1.1

Using Baye’s Rule:

Simplifying by :

Simplifying by :

can be cancelled in the denominator, and

Combining fractions

Changing into a form that will allow it o be put inside the exponential

Isolating x values

Converting to desired form (µ signs are switched to accommodate).

1.2

Finding

Finding

Subbing in values of

Taking ln and negating

Simplifying using the log rule that

Simplifying using the log rule that

Simplifying by grouping brackets

Using chain rule

1.3

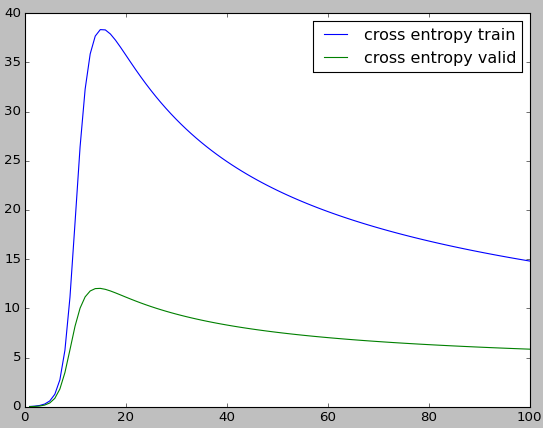
2.1

The training classification rates for 1, 3, 5, 7, and 9 are 0.82, .086, 0.86, 0.86, and 0.84, respectively. I would choose 5 as the best value of k, as it’s tied for the highest classification rate, and is in the middle so it is farther form the dropoff at both ends. K+2 (7) and K-2 (3) both have the same rate of 0.86. The test classification rate for all of these three is higher, at .92 for 3 and .94 for 5 and 7.

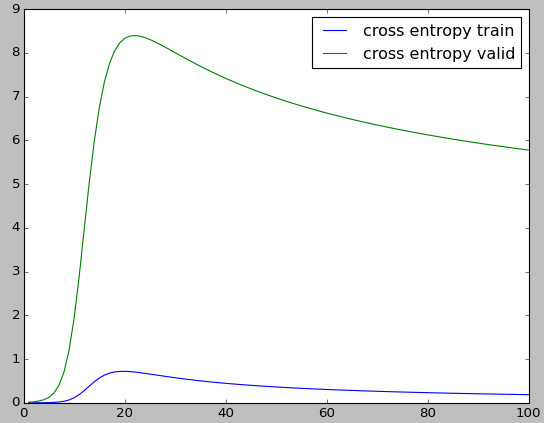
2.2

The hyperparameters which I determined to be best were a learning rate of 0.05, and 100 iterations. The final cross entropy and classification error for train, valid and test, and mnist\_train and mnist\_trian\_small respectively, are: mnist\_train: (train cross entropy: 14.811450, train classification error: 0.0373, valid cross entropy: 5.856860, valid classification error: 0.0980, test cross entropy: 4.85316136724, test classification error: 0.0785), mnist\_train\_small: (train cross entropy: 0.186111, train classification error: 0.0909, valid cross entropy: 5.778267, valid classification error: 0.3529, test cross entropy: 4.16516872097, test classification error: 0.2942)

mnist\_train:



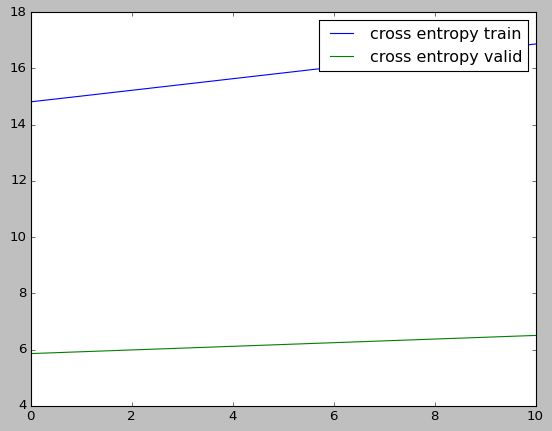
mnist\_train\_small:

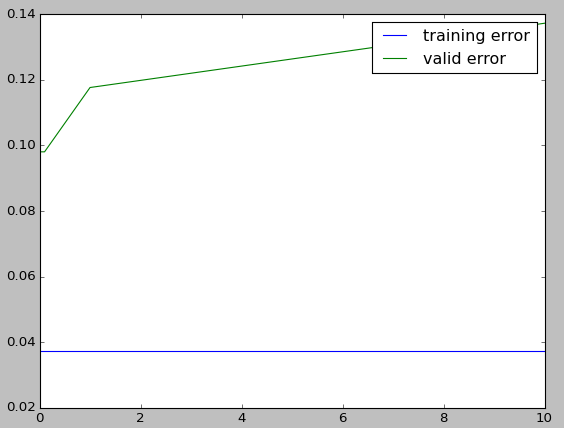


The results don’t change on different runs because I initialize my weights to set, relatively low numbers (0.1). I chose static weights rather than random weights because they reliably provided better classification rates than any randomly generated weights, and generally with fewer iterations, too.

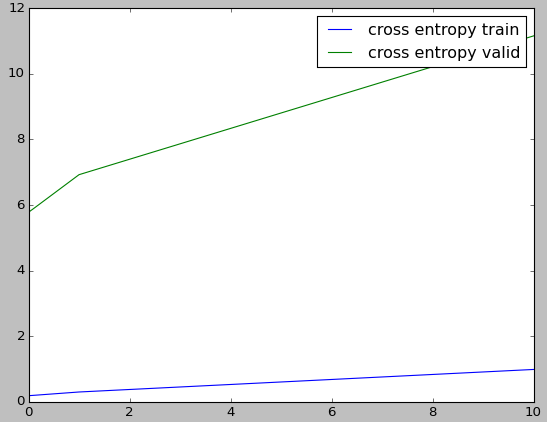
2.3

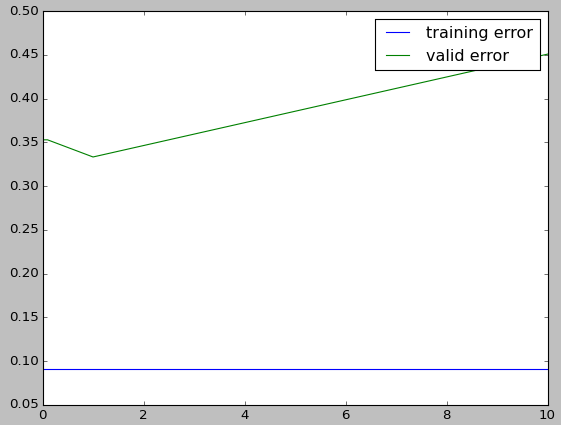
mnist\_train:





mnist\_train\_small:



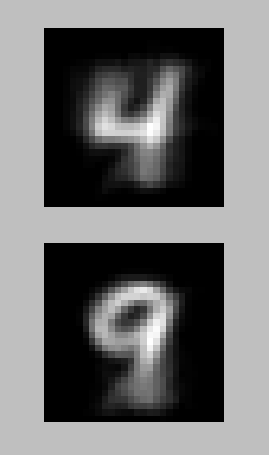


Cross entropy and classification error increase as λ increases, with one exception in valid\_error when using the small training set. This is probably because the model is overfitting, and the values of λ to the left of the minimum are too small to prevent the overfitting, whereas the values to the right hinder the fitting in the first place. So, the best value of λ is probably based on the size of your dataset; as can be seen in the graphs above, valid error decreases to the left of λ = 1 in the large training set, but increases in the small dataset; therefore, there is not a catch-all value that is the best, but in this case, I would have to say that the better value is the one that works for larger datasets, meaning λ = 0.01. The test error for the large dataset is 0.07843137254901966 and 0.29411764705882343 for the small set.

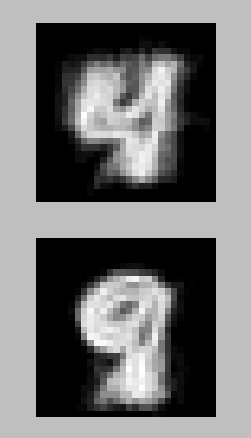
2.4

The train accuracy is 0.8625, and the test accuracy is 0.8

Mean:



Var:



The bright spots in the mean and the dark spots in the var indicate where the most reliability is. The var is quite fuzzy overall, indicating that there is a lot of uncertainty.

2.5

Logistic regression has the best classification rate, followed by k-NN, followed by naïve Bayes.