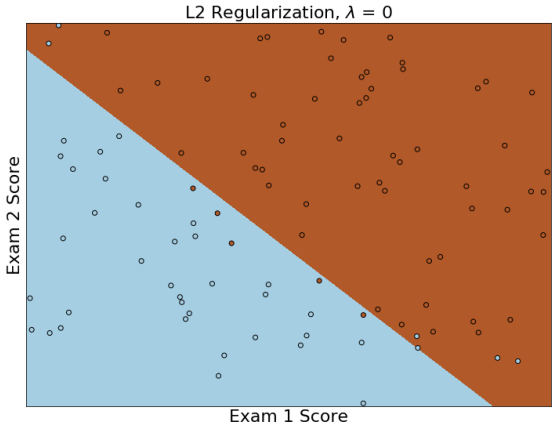
CIS 419/519: Homework 3

Jiatong Sun

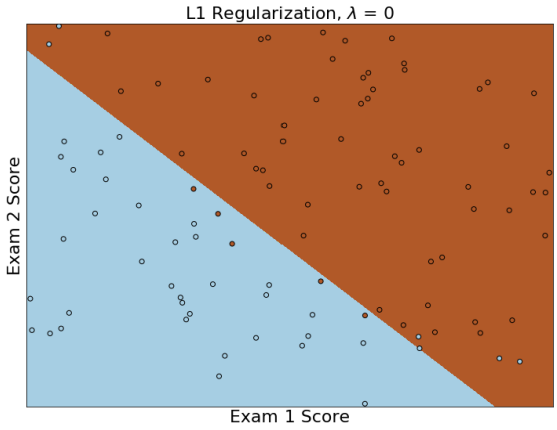
Although the solutions are entirely my own, I consulted with the following people and sources while working on this homework: $JunfanPan$, $ZhuoyuHe$, $YuchenSun$, $YupengLi$, $JingZhao$, $TianjiaZhu$ $https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html$

1 Logistic Regression

1.3 Analysis

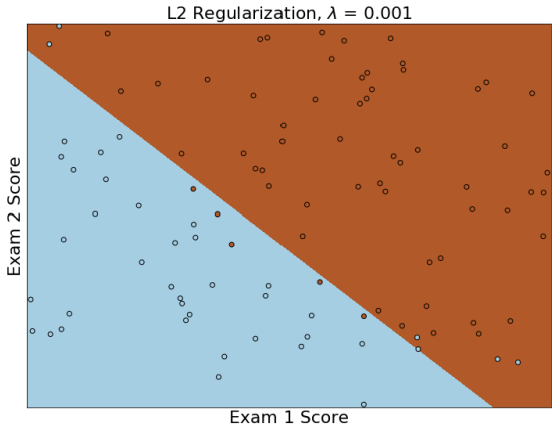


(a) $\lambda = 0, regNorm = 2$

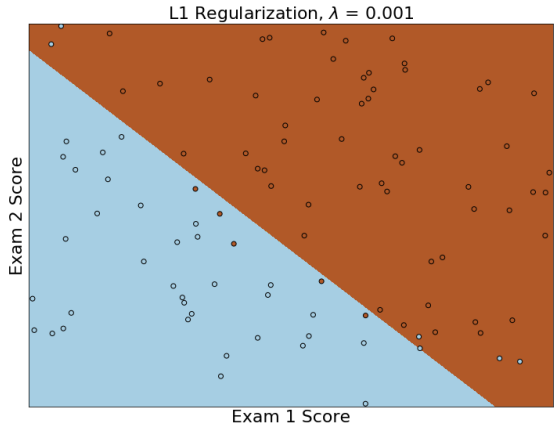


(b) $\lambda = 0, regNorm = 1$

Figure 1: $\lambda = 0$

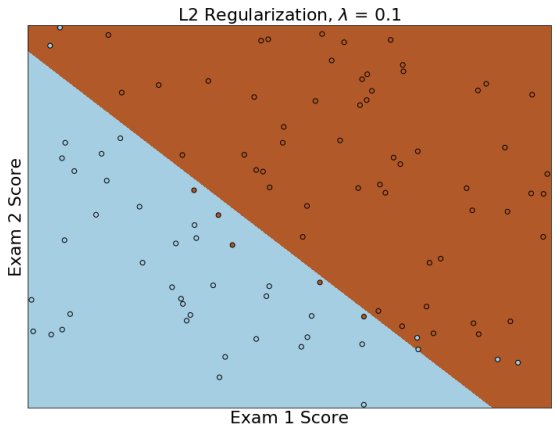


(a) $\lambda = 0.001, regNorm = 2$

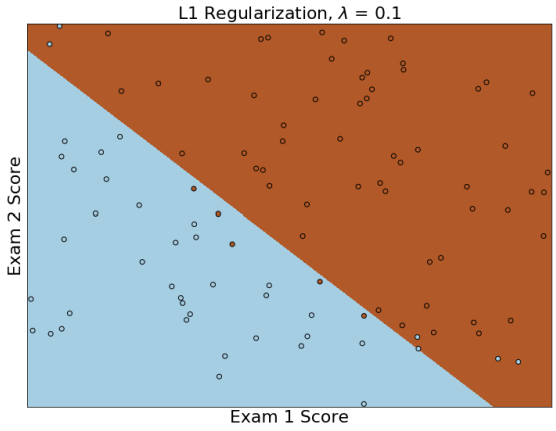


(b) $\lambda = 0.001, regNorm = 1$

Figure 2: $\lambda = 0.001$

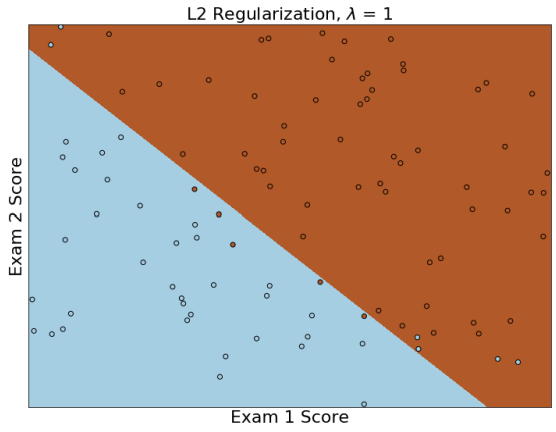


(a) $\lambda = 0.1, regNorm = 2$

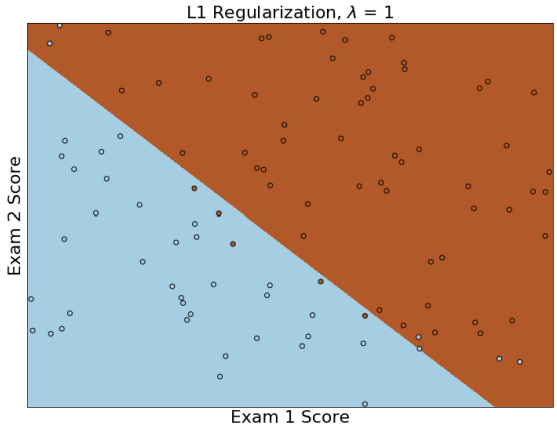


(b) $\lambda = 0.1, regNorm = 1$

Figure 3: $\lambda = 0.1$

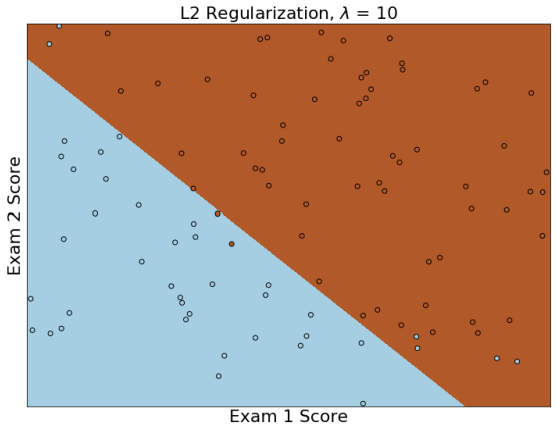


(a) $\lambda = 1, regNorm = 2$

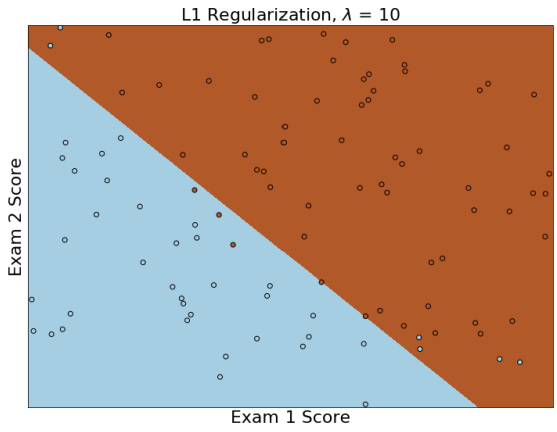


(b) $\lambda = 1, regNorm = 1$

Figure 4: $\lambda = 1$



(a) $\lambda = 10, regNorm = 2$



(b) $\lambda = 10, regNorm = 1$

Figure 5: $\lambda = 10$

Conclusion:
As the λ increases, the border slightly moves to bottom left corner, which means more data are recognized as 1. This is because a bigger λ causes the regularization term to be bigger and the convergence to be more quick, and thus the boarder stops moving at a relatively small iteration number. For l1 and L2, when λ is small, the difference is hard to observe; but when λ is greater than 1, L2 norm converges more quickly than L1 norm and thus its boarder is more close to the bottom left corner.

2 Comparing Algorithms

2.2 Comparing Algorithms

Data Set Description				
Data Set	**number of features**	**number of instances**	**missing data**	**feature properties**
wdbc	30	569	No	Computed from a digitized image of a fine needle aspirate (FNA) of a breast mass
retinopathy	19	1151	No	Represent either a detected lesion, a descriptive feature of a anatomical part or an image-level descriptor
diabetes	8	768	No	Predict whether or not a patient has diabetes based on certain diagnostic measurements

Setup:
All of the following learning processes use those parameters:
$\alpha = 0.001$, $maxNumIters = 10000$, $\epsilon = 0.0001$, 3 trials, 5 folders

Result:

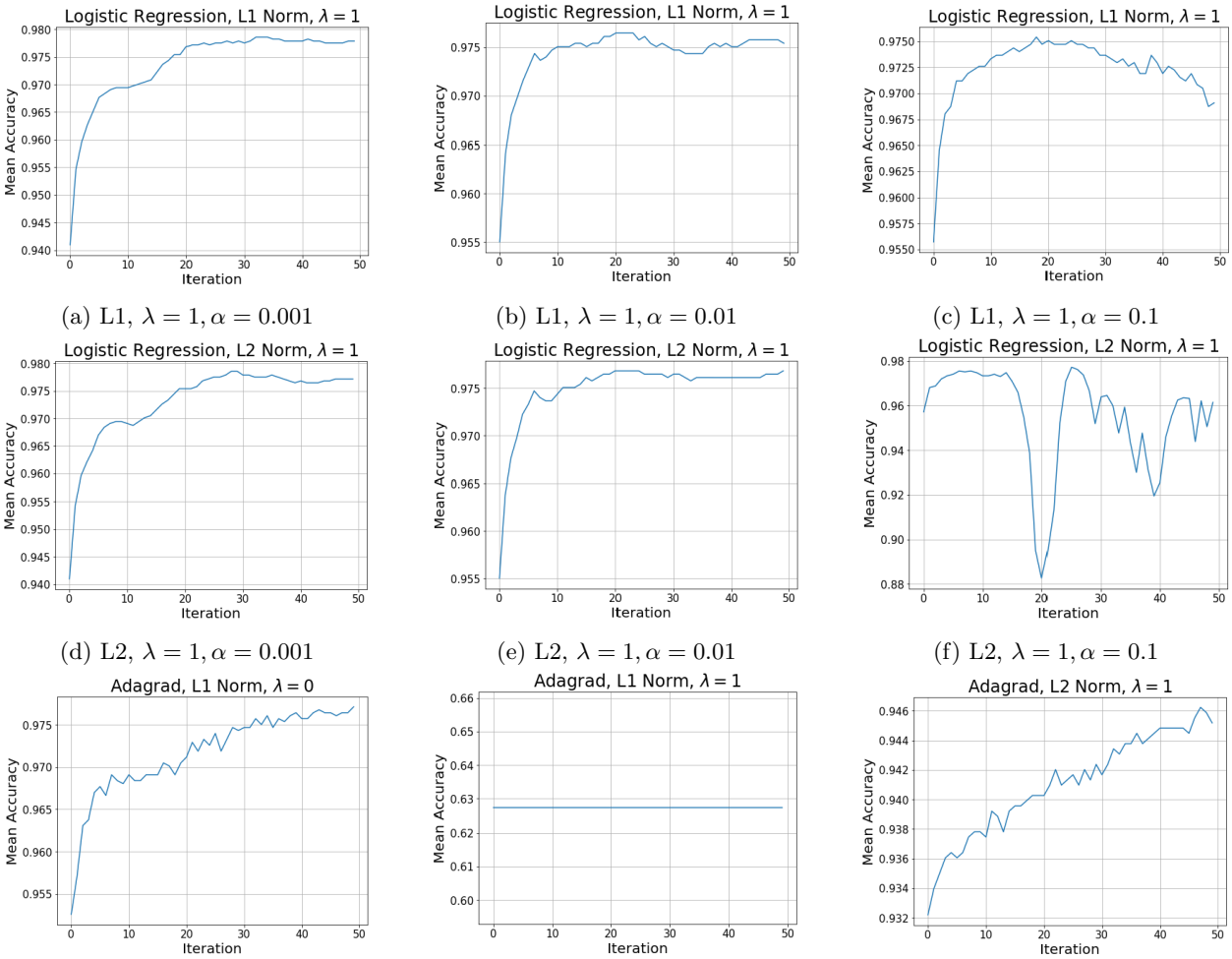
wdbc data set		
	$\lambda = 1$	$\lambda = 10$
Logistic Regression L1	0.9701	0.9701
Logistic Regression L2	0.9748	0.9736
Adagrad L1	0.6274	0.6290
Adagrad L2	0.9467	0.8594

retinopathy data set		
	$\lambda = 1$	$\lambda = 10$
Logistic Regression L1	0.7324	0.6904
Logistic Regression L2	0.7261	0.6901
Adagrad L1	0.5308	0.5308
Adagrad L2	0.5995	0.5946

diabetes data set		
	$\lambda = 1$	$\lambda = 10$
Logistic Regression L1	0.7704	0.7687
Logistic Regression L2	0.7717	0.7682
Adagrad L1	0.6510	0.6510
Adagrad L2	0.7214	0.6510

2.3 Understanding Regularization and Adagrads

Setup: All of the following learning curves are generated on "wdbc.csv"



(a) L1, $\lambda = 1, \alpha = 0.001$ (b) L1, $\lambda = 1, \alpha = 0.01$ (c) L1, $\lambda = 1, \alpha = 0.1$

(d) L2, $\lambda = 1, \alpha = 0.001$ (e) L2, $\lambda = 1, \alpha = 0.01$ (f) L2, $\lambda = 1, \alpha = 0.1$

(g) L1, $\lambda = 0, \alpha = 0.01$ (h) L1, $\lambda = 1, \alpha = 0.01$ (i) L2, $\lambda = 1, \alpha = 0.001$

Figure 6: Learning Curves

Conclusion:

- From the first two rows of graphs, we can discover the influence of the learning rate α. When the learning rate goes bigger, learning process becomes faster, but a too big α may result in an overshooting, and thus make the mean accuracy to drop down at some point.
- From 6g and 6h, we can discover the influence of regularization. When $\lambda = 0$, the learning accuracy can go high continuously. But for $\lambda = 1$, the learning performance will remain a low level. This is because the regularization parameter is set too big and prevent over-fitting excessively so the learning can not happen.
- From the first two rows of graphs, we can discover the difference between L1 norm and L2 norm. When the learning rate α goes bigger, L2 norm is more sensitive than L1 norm, which is because the regularization term takes more ratio in L2 norm than in L1 norm.
- From the first two rows and the last row, we can discover the difference between a fixed learning rate and adagrad. When using a fixed learning rate, the learning curve is much more smooth,

while using adagrad makes the curve have many spikes. This is because adagrad adjusts gradient according to the frequency of feature properties and only goes a large step when the current feature has a high frequency. So there is more oscillation in the learning process in adagrad than with a fixed learning rate.