CSE 446 HW2

Rohan Mukherjee

May 1, 2024

A1.

a. In the L1 norm, small numbers are punished more than in the L2 norm. For example, in the L2 norm a coordinate of 0.1 translates to an increase in only 0.01, which is almost negligible. But in the L1 norm it still has a lot of weight.

b. This is true, since we might step past the minimum with too large a step size, and then we will try to step back towards it, but just end up oscillating far from the minimum.

c. SGD runs much, much faster than GD, bringing the running time from $O(n \cdot d)$ per iteration to only $O(d)$. GD on the other hand is much more stable, not relying on randomness and we don't have to worry about changing the step size in the process.

d. Logistic regression does not have a closed form solution, while linear regression does.

A2.

a. Since $|ax| = |a||x|$ holds for all $a, x \in \mathbb{R}$, we must have $\sum_{i=1}^n |ax_i| = \sum_{i=1}^n |a||x_i| = |a| \sum_{i=1}^n |x_i|$. Also, $|x| \geq 0$ for every $x \in \mathbb{R}$, so $\sum_{i=1}^n |x_i| \geq 0$. Finally, notice that

$$(|a+b|)^2 = (a+b)^2 = a^2 + b^2 + 2ab \leq a^2 + b^2 + 2|a||b| = (|a| + |b|)^2$$

Since $f(x) = \sqrt{x}$ is increasing, and both the LHS and RHS are positive, taking square roots on both sides shows that $|a+b| \leq |a| + |b|$. Using this repeatedly yields:

$$\sum_{i=1}^n |x_i + y_i| \leq \sum_{i=1}^n |x_i| + |y_i| = \sum_{i=1}^n |x_i| + \sum_{i=1}^n |y_i|$$

b. Consider the vectors $x = (1/4, 0)$ and $y = (0, 1/4)$. Then,

$$g(x+y) = \left(\frac{1}{2} + \frac{1}{2}\right)^2 = 1$$

while,

$$g(x) + g(y) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

Which shows that the triangle inequality does not hold.

A3.

Note that the set of points $\{\lambda a + (1-\lambda)b \mid 0 \leq \lambda \leq 1\} = \{a + \lambda(b-a) \mid 0 \leq \lambda \leq 1\}$ is precisely the lien conneting a with b. Since in part I, the line connecting the point b with c does not live fully within the set, we see that I is not convex. Similarly, applying the same logic to points a and d in the second picture shows that the II is not convex either.

A4.

The first function I is convex because it is a quadratic function pointing upwards, with always-positive second derivative. The second function is not convex because The secant line between b and c lies below the curve, not above it (Recall that the line $\lambda f(x) + (1-\lambda)f(y)$ for $0 \leq \lambda \leq 1$ is the secant line of f between x and y).

A5.

(a/b). Here are my graphs, where nnz(w) is the number of non-zero elements in w:

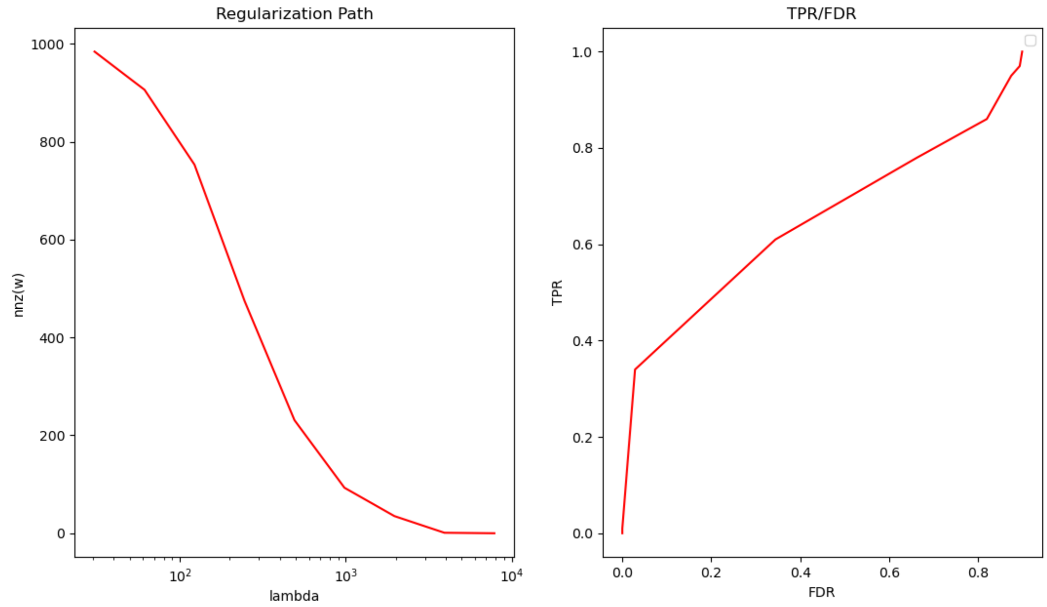

Figure 1: nnz(w) as a function of λ and FDR vs TPR

c. We can see on the left that increasing λ decreases the number of nonzero entries in w, as expected. On the right, recall that λ_{max} was chosen to have all zero entries, equivalently having a true positive rate of 0. Thus as λ shrinks, the true positive rate and the false positive rate both increase. I would guess that the sweet spot is where the FDR is really small and the TPR is reasonably large, equivalently when the slope is really high. This occurs around FDR = 0.05. Thus there is a trade off with λ–smaller λ makes the number of nonzeros much smaller while at the same time making the false positive rate and true positive rate higher.

A6.

a. PCTNotHSGrad, PCTLargHouseOccup, and PCTUsePubTrans In a lot of less wealthy communities, the education system is either a lot worse due to little investment, so due to government policies like not having enough tax dollars to invest in schools the graduation rate can be lower. Similarly, wealther communities tend to have less single-family housing, and even some policies that prevent more single family housing, so the percentange of large house occupancy can go down. Finally, the percentage of people using public transportation is entirely dependent on the amount of investment the government has put into public transportation, so that varies a lot depending on policy.

b. Three features that might be interpreted as reasons for higher levels of violent crime are the PctUnemployed, pctNotHSGrad, and LemasSwornFT. When working at a store is more dangerous less people will want to, making PctUnemployed go up, violent crime in the neighborhood can prevent kids from going to school because they want to stay safe, and finally, police offers are not commiting crimes but are instead are trying to stop them, so show up near crime a lot, so a higher LemasSwornFT could be misinterpreted as a cause of violent crime.

c. The nonzero weights as a function of λ can be seen in the following graph:

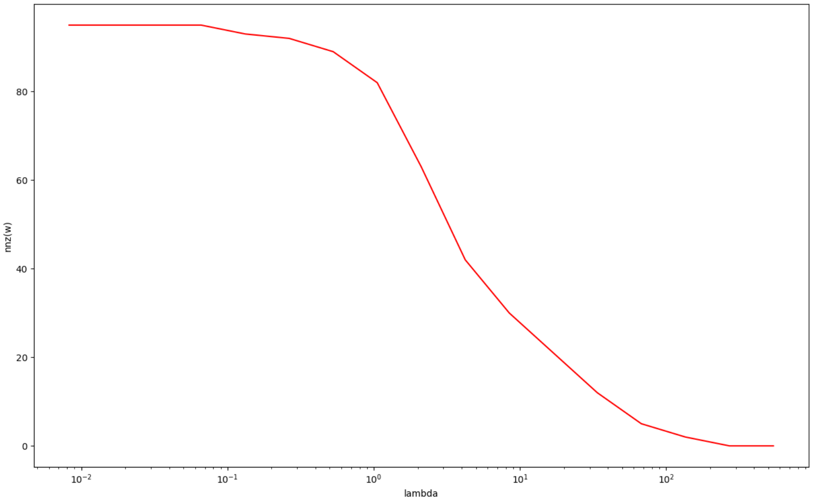

Figure 2: Nonzero weights as a function of λ

d. The regularization paths for the coefficients of those variables is:

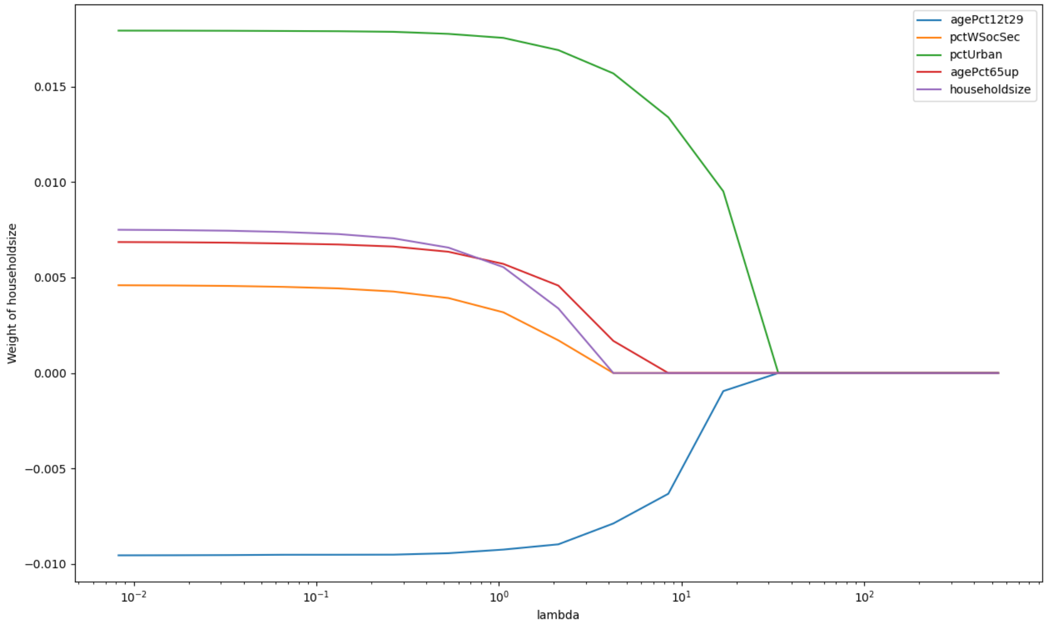

Figure 3: Regularization paths for the coefficients of the variables

e. Here is the MSE on the training and test data:

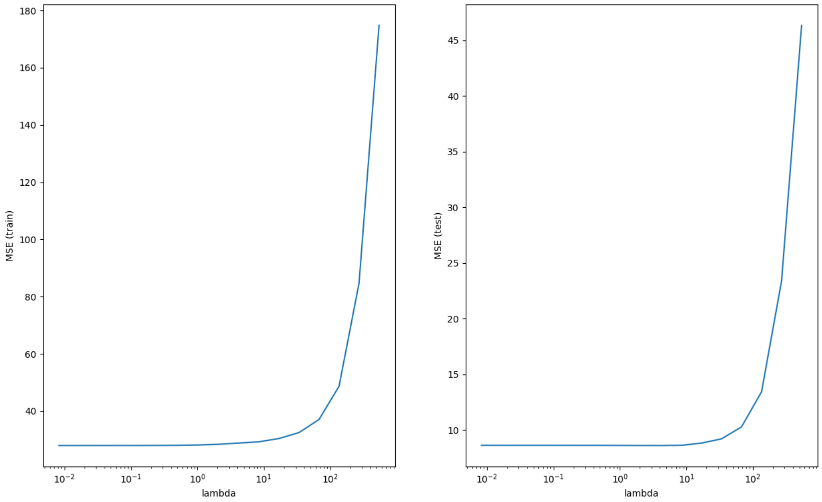

Figure 4: MSE on the training and test data

f. The most positive feature was PctIlleg and the most negative was PctKids2Par. I briefly mentioned this above when I was discussing the ethics, but this means that, under this model, the percentage of kids born to unmarried parents has the biggest correlation with violent crime, and the percentage of kids with 2 parents has the least correlation with violent crime.

g. This is a correlation ≠ causation mistake. Fire trucks are often seen near burning buildings because they help clear fire. Fire trucks are extremely correlated with fire but instead of starting fires they put them out. Similarly, the agePct65up variable is not causing the crime, but instead could be a result of lower crime, such as less violent crime means less people are victims of murder, so more people will live longer, thus agePct65up goes up. In this manner agePct65up is correlated with less violent crime but is not the cause.

A7.

a. Notice that

$$\frac{\mathrm{d}}{\mathrm{d}x}\log(1+e^{-x}) = \frac{-e^{-x}}{1+e^{-x}} = \frac{1}{1+e^{-x}} - 1$$

We recall the chain rule: if $f : \mathbb{R} \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}$ are differentiable, then $\nabla_x(f \circ g)(x) = f'(g(x))\nabla_x g(x)$. Taking $f(x) = \log(1+e^{-x})$ and $g(x) = y_i(b + x_i^T w)$, we get that

$$\nabla_w \log\Big(1 + \exp\big(-y_i(b + x_i^T w)\big)\Big) = \nabla_w f(g(x)) = \left(\frac{1}{1+e^{-y_i(b+x_i^T w)}} - 1\right) y_i x_i = (\mu_i(w,b) - 1) y_i x_i$$

Because $\nabla_w y_i(b + x_i^T w) = \nabla_w y_i x_i^T w = y_i x_i$. Similarly,

$$\frac{\partial}{\partial b} f(g(x)) = \left(\frac{1}{1+e^{-y_i(b+x_i^T w)}} - 1\right) y_i = (\mu_i(w,b) - 1) y_i$$

Since $\nabla_w \lambda w^T w = \lambda 2w$, we have that

$$\nabla_w J(w,b) = \frac{1}{n}\sum_{i=1}^{n} (\mu_i(w,b) - 1) y_i x_i + 2\lambda w$$
$$\frac{\partial}{\partial b} J(w,b) = \frac{1}{n}\sum_{i=1}^{n} (\mu_i(w,b) - 1) y_i$$

b. (i) Here is the plot of $J(w,b)$ as a function of the iteration number:

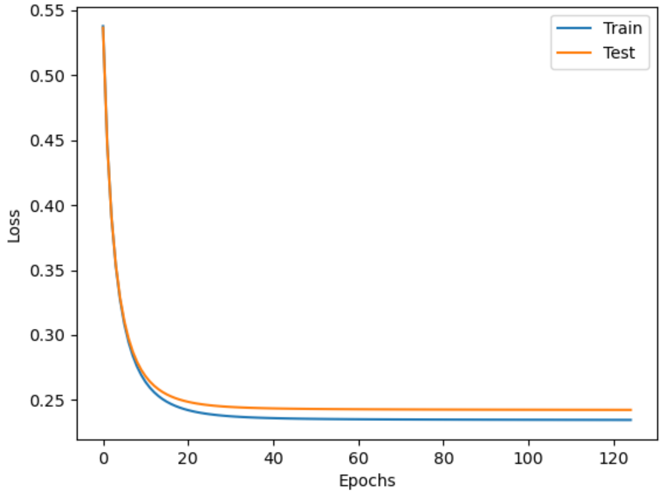

Figure 5: Loss as a function of the iteration number

(ii) Here is the plot of the miscalculation error as a function of the iteration number:

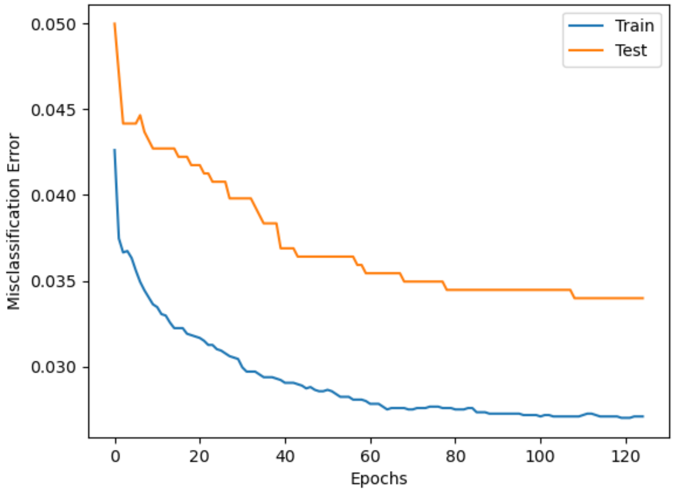

Figure 6: Misclassification Error as a function of the iteration number

c. (i) Loss for batch size 100:

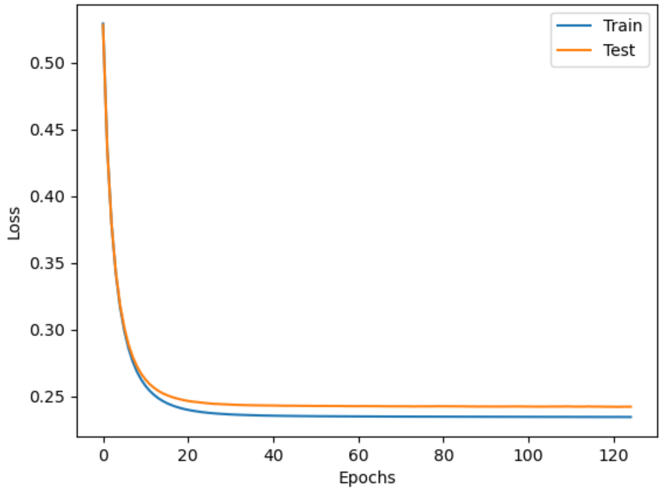

Figure 7: Loss as a function of the iteration number

(ii) Misclassification error for batch size 100:

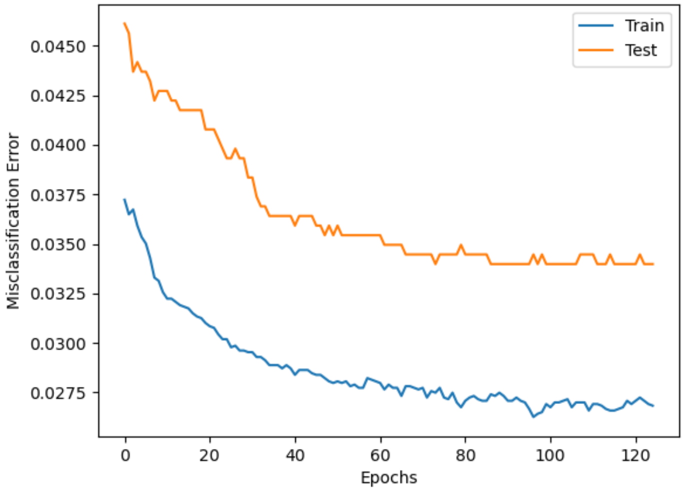

Figure 8: Misclassification Error as a function of the iteration number

d. (i) Loss for batch size 1:

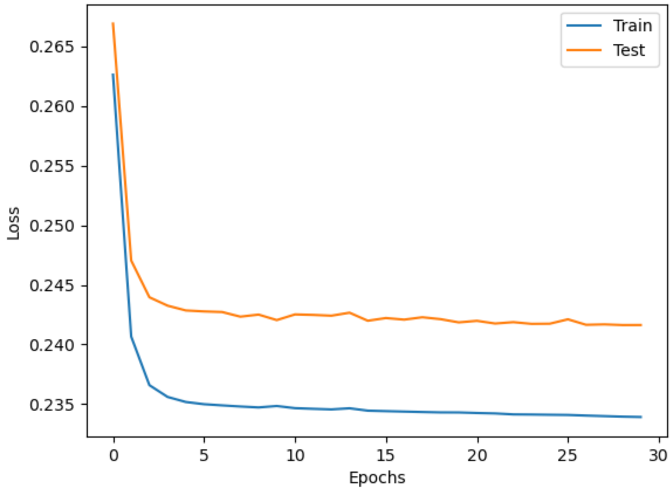

Figure 9: Loss as a function of the iteration number

(ii) Misclassification error for batch size 1:

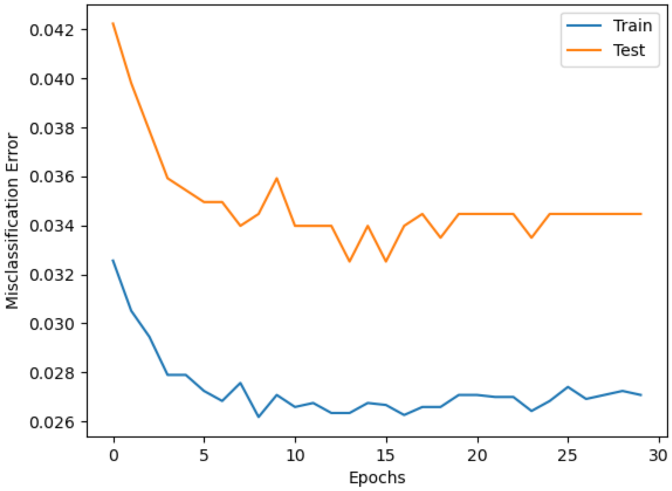

Figure 10: Misclassification Error as a function of the iteration number

A8.

1. This homework took me around 12-15 hours to complete.