COMS 4772 Fall 2015: Homework #2

Daniel Kronovet - dbk2123@columbia.edu
Discussants: Stephen Ra, Steve Royce

October 16, 2015

Problem 1

We set up the EM master equation across variables X, W, Z as follows:

$$lnp(X,W) = \int q(Z) ln \frac{p(X,W,Z)}{q(Z)} dZ + \int q(Z) ln \frac{q(Z)}{p(Z|X,W)} dZ$$

Given that $X = (x_1, ..., x_n)$ and $Z = (z_1, ..., z_n)$, and are independent, we can write the joint likelihood as follows:

$$ln \sum_i^n p(x_i, W) = \sum_i^n \int q(z_i) ln \frac{p(x_i, W, z_i)}{q(z_i)} dz_i + \sum_i^n \int q(z_i) ln \frac{q(z_i)}{p(z_i|x_i, W)} dz_i$$

Finally, we expand the joint distribution into the product of distributions across x_i, W, z_i:

$$ln \sum_i^n p(x_i, W) = \sum_i^n \int q(z_i) ln \frac{p(x_i|W, z_i)p(W)P(z_i)}{q(z_i)} dz_i + \sum_i^n \int q(z_i) ln \frac{q(z_i)}{p(z_i|x_i, W)} dz_i$$

Now, for the E-step of the algorithm, we must derive the posterior distribution $p(z_i|x_i, W)$. We begin with:

$$p(z_i|x_i, W) = \frac{p(x_i|W, z_i)p(W)p(z_i)}{\int p(x_i|W, z_i)p(W)p(z_i)dz_i}$$

As we are not interested in a posterior on W, we can treat it as a parameter, and thus cancel it out:

$$p(z_i|x_i, W) = \frac{p(x_i|W, z_i)p(z_i)}{\int p(x_i|W, z_i)p(z_i)dz_i}$$

Now we are dealing with the posterior over two Gaussians, which is itself Gaussian. We focus on the numerator, with the denominator ultimately providing the normalizing constant.

$$p(x_i|W, z_i)p(z_i)$$

$$(2\pi)^{-d/2}(|\sigma^2 I|)^{-1/2}exp\left\{-\frac{1}{2}(x_i - Wz_i)^T(\sigma^2 I)^{-1}(x_i - Wz_i)\right\}(2\pi)^{-k/2}(|I|)^{-1/2}exp\left\{-\frac{1}{2}(z_i)^T(I)^{-1}(z_i)\right\}$$

$$(2\pi)^{\frac{-d+k}{2}}(|\sigma^2 I|)^{-1/2}exp\left\{-\frac{1}{2}\left[(x_i - Wz_i)^T(\sigma^2 I)^{-1}(x_i - Wz_i) + (z_i)^T(z_i)\right]\right\}$$

To finish the derivation, we turn to Bishop 2.3, which gives $z_i \sim N(\mu_i', \Sigma_i')$:

$$\Sigma_i' = \left(I + \frac{W^T W}{\sigma^2}\right), \mu_i' = \Sigma_i' \frac{W^T x_i}{\sigma^2}$$

With the posterior in hand, we can return to the EM master equation to complete the E-step by setting $q(z_i) = p(z_i|x_i, W)$, with $W = W_t$, the value of W at the t-th iteration of our algorithm:

$$ln\sum_i^n p(x_i, W) = \underbrace{\sum_i^n \int p(z_i|x_i, W) ln\frac{p(x_i|W, z_i)p(W)p(z_i)}{p(z_i|x_i, W)}dz_i}_{\mathcal{L}(W_t)} + \underbrace{\sum_i^n \int p(z_i|x_i, W) ln\frac{p(z_i|x_i, W)}{p(z_i|x_i, W)}dz_i}_{\text{KL-Divergence}}$$

We now attempt to find W_{t+1} by maximizing $\mathcal{L}(W)$.

$$\mathcal{L}(W) = \sum_i^n \int p(z_i|x_i, W) ln\frac{p(x_i|W, z_i)p(W)p(z_i)}{p(z_i|x_i, W)}dz_i$$

$$= \sum_i^n \int p(z_i|x_i, W) lnp(x_i|W, z_i)p(W)p(z_i)dz_i - \int p(z_i|x_i, W) lnp(z_i|x_i, W)dz_i$$

The left-hand term, the entropy of $q(z_i)$, does not depend on W, and can be excluded from the optimization.

$$\sum_i^n \int p(z_i|x_i, W) lnp(x_i|W, z_i)p(W)p(z_i)dz_i + const$$

$$= \sum_i^n \mathbb{E}_q\left[lnp(x_i|W, z_i)p(W)p(z_i)\right] + const$$

$$= \sum_i^n \mathbb{E}_q\left[lnp(x_i|W, z_i)\right] + \mathbb{E}_q\left[lnp(W)\right] + \mathbb{E}_q\left[lnp(z_i)\right] + const$$

Again, we can drop the left-most term, which does not vary with W and will disappear once we take the derivative.

$$= \sum_i^n \mathbb{E}_q\left[lnp(x_i|W, z_i)\right] + \mathbb{E}_q\left[lnp(W)\right] + const$$

We now expand into the distributions:

$$= \sum_i^n \mathbb{E}_q \left[-\frac{d}{2} ln(2\pi) - \frac{1}{2} ln(|I\sigma^2|) - \frac{1}{2\sigma^2} ||x_i - W z_i||_2^2 \right] + \mathbb{E}_q \left[\frac{dk}{2} ln\left(\frac{\lambda}{2\pi}\right) - \frac{\lambda}{2} trace(W^T W) \right] + const$$

We take advantage of the linearity of expectation:

$$= \sum_i^n -\frac{d}{2} ln(2\pi) - \frac{1}{2} ln(|I\sigma^2|) - \frac{1}{2\sigma^2} \mathbb{E}_q \left[||x_i - W z_i||_2^2 \right] + \frac{dk}{2} ln\left(\frac{\lambda}{2\pi}\right) - \frac{\lambda}{2} trace(W^T W) + const$$

And again exclude every term not involving W.

$$= \sum_i^n -\frac{1}{2\sigma^2} \mathbb{E}_q \left[||x_i - W z_i||_2^2 \right] - \frac{\lambda}{2} trace(W^T W) + const$$

We now expand the quadratic term.

$$= \sum_i^n -\frac{1}{2\sigma^2} \mathbb{E}_q \left[||x_i|| - 2 x_i^T W z_i + ||W z_i|| \right] - \frac{\lambda}{2} trace(W^T W) + const$$

And carry through the expectation.

$$= \sum_i^n -\frac{1}{2\sigma^2} \left(||x_i|| - 2 x_i^T W \mathbb{E}_q \left[z_i \right] + \mathbb{E}_q \left[||W z_i|| \right] \right) - \frac{\lambda}{2} trace(W^T W) + const$$

$$= \sum_i^n -\frac{1}{2\sigma^2} \left(||x_i|| - 2 x_i^T W \mu_i' + \mathbb{E}_q \left[||W z_i|| \right] \right) - \frac{\lambda}{2} trace(W^T W) + const$$

$$= \sum_i^n -\frac{1}{2\sigma^2} ||x_i|| + \sum_i^n \frac{1}{\sigma^2} x_i^T W \mu_i' - \sum_i^n \frac{1}{2\sigma^2} \mathbb{E}_q \left[||W z_i|| \right] - \frac{\lambda}{2} trace(W^T W) + const$$

At last, we take the derivative of $\mathcal{L}(W)$:

$$\mathcal{L}_{\nabla_W} = \sum_i^n \frac{1}{\sigma^2} x_i \mu_i'^T - \sum_i^n \frac{1}{2\sigma^2} \mathbb{E}_q \left[2(W z_i) z_i^T \right] - \frac{\lambda}{2} 2W = 0$$

$$\sum_i^n \frac{1}{\sigma^2} x_i \mu_i'^T - \sum_i^n \frac{1}{\sigma^2} W \mathbb{E}_q \left[z_i z_i^T \right] - \lambda W = 0$$

$$\sum_i^n \frac{1}{\sigma^2} x_i \mu_i'^T - \sum_i^n \frac{1}{\sigma^2} W (\Sigma_i' + \mu_i' \mu_i'^T) - \lambda W = 0$$

Reorganizing:

$$\sum_i^n x_i \mu_i'^T = W \left[\sum_i^n (\Sigma_i' + \mu_i' \mu_i'^T) + \sigma^2 \lambda \right]$$

$$\left[\sum_i^n x_i \mu_i'^T \right] \left[\sum_i^n (\Sigma_i' + \mu_i' \mu_i'^T) + \sigma^2 \lambda \right]^{-1} = W = W_{t+1}$$

Finishing the derivation.

To run the actual algorithm, we first set $q(z_i) = p(z_i|x_i, W_t)$, which allows us to generate $\mathcal{L}_{W_t}$.

Then, we take the derivative of this $\mathcal{L}_{W_t}$ with respect to W to find the value W_{t+1} which maximizes $\mathcal{L}_{W_t(W)}$.

Once we have this new W_{t+1}, we can calculate the new $q(z_i) = p(z_i|x_i, W_{t+!})$. From here, we can repeat the process to derive better and better value of W, our model variable of interest.

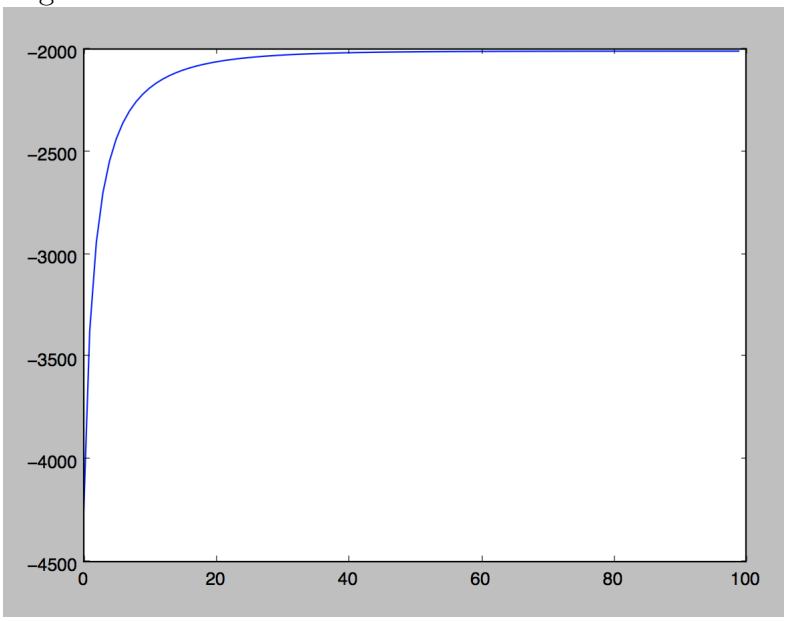
Problem 2

Part a

I did this it was fun.

Part b

Log likelihood at T=100



Part c

	0	1
0	931	51
1	77	932

$$Accuracy = 0.935710698142$$

Part d: Misclassified Images

Image 156, $\sim Bern(0.836626)$

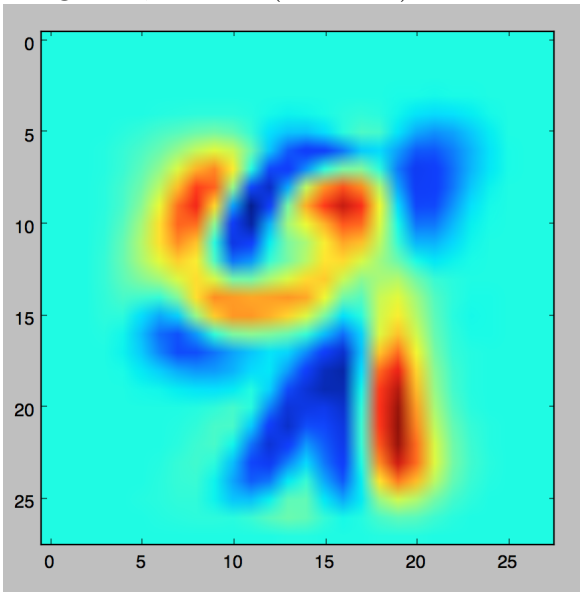


Image 564, $\sim Bern(0.584582)$

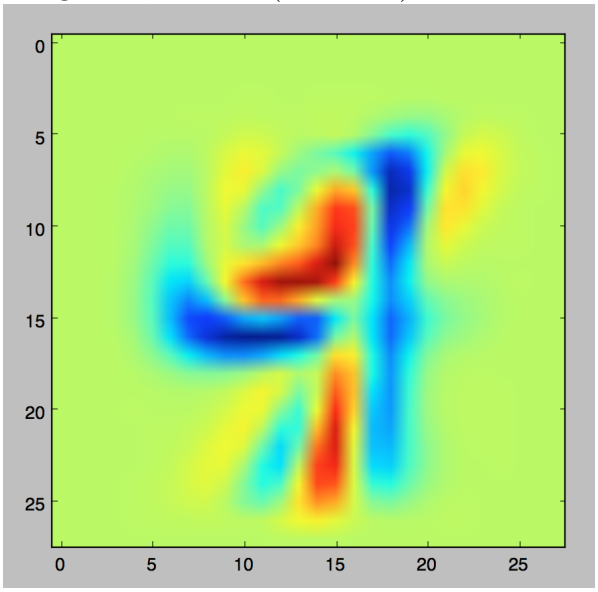
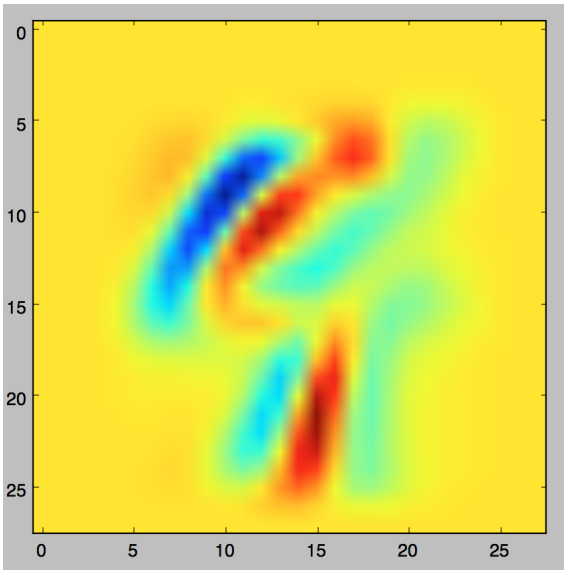


Image 1658, $\sim Bern(0.47498)$



Part e: Ambiguous Predictions

Image 210, $\sim Bern(0.502074)$

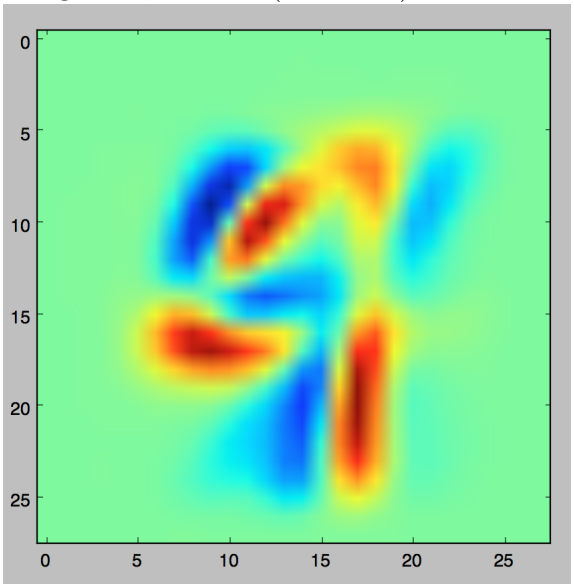


Image 340, $\sim Bern(0.505406)$

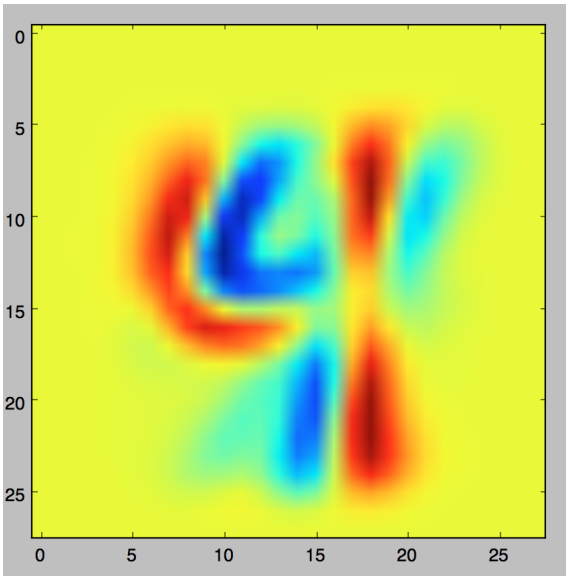
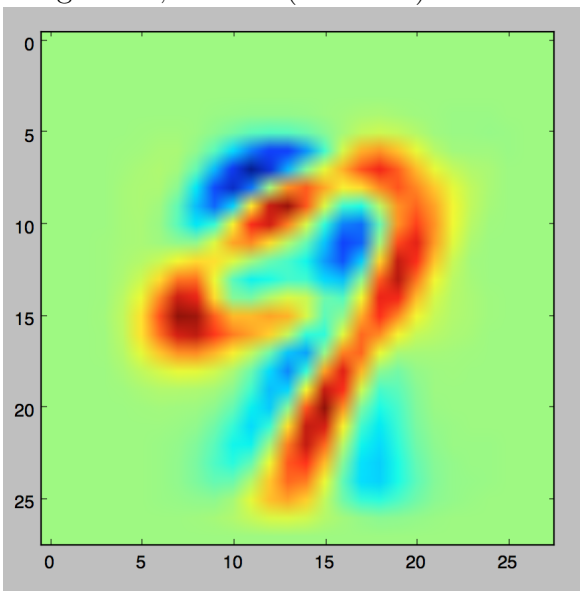
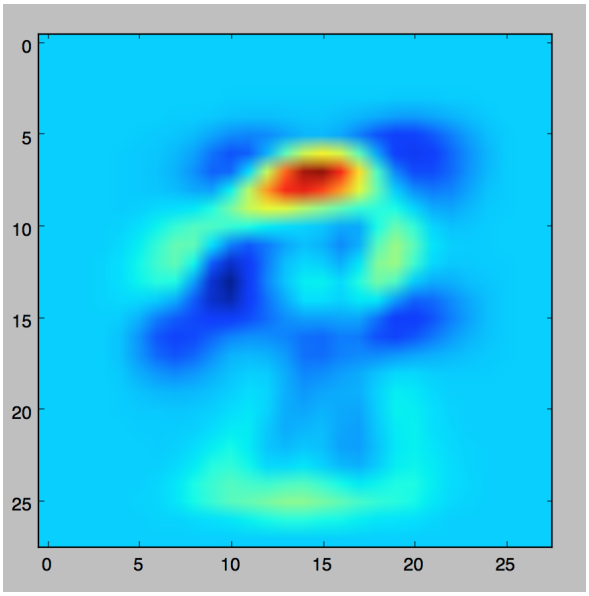


Image 1990, $\sim Bern(0.501256)$

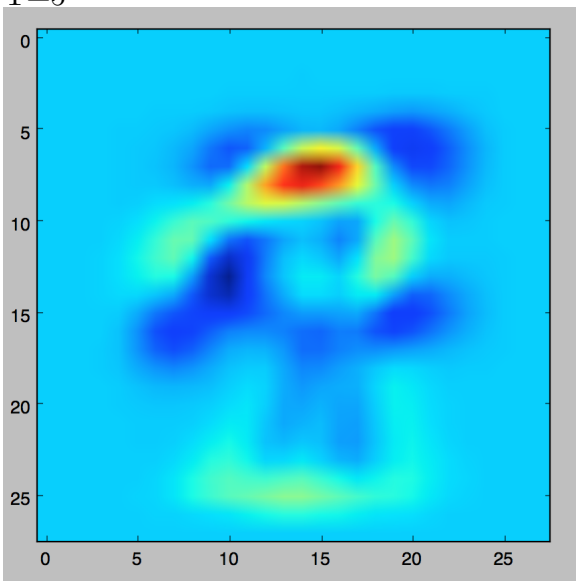


Part e: Changes to w over time

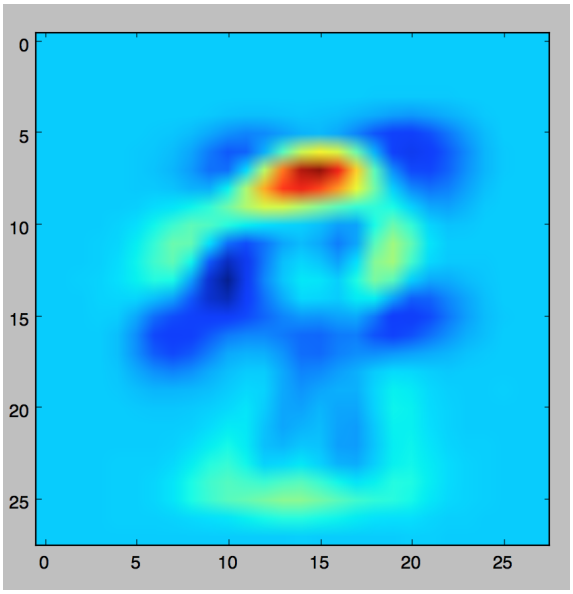
T=1



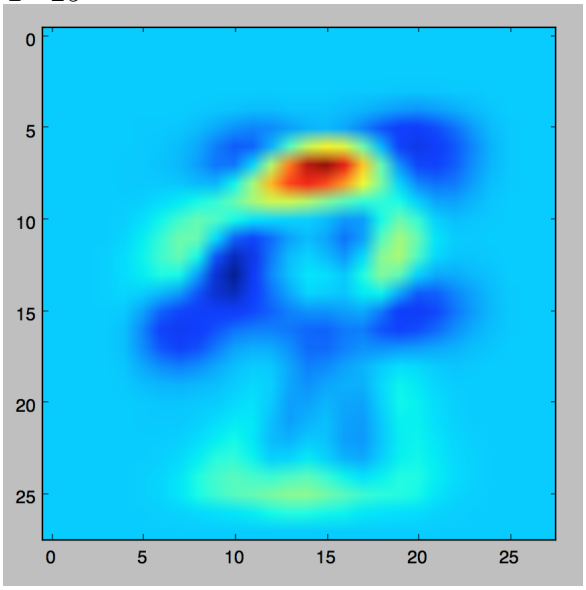
T=5



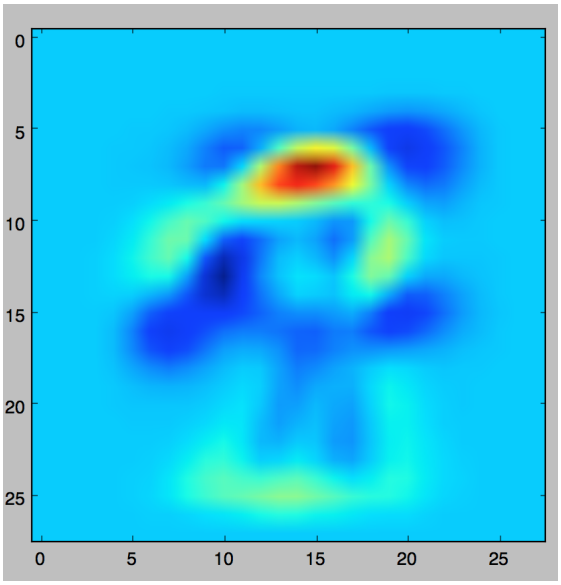
T=10



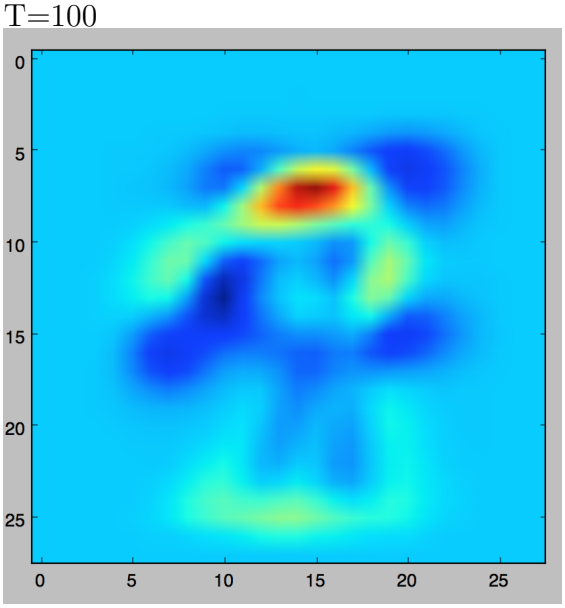
T=25



T=50



T=100



I noticed relatively little change in the rendered w. There were small changes in the distribution of heat across the image, which I interpret as the algorithm learning more precisely which components of a handwritten image are the most powerful discriminators.