# CS 688: Homework 2 Solutions

February 26, 2024

**Note:** These are not intended to be comprehensive, just to help you see what the answers should be.

1 $E_{\text{in}}$ (the cross-entropy error) and the classification error (0-1 loss) for the non-normalized data are reported below:

| Algorithm | $E_{\text{in}}$ | Classification Error | |
|---|---|---|---|
| | | Training | Testing |
| GD 10000 | .5847 | .3092 | .3172 |
| GD 100000 | .4937 | .2237 | .2069 |
| GD 1000000 | .4354 | .1513 | .1310 |

Main takeaways: the generalization performance is excellent, in the sense that testing error is very close to training error, so the model is definitely not overfitting. The performance improves significantly with lower $E_{\text{in}}$, which makes sense given that the model generalizes well. Therefore, the benefits of optimizing the model for the training set cross-entropy loss are good in this case.

Note that test error is lower than training error in some cases. This is OK – it means the model is generalizing well, which you would expect sometimes for a simple model. In those cases, training error and test error will be close, and it's random whether one is better or worse.

When using normalized data, you should achieve the optimal cross-entropy loss ($E_{\text{in}}$) very fast (n fact, in time comparable to a standard package for some settings of $\eta$) the learning rate). Here are some examples of the number of iterations needed (doesn't terminate by 10000 much beyond 7).

| $\eta$ | $E_{\text{in}}$ | # iterations |
|---|---|---|
| 0.01 | 0.4074 | 23368 |
| 0.1 | 0.4074 | 2333 |
| 1 | 0.4074 | 230 |
| 4 | 0.4074 | 54 |
| 6 | 0.4074 | 32 |
| 7 | 0.4074 | 43 |

2 When $y_i \mathbf{w}^T \mathbf{x}_i \geq 1$ the term inside the parentheses is 0, so the gradient is 0 and there will be no update. If $y_i \mathbf{w}^T \mathbf{x}_i < 1$ then $\nabla_i \mathbf{w} = 2(1 - y_i \mathbf{w}^T \mathbf{x}_i)(-y_i \mathbf{x}_i) = 2\mathbf{x}_i(\mathbf{w}^T \mathbf{x}_i - y_i)$. (The last step is just simplifying a bit). Thus the gradient descent update rule becomes $\mathbf{w} = \mathbf{w} - \eta(\mathbf{w}^T \mathbf{x}_i - y_i)\mathbf{x}_i$.

3 There are many reasonable answers to this question. One thought is that the calibration metric is perhaps more fair on an *individual* level, because given your risk group it doesn't matter if you are black or white, you are equally likely to recidivate. Therefore, if decisions are made by judges on the basis of risk group, this could be thought of as more fair to the individual. One counterargument here is that, even though this may be true, judges using this method will unfairly keep many more black people in jail than white people, and since being in jail is bad for them, this is having bad overall effects on the black community at large, which may become entrenched over time, hurting that community significantly more than the white community.