Machine Learning & Data Mining
Caltech CS/CNS/EE 155
Homework 2

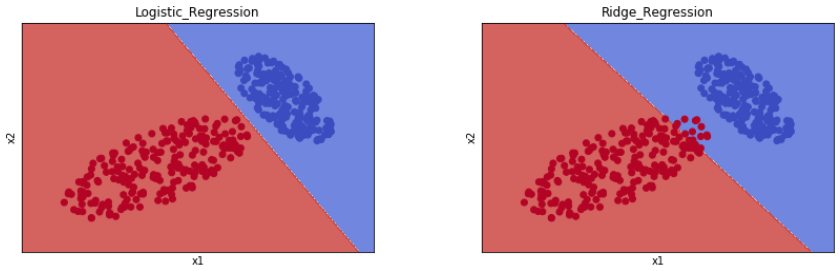
Changhao Xu
UID: 2103530
January 20th, 2020

---

1 Comparing Different Loss Functions

Question A: Squared loss is to minimize the distance between data points and model boundary. In this case, points that are far away from the classification boundary are heavily penalized regardless of whether they are misclassified or not.

Question B: Logistic regression behaves better than ridge regression here. Logistic regression uses log loss while ridge regression uses squared loss; as explained in Question A, squared loss are heavily affected by the points far away from the classification boundary and thus cannot classify correctly.



Question C:

$$\nabla_w L_{hinge} = \begin{cases} 0 & y\mathbf{w}^T\mathbf{x} \geq 1 \\ -y\mathbf{x} & y\mathbf{w}^T\mathbf{x} < 1 \end{cases}$$

$$\nabla_w L_{log} = \frac{-y\mathbf{x}}{e^{y\mathbf{w}^T\mathbf{x}} + 1}$$

(1/2, 3): $\nabla_w L_{hinge}$ = (-1, -1/2, -3), $\nabla_w L_{log}$ = (-0.37754, -0.18877, -1.13262)
(2, -2): $\nabla_w L_{hinge}$ = 0, $\nabla_w L_{log}$ = (-0.11920, -0.23841, 0.23841)
(-3, 1): $\nabla_w L_{hinge}$ = 0, $\nabla_w L_{log}$ = (0.04743, -0.14228, 0.04743)

Question D: Hinge loss gradient will converge to 0 when the point is correctly classified ($y\mathbf{w}^T\mathbf{x} \geq 1$); while log loss gradient will not converge to 0 unless $\mathbf{x} = 0$. For a linearly separable dataset, suppose a hyperplane with weight $\mathbf{w}$ that classify all points correctly. In order to eliminate training error, simply scale the weight vector $\mathbf{w}$ to a sufficiently large vector, so that $y\mathbf{w}^T\mathbf{x} \geq 1$ for all points $\mathbf{x}$ in the dataset.

Question E: If learning objective of SVM is to minimize just L_{hinge}, then any hyperplane with a properly scaled weight vector $\mathbf{w}$ can eliminate training error and result in 0 gradient. Therefore, in order for an SVM to be a "maximum margin" classifier, an additional penalty term $\lambda\|w\|^2$ must be added, where $\|w\|^2$ addresses the margin between the training points and the boundary.

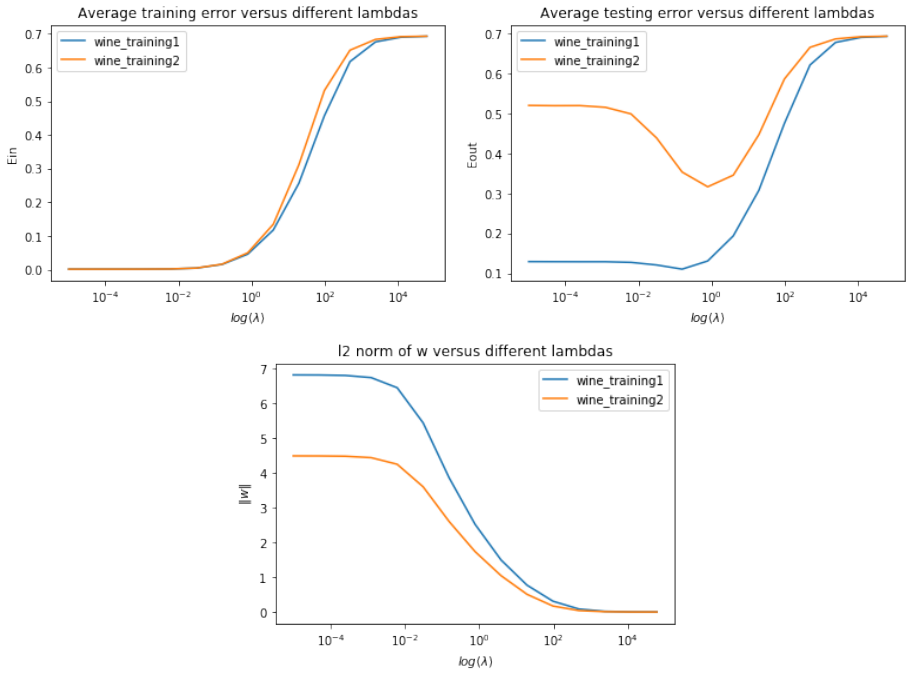
Changhao Xu
UID: 2103530

2 Effects of Regularization

Question A : Adding the penalty term will increase the training (in-sample) error, because training without regularization will minimize training error, while regularization term will limit this fitting performance. Adding a penalty term cannot always decrease the out-of-sample errors (this only happens when overfitting occurs). When the model is simple and underfitting occurs, adding regularization will increase out-of-sample error.

Question B: l_0 regularization is not continuous, and thus is hard to use methods like stochastic gradient descent to optimize.

Question C:



Question D : Training error with training1 is overall smaller than that of training2. Initially with little regularization, both have overfitting ($E_{in} = 0$), then with better regularization, training1 have more data points and thus perform better. Testing error with training1 is overall smaller than that of training2, and both experience a decrease and then increase of E_{out}. Initially with little regularization, both have overfittings (E_{out} is big compared with E_{in}), then with better regularization, both model have a decrease in E_{out}, while training1 have more data points and thus perform better. As λ becomes too big, both model suffer from underfitting, and have an increase in E_{out}.

Question E : When λ increase, training error always increase due to adding of penalty term. As for

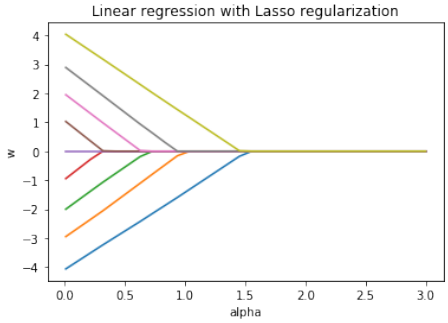
testing error, initially testing error is big due to overfitting, then with regularization the model performs better, so testing error decrease, but with too large λ the model suffer from underfitting, so testing error increase.

Question F: When λ increase, l_2 norm of **w** decrease because w have fewer coefficients as model becomes simpler.

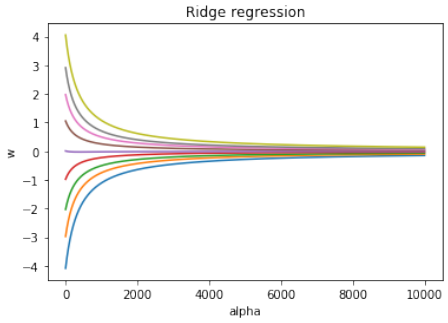
Question G: I would choose $\lambda = 0.78125$, which has smallest testing error.

3 Lasso (l_1) vs. Ridge (l_2) Regularization

Question A: i.



ii.



iii. As alpha increases, the weights in Lasso regression decrease linearly until they are close to 0, and then they reach exactly 0, so number of model weights that are exactly zero increases nearly linearly with the increase of alpha. Meanwhile, in Ridge regression, as alpha increases, the weights decrease exponentially

---

but the weights doesn't reach exactly 0.

Question B: i.

$$\begin{aligned}\nabla_w\Big[\|\mathbf{y}-\mathbf{x}w\|^2+\lambda\|w\|_1\Big] &= \nabla_w\Big[(\mathbf{y}^T-\mathbf{x}^Tw)(\mathbf{y}-\mathbf{x}w)+\lambda\|w\|_1\Big]\\ &= \nabla_w\Big[\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\mathbf{x}w-\mathbf{x}^T\mathbf{y}w+\mathbf{x}^T\mathbf{x}w^2+\lambda\|w\|_1\Big]\\ &= -\mathbf{y}^T\mathbf{x}-\mathbf{x}^T\mathbf{y}+2w\mathbf{x}^T\mathbf{x}+\lambda\nabla_w\|w\|_1\\ &= 0\end{aligned}$$

$$w=\frac{1}{2}(\mathbf{x}^T\mathbf{x})^{-1}\Big[\mathbf{y}^T\mathbf{x}+\mathbf{x}^T\mathbf{y}-\lambda\nabla_w\|w\|_1\Big], where\nabla_w\|w\|_1=\begin{cases}sgn(w) & w\neq 0\\ [-1,+1] & w=0\end{cases}$$

In this case, when w = 0, $w=\frac{1}{2}(\mathbf{x}^T\mathbf{x})^{-1}\Big[\mathbf{y}^T\mathbf{x}+\mathbf{x}^T\mathbf{y}-\lambda[-1,+1]\Big]=0$, which requires that $\mathbf{y}^T\mathbf{x}+\mathbf{x}^T\mathbf{y}-\lambda[-1,+1]=0$, i.e. $\mathbf{y}^T\mathbf{x}+\mathbf{x}^T\mathbf{y}\leq\lambda$. Therefore,

$$w=\begin{cases}\frac{1}{2}(\mathbf{x}^T\mathbf{x})^{-1}\Big[\mathbf{y}^T\mathbf{x}+\mathbf{x}^T\mathbf{y}-\lambda sgn(w)\Big] & \mathbf{y}^T\mathbf{x}+\mathbf{x}^T\mathbf{y}>\lambda\\ 0 & \mathbf{y}^T\mathbf{x}+\mathbf{x}^T\mathbf{y}\leq\lambda\end{cases}$$

ii. When $\lambda=0$, $\mathbf{y}^T\mathbf{x}+\mathbf{x}^T\mathbf{y}>\lambda$ and w = $\frac{1}{2}(\mathbf{x}^T\mathbf{x})^{-1}\Big[\mathbf{y}^T\mathbf{x}+\mathbf{x}^T\mathbf{y}\Big]\neq 0$.
Therefore, when w = 0, $\lambda_{min}=\mathbf{y}^T\mathbf{x}+\mathbf{x}^T\mathbf{y}$

iii. When w becomes **w** and Lasso becomes Ridge,

$$\begin{aligned}\nabla_{\mathbf{w}}\Big[\|\mathbf{y}-\mathbf{x}\mathbf{w}\|^2+\lambda\|\mathbf{w}\|_2^2\Big] &= \nabla_{\mathbf{w}}\Big[(\mathbf{y}^T-\mathbf{w}^T\mathbf{x}^T)(\mathbf{y}-\mathbf{x}\mathbf{w})+\lambda\mathbf{w}^T\mathbf{w}\Big]\\ &= \nabla_{\mathbf{w}}\Big[\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\mathbf{x}\mathbf{w}-\mathbf{w}^T\mathbf{x}^T\mathbf{y}+\mathbf{w}^T\mathbf{x}^T\mathbf{x}\mathbf{w}+\lambda\mathbf{w}^T\mathbf{w}\Big]\\ &= -\mathbf{y}^T\mathbf{x}-\mathbf{x}^T\mathbf{y}+2\mathbf{x}^T\mathbf{x}\mathbf{w}+2\lambda\mathbf{w}\\ &= 0\end{aligned}$$

$$w=\frac{1}{2}(\mathbf{x}^T\mathbf{x}+\lambda\mathbf{I})^{-1}(\mathbf{y}^T\mathbf{x}+\mathbf{x}^T\mathbf{y})$$

iv. When $\lambda=0$, w = $\frac{1}{2}(\mathbf{x}^T\mathbf{x})^{-1}\Big[\mathbf{y}^T\mathbf{x}+\mathbf{x}^T\mathbf{y}\Big]\neq 0$.
Therefore, since $\lambda>0$, there doesn't exist such λ such that w = 0.