875

What do you notice as λ increases? Explain what is happening and why.

- For the λ selection, I choose the range start from -2 to 0.5 with 0.5 between each step. I notice the error rate go down at first and reach the lowest when λ is equal to -1. Then the error rate goes up.
- The reason why is that we project dataset in to lower dimension, and when λ is equal to -1, can separate two class the most.

Which value of λ is optimal?

- the optimal value for lambda is -1 and loss value is: 0.005