CSC467 HW2

#1

$$p(x|\pi,\mu,\Sigma) = \sum_{j=1}^{k} \pi_j N(x|\mu_j,\Sigma)$$

E: $\gamma(z_k) = p(z_k=1|x) = \frac{p(z_k=1)\,p(x|z_k=1)}{p(x)}$

$$= \frac{\pi_k N(x|\mu_k,\Sigma)}{\sum_{j=1}^{k}\pi_j N(x|\mu_j,\Sigma)}$$

M: Want to have $\frac{\partial \ln p(x|\pi,\mu,\Sigma)}{\partial \pi_k} = \frac{\partial \ln P(x|\pi,\mu,\Sigma)}{\partial \mu_k} = \frac{\partial \ln P(x|\pi,\mu,\Sigma)}{\partial \Sigma} = 0$

$$\frac{\partial \ln P(x|\pi,\mu,\Sigma)}{\partial \mu_k} = \sum_{n=1}^{N} \frac{1}{\sum_{j=1}^{k}\pi_j N(x^{(n)}|\mu_j,\Sigma)} \cdot \frac{\partial}{\partial \mu_k}\pi_k N(x^{(n)}|\mu_k,\Sigma)$$

$$= \sum_{n=1}^{N} \frac{1}{\sum_{j=1}^{k}\pi_j N(x^{(n)}|\mu_j,\Sigma)}\,\pi_k \frac{\partial}{\partial \mu_k}\frac{1}{(2\pi)^{1/2}|\Sigma|^{1/2}}\exp\left(-\tfrac{1}{2}(x^{(n)}-\mu_k)^T\Sigma^{-1}(x^{(n)}-\mu_k)\right)$$

$$= \sum_{n=1}^{N} \frac{1}{\sum_{j=1}^{k}\pi_j N(x^{(n)}|\mu_j,\Sigma)}\cdot\pi_k \frac{\partial}{\partial \mu_k}\frac{1}{2\pi|\Sigma|^{1/2}}\exp\left(-\tfrac{1}{2}(x^{(n)}-\mu_k)^T\Sigma^{-1}(x^{(n)}-\mu_k)\right)$$

$$= \sum_{n=1}^{N} \frac{\pi_k \frac{1}{2\pi|\Sigma|^{1/2}}}{\sum_{j=1}^{k}\pi_j N(x^{(n)}|\mu_j,\Sigma)}\exp\left(-\tfrac{1}{2}(x^{(n)}-\mu_k)^T\Sigma^{-1}(x^{(n)}-\mu_k)\right)\frac{\partial}{\partial \mu_k}\left[-\tfrac{1}{2}(x^{(n)}-\mu_k)^T\Sigma^{-1}(x^{(n)}-\mu_k)\right]$$

$$= \sum_{n=1}^{N} \frac{\pi_k \frac{1}{2\pi|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x^{(n)}-\mu_k)^T\Sigma^{-1}(x^{(n)}-\mu_k)\right)}{\sum_{j=1}^{k}\pi_j N(x^{(n)}|\mu_j,\Sigma)}\left[-\frac{1}{2}\left(\frac{\partial}{\partial \mu_k}(x^{(n)}-\mu_k)^T\Sigma^{-1}(x^{(n)}-\mu_k)\right)\right]$$

$$= \sum_{n=1}^{N} \underbrace{\frac{\pi_k N(x^{(n)}|\mu_k,\Sigma)}{\sum_{j=1}^{k}\pi_j N(x^{(n)}|\mu_j,\Sigma)}}_{\gamma(z_k^{(n)})}\cdot(x^{(n)}-\mu_k)^T\Sigma^{-1}$$

proof:
A is symmetric
$\Rightarrow A = A^T$
$AA^{-1} = I$
$(A^{-1})^T A^T = I$
$(A^{-1})^T A = I$
$\Rightarrow (A^{-1})^T = A^{-1}$ ∎

TAKE transpose on both sides, using the property $(A+B)^T = A^T + B^T$

$$= \sum_{n=1}^{N}\gamma(z_k^{(n)})(\Sigma^{-1})^T(x^{(n)}-\mu_k),$$

(a) Σ is symmetric b/c $\Sigma(i,j) = \Sigma(j,i)\ \forall i,j$, $(\Sigma^{-1})^T = \Sigma^{-1}$

$$0 = \sum_{n=1}^{N}\gamma(z_k^{(n)})\Sigma^{-1}(x^{(n)}-\mu_k)$$

$$0 = \sum_{n=1}^{N} \gamma(z_k^{(n)}) \Sigma^{-1} (x^{(n)} - \mu_k)$$

$$= \sum_{n=1}^{N} \gamma(z_k^{(n)}) \Sigma^{-1} x^{(n)} - \sum_{n=1}^{N} \gamma(z_k^{(n)}) \Sigma^{-1} \mu_k$$

$$\sum_{n=1}^{N} \gamma(z_k^{(n)}) \Sigma^{-1} \mu_k = \sum_{n=1}^{N} \gamma(z_k^{(n)}) \Sigma^{-1} x^{(n)}$$

$$\sum_{n=1}^{N} \gamma(z_k^{(n)}) \mu_k = \sum_{n=1}^{N} \gamma(z_k^{(n)}) x^{(n)}$$

$$\mu_k = \frac{\sum_{n=1}^{N} \gamma(z_k^{(n)}) x^{(n)}}{\sum_{n=1}^{N} \gamma(z_k^{(n)})}$$

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_k^{(n)}) x^{(n)}, \quad N_k = \sum_{n=1}^{N} \gamma(z_k^{(n)})$$

$$\frac{\partial \ln p(x|\pi,\mu,\Sigma)}{\partial \pi_k} \neq 0 \quad \text{with constraint } \sum \pi_k = 1$$

⇒ use lagrange multiplier

$$\frac{\partial \ln p(x|\pi,\mu,\Sigma)}{\partial \pi_k} = 0 \quad \frac{\partial}{\partial \pi_k} \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k N(x^{(n)}|\mu_k,\Sigma) + \lambda_1 (1 - \sum_k \pi_k)$$

$$= \sum_{n=1}^{N} \frac{\partial}{\partial \pi_k} \ln \sum_{k=1}^{K} \pi_k N(x^{(n)}|\mu_k,\Sigma) + \lambda_1 (1 - \sum_k \pi_k)$$

$$0 = \sum_{n=1}^{N} \frac{N(x^{(n)}|\mu_k,\Sigma)}{\sum_{k=1}^{K} \pi_k N(x^{(n)}|\mu_k,\Sigma)} - \lambda_1$$

$$0 = \sum_{n=1}^{N} \frac{\pi_k N(x^{(n)}|\mu_k,\Sigma)}{\sum_{k=1}^{K} \pi_k N(x^{(n)}|\mu_k,\Sigma)} - \pi_k \lambda_1$$

$$\pi_k \lambda_1 = \sum_{n=1}^{N} \gamma(z_k^{(n)}) = N_k$$

$$\sum_{k=1}^{K} \pi_k \lambda_1 = \sum_{k=1}^{K} \sum_{n=1}^{N} \gamma(z_k^{(n)})$$

$$\lambda_1 = N$$

$$\Rightarrow \pi_k = \frac{N_k}{\lambda_1} = \frac{N_k}{N}$$

$$\frac{\partial \ln P(x|\mu,\Sigma,\pi)}{\partial \Sigma} = 0$$

$$= \frac{\partial}{\partial \Sigma} \sum_{n=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k N(x^{(n)}|\mu_k,\Sigma)\right)$$

$$= \sum_{n=1}^{N} \frac{\pi_k}{\sum_{k=1}^{K} \pi_k N(x^{(n)}|\mu_k,\Sigma)} \cdot \frac{\partial}{\partial \Sigma} N(x^{(n)}|\mu_k,\Sigma)$$

$$\frac{\partial}{\partial \Sigma} N(x^{(n)}|\mu_k,\Sigma)$$

$$= \frac{\partial}{\partial \Sigma} \frac{1}{2\pi\sqrt{\det\Sigma}} \exp\left(-\frac{1}{2}(x^{(n)}-\mu_k)^T \Sigma^{-1}(x^{(n)}-\mu_k)\right)$$

$$= \left(\frac{\partial}{\partial \Sigma} \frac{1}{2\pi\sqrt{\det\Sigma}}\right) \exp\left(-\frac{1}{2}(x^{(n)}-\mu_k)^T \Sigma^{-1}(x^{(n)}-\mu_k)\right) + \frac{1}{2\pi\sqrt{\det\Sigma}} \exp\left(-\frac{1}{2}(x^{(n)}-\mu_k)^T \Sigma^{-1}(x^{(n)}-\mu_k)\right) \cdot \frac{\partial}{\partial \Sigma}\left(-\frac{1}{2}(x^{(n)}-\mu_k)^T \Sigma^{-1}(x^{(n)}-\mu_k)\right)$$

$$= \frac{1}{2\pi} \cdot \left(-\frac{1}{2}\right)(\det\Sigma)^{-1/2}(\Sigma^{-1})^T \exp\left(-\frac{1}{2}(x^{(n)}-\mu_k)^T \Sigma^{-1}(x^{(n)}-\mu_k)\right) + N(x^{(n)}|\mu_k,\Sigma) \cdot \frac{1}{2}\left(\Sigma^{-1}(x_n^{(n)}-\mu_k)(x_n^{(n)}-\mu_k)^T \Sigma^{-1}\right)^T$$

$$= \frac{-\frac{1}{2}(\Sigma^{-1})^T}{2\pi\sqrt{\det\Sigma}} \exp\left(-\frac{1}{2}(x^{(n)}-\mu_k)^T \Sigma^{-1}(x^{(n)}-\mu_k)\right) + N(x^{(n)}|\mu_k,\Sigma) \cdot \frac{1}{2}\left(\Sigma^{-1}(x^{(n)}-\mu_k)(x^{(n)}-\mu_k)^T \Sigma^{-1}\right)^T$$

$$= -\frac{1}{2} N(x^{(n)}|\mu_k,\Sigma)\left[(\Sigma^{-1})^T - \left(\Sigma^{-1}(x^{(n)}-\mu_k)(x^{(n)}-\mu_k)^T \Sigma^{-1}\right)^T\right]$$

$$\Rightarrow \frac{\partial \ln P(X|\mu,\Sigma,\pi)}{\partial \Sigma} = \sum_{n=1}^{N} \frac{\pi_k N(x^{(n)}|\mu_k,\Sigma)}{\sum_{k=1}^{K} \pi_k N(x^{(n)}|\mu_k,\Sigma)} \cdot \left(-\frac{1}{2}\right)\left[(\Sigma^{-1})^T - \left(\Sigma^{-1}(x^{(n)}-\mu_k)(x^{(n)}-\mu_k)^T \Sigma^{-1}\right)^T\right]$$

$$0 = -\frac{1}{2}\sum_{n=1}^{N} \gamma(z_k^{(n)})\left[(\Sigma^{-1})^T - \left(\Sigma^{-1}(x^{(n)}-\mu_k)(x^{(n)}-\mu_k)^T \Sigma^{-1}\right)^T\right]$$

$$0 = -\frac{1}{2}(\Sigma^{-1})^T \sum_{n=1}^{N} \gamma(z_k^{(n)})\left[I - \left(\Sigma^{-1}(x^{(n)}-\mu_k)(x^{(n)}-\mu_k)^T\right)^T\right], \quad (\Sigma^{-1})^T = \Sigma^{-1} \text{ b/c } \Sigma \text{ is symmetric.}$$

$$0 = \sum_{n=1}^{N} \gamma(z_k^{(n)}) - \sum_{n=1}^{N} \gamma(z_k^{(n)})\left(\Sigma^{-1}(x^{(n)}-\mu_k)(x^{(n)}-\mu_k)^T\right)^T$$

$$\left(\sum_{n=1}^{N} \gamma(z_k^{(n)})\right)^T = \left(\sum_{n=1}^{N} \gamma(z_k^{(n)})\left(\Sigma^{-1}(x^{(n)}-\mu_k)(x^{(n)}-\mu_k)^T\right)^T\right)^T \Rightarrow N_k = \Sigma^{-1}\sum_{n=1}^{N} \gamma(z_k^{(n)})(x^{(n)}-\mu_k)(x^{(n)}-\mu_k)^T$$

$$\Sigma = \frac{1}{N_k}\sum_{n=1}^{N} \gamma(z_k^{(n)})(x^{(n)}-\mu_k)(x^{(n)}-\mu_k)^T$$

#2.1, #2.2

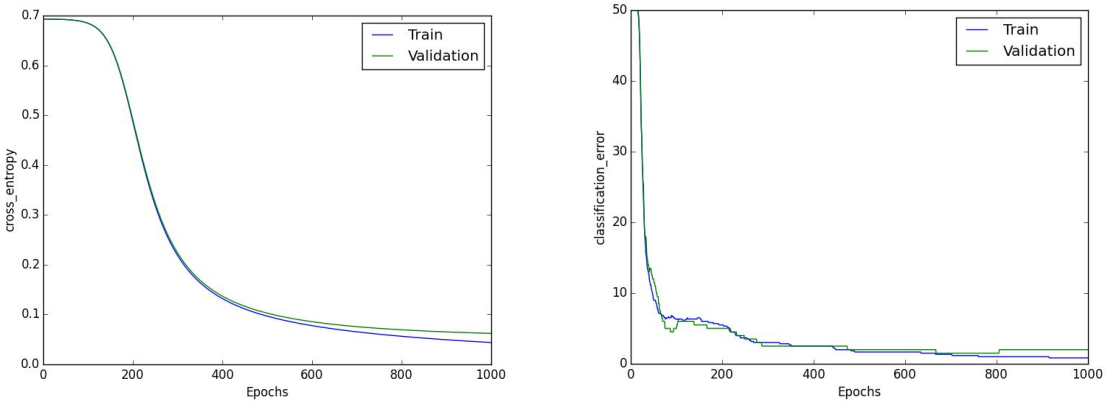

*Fig1: **Cross entropy** and classification error vs. epochs on both train and validation sets*

Validation set will generally have higher cross entropy and classification error compared to the training set. This is because we only look at the training set during training, and therefore parameters are more optimized for training set (possibly overfit after 500 iterations, since the two curves diverge at that point)

#2.3

Learning rate:

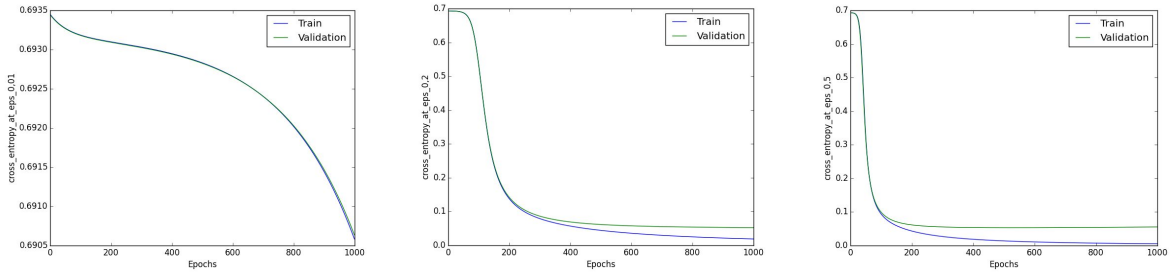

*Fig2: **Cross entropy** vs epoch on test/valid set, learning rate of 0.01 0.2 0.5 from left to right*

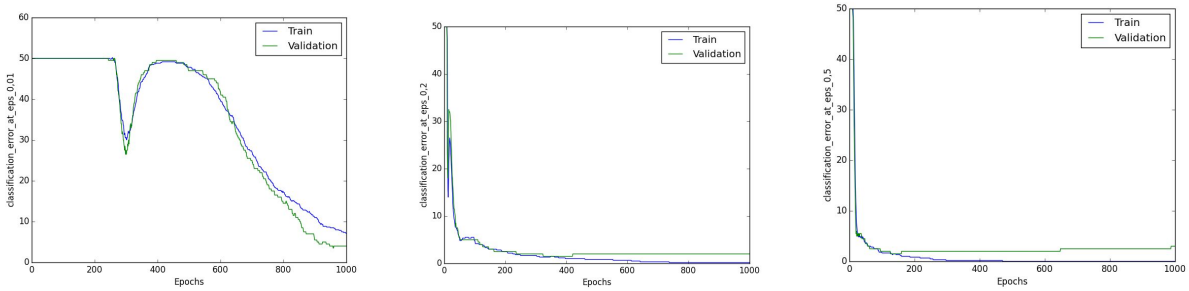

*Fig3: **Class error** vs epoch on train/valid set, learning rate of 0.01 0.2 0.5 from left to right*

NN with higher learning rate converges faster. However the effect is less significant at higher learning rates and both cross entropy and classification error increases after 500 epochs when learning rate is 0.5. In fact, classification error fails to converge at very high learning rates, as shown below. Therefore I would choose the **learning_rate=0.2**

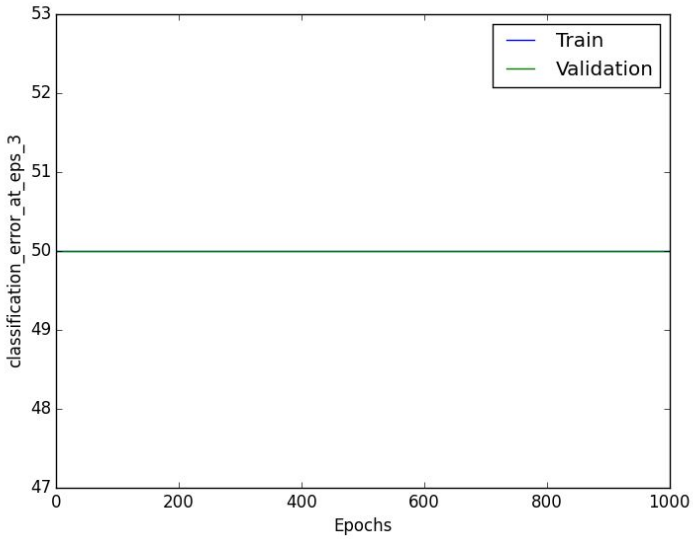

Fig4: Classification error vs epoch at learning rate of 3

Momentum:

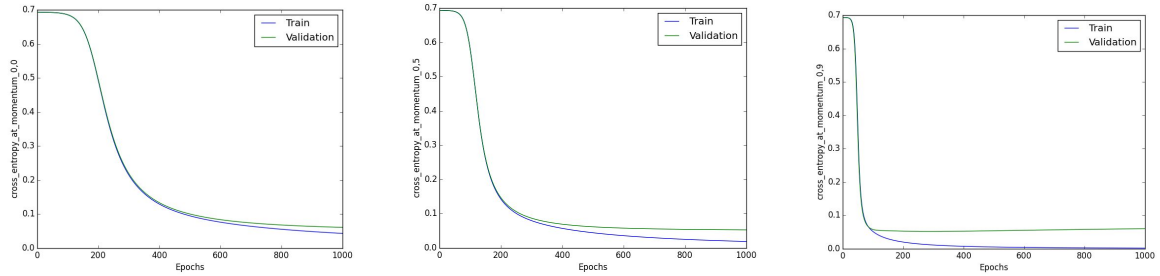

*Fig5: **Cross entropy** vs epoch on train/valid set, momentum of 0 0.5 0.9 from left to right*

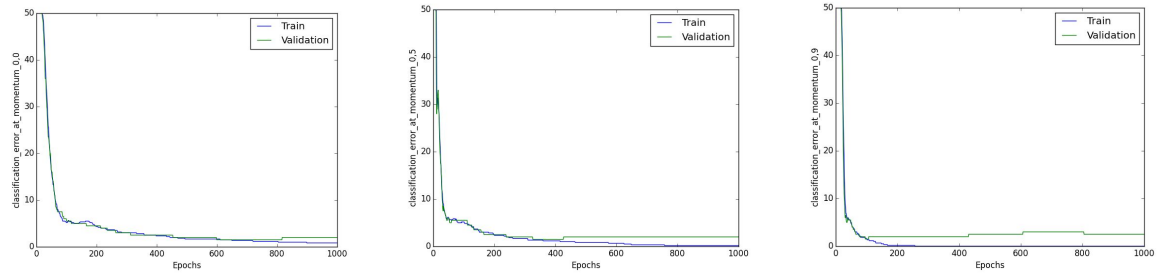

*Fig6: **Class error** vs epoch on train/valid set, momentum of 0 0.5 0.9 from left to right*

Similar to learning rate, NN with higher momentum converges faster. However at higher momentum (0.9), cross entropy increases after 500 epochs. I would choose **momentum=0.5**

#2.4

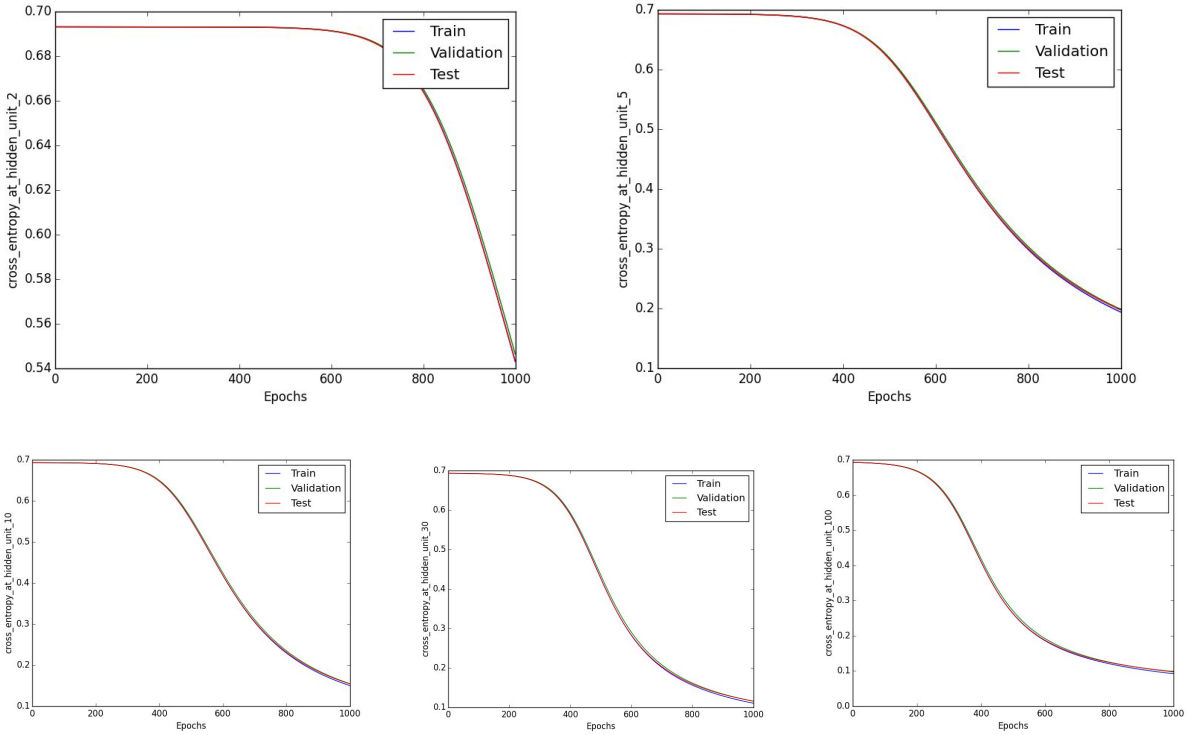

*Fig7: **Cross entropy** vs hidden units on train/test/valid/ set with 2 5 10 30 100 hidden units*

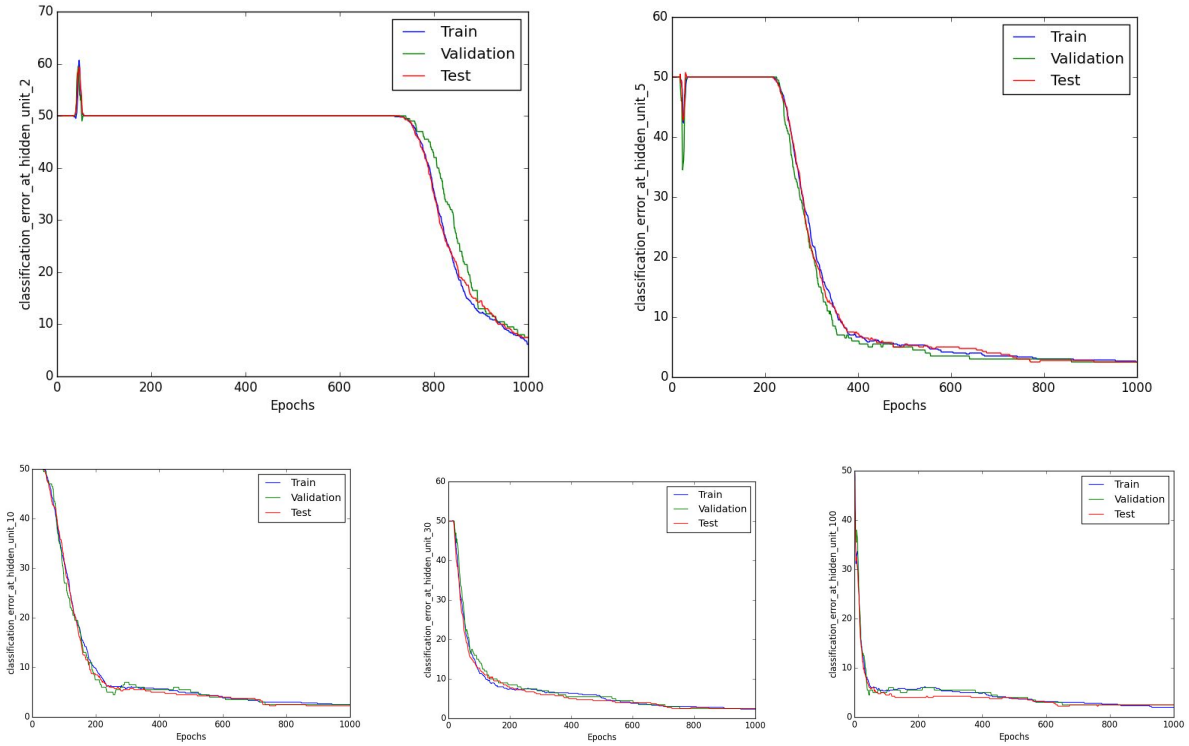

*Fig8: **Cross entropy** vs hidden units on train/test/valid/ set with 2 5 10 30 100 hidden units*

Cross entropy for train/valid/test sets converges to a lower value when the number of hidden unit is not too big or small, as seen in the plot with 30 hidden units. At higher or lower number of hidden units, cross entropy converges to a higher value, as seen in the plots with 2, 5, 10 and 100 hidden units. Classification error converges to roughly the same value regardless of the number of hidden units, but the rate of convergence is higher with more hidden units The best number of hidden units to use is 20.

#2.5

KNN produced the following results

K	Validation class error %	Test class error %
1	1.5	1.5
3	1.0	1.25
5	2.0	1.25
7	1.5	1.0
9	1.5	1.25

KNN produced slightly lower classification error compared to NN. In terms of running time, NN spends significant amount of time to train, as shown below

hidden nodes	Training time in seconds
2	1.25
5	1.44
10	1.73
30	2.71
100	6.05

While KNN took 0.012 seconds to classify all 400 images in the validation set, NN took 0.00014 which is much faster than KNN. NN is more scalable than KNN where the break-even load is 48K images, at which point KNN and NN will take the same amount of time to train and classify all images.

#3.2

Model 2:

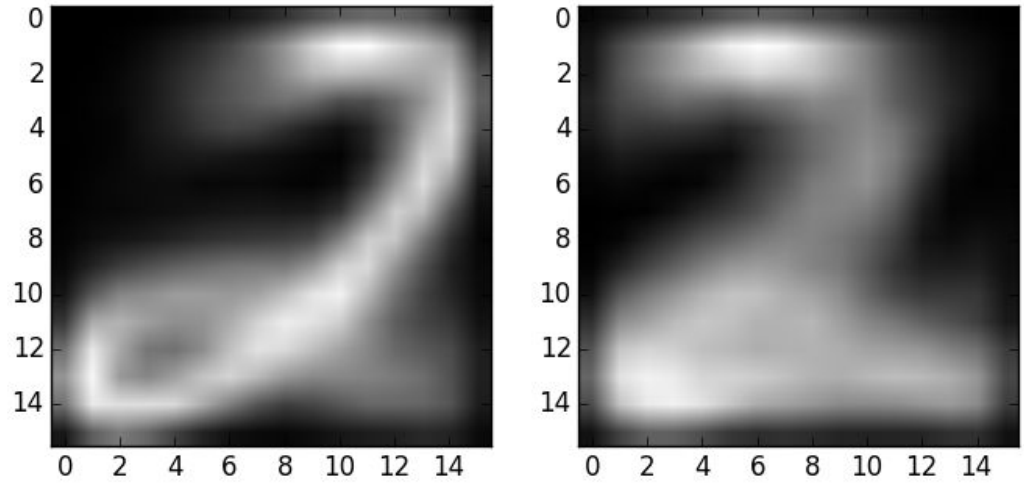

Fig8: Mean vector for the two clusters in 2

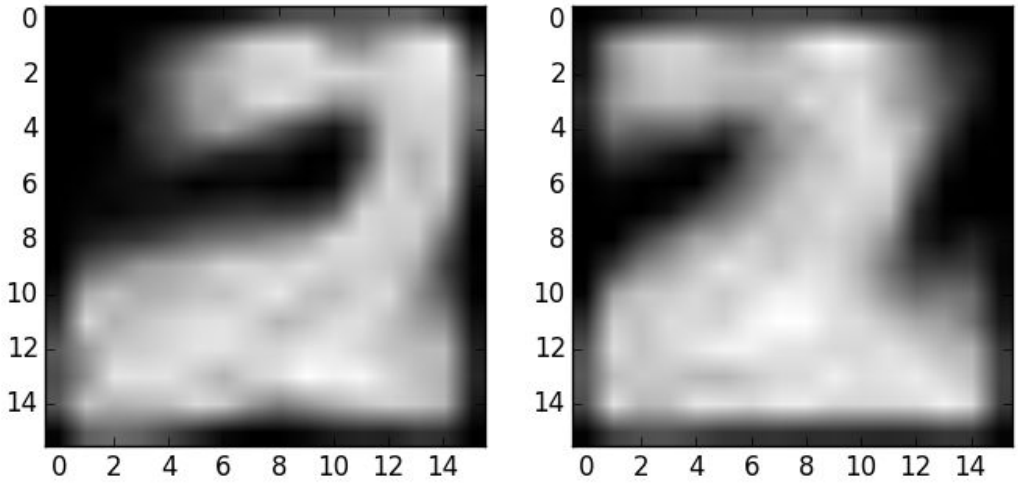

Fig9: Variance vector for the two clusters in 2

LogProbX for 2 is: [-3868.40633613]
Mixing coefficients for 2 are: [[0.50664613], [0.49335387]]

Model 3:

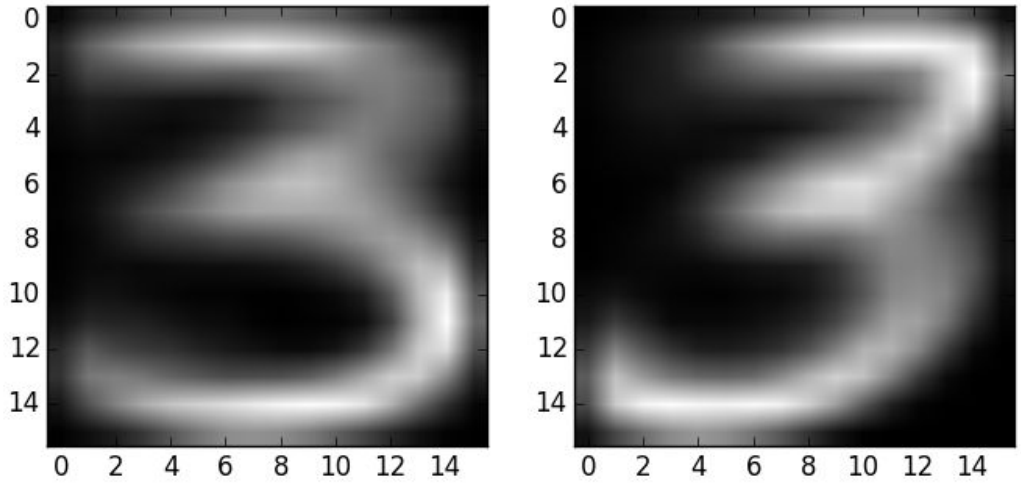

Fig10: Mean vector for the two clusters in 3

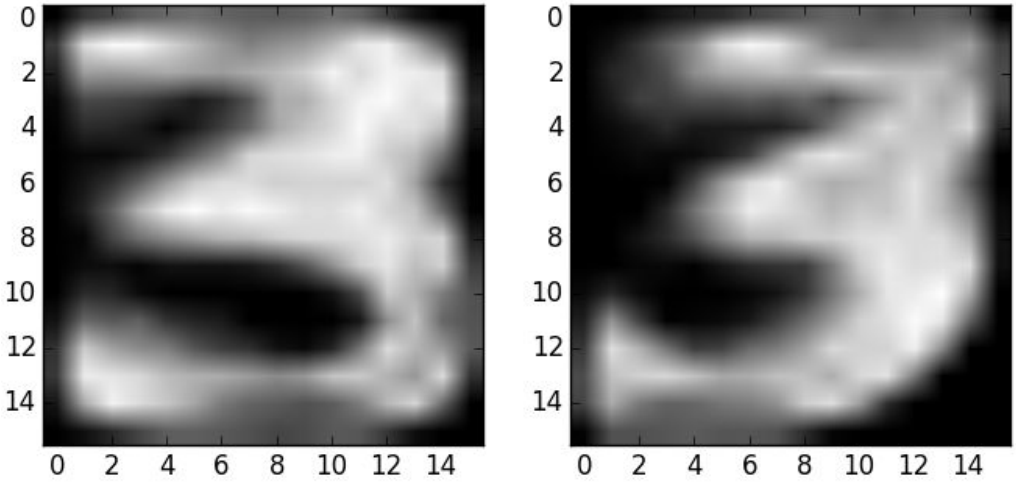

Fig11: Variance vector for the two clusters in 3

LogProbX for 3 is: [2285.5904302]
Mixing coefficients for 3 are: [[0.49974222], [0.50025778]]

Results are taken with **randConst = 10**

#3.3

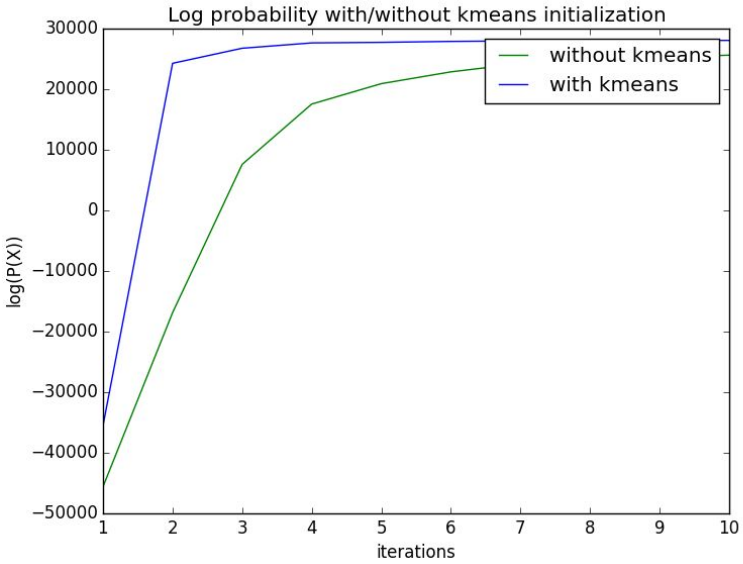

Fig12: log(p(x)) with/without kmeans initialization

Final log(p(x)) without kmeans is 25615, lower than with kmeans, which is 28050. The speed of convergence is faster with kmeans initialization

#3.4

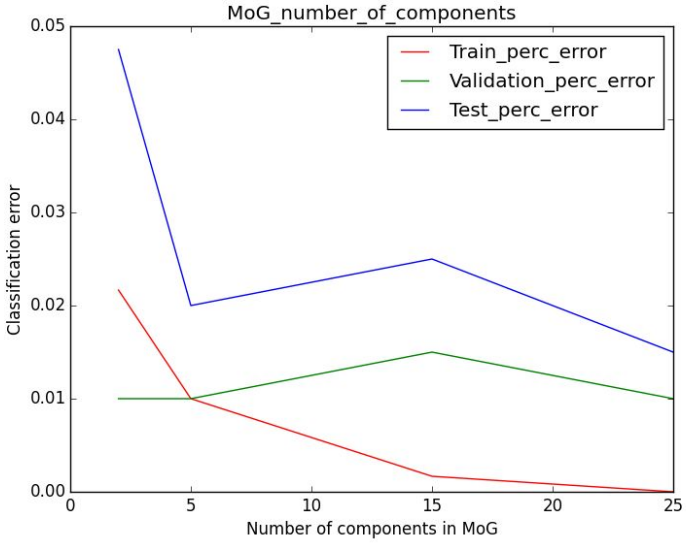

Fig13: Classification error of train/valid/test sets vs. number of components in MoG

1- More components will fit the training set better, and therefore result in a lower classification error (but might overfit). In the limiting case when the number of components equal to the number of training examples, we would get a classification error of 0 if the minimum variance is low enough.

2- Error rate of test set is generally higher than that of training set mostly due to the fact that the MoG model is trained on test set and therefore suits test set better. Other factors that might affect test set performance are:

- Training set might be too small (not representative enough)
- Did not capture all features of test set (still, not representative enough)

Error rate went down when the number of components is increased to 5 (from 2) suggesting better fit of data at nCluster=5. However the error rate went up at 15 components suggesting possible overfitting of data.

3- To achieve the best result for this system, 25 components should be used because it has the lowest train/valid/test classification error. However, 5 components should be used on new images because it has a lower chance of overfitting the current system (training data) and therefore generalizes better

#3.5

The hidden weights (for all 5 hidden units)

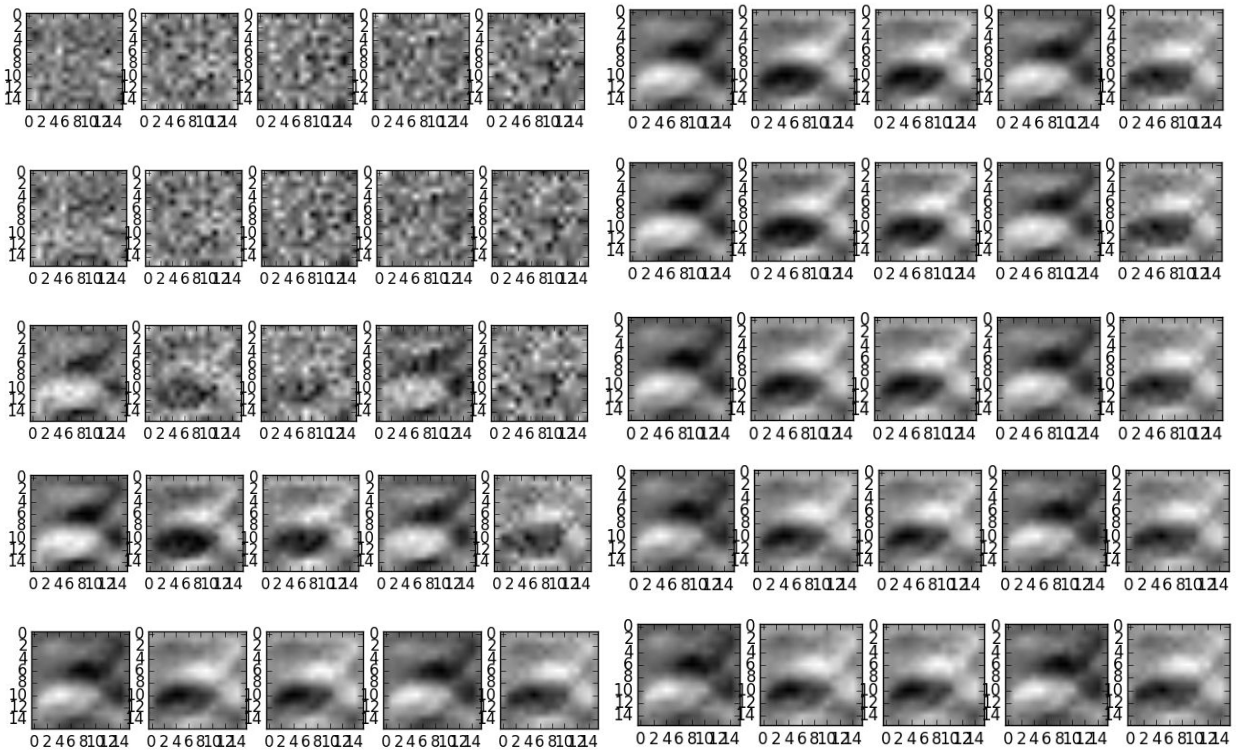

Fig13: Weights after every 900 iterations (starting at 0, ending at 900)

NN learned to recognize 2's and 3's after ~500 iterations, as seen in Fig 13, the weights on image 2, 3, 5 resemble the outline of 3 (where white = 1), image 1, 4 resemble the outline of 2. This is similar to MoG components, where each cluster represents some variation of handwritten 2 and 3. NN and MoG have similar performance in terms of classification error on test and validation sets with MoG achieving a slightly lower error rate. On training set however, MoG had close to zero classification rate with 25 components while NN remained at 5% even with 100 hidden units.