Zoltan Onodi-Szucs
zo2131

Homework 2
Bayesian Machine Learning

1.

Assume we have $\{x_1, \ldots, x_N\}$ with $x \in \mathbb{R}^d$. We model them as independent random variables with $x_n \sim N(Wz_n, \sigma^2 I)$. The matrix $W \in \mathbb{R}^{d \times k}$ is unknown and we model $z_n \sim N(0, I)$. The columns of W have zero-mean spherical Gaussian priors with variance λ^{-1}, which can be written according to a matrix Gaussian prior as

$$p(W) = \left(\frac{\lambda}{2\pi}\right)^{dk/2} \exp\left\{-\frac{\lambda}{2}\text{trace}(W^T W)\right\}.$$

Our goal is to find

$$W' = \arg\max_W \ \ln p(x_1, \ldots, x_N, W).$$

Derive an EM algorithm for doing this where $z_1, \ldots, z_N$ function as the "hidden variables" we integrate out. As a reminder, at iteration t this involves calculating posteriors for each z_n given W_{t-1}, taking expectations of the log joint likelihood using each $q(z_n)$, and maximizing over the result with respect to W. Clearly describe every step in your derivation of the algorithm for full credit, both mathematically and in words. Also, summarize the steps of the algorithm at the end using pseudo-code.

(Side comment: There is no correct pseudo-code language we are looking for. We simply want to see a clear picture of the algorithm you would implement without having to search for the pieces in your response. This should be in addition to your detailed derivation.)

Writing out the EM master equation treating z_i as the hidden variable and the problem setting:

$$\ln p(X, W) = \sum_{i=1}^{N} \int q(z_i) \ln \frac{p(x_i, W, z_i)}{q(z_i)} dz_i + \sum_{i=1}^{N} \int q(z_i) \ln \frac{q(z_i)}{p(z_i | x_i, W)}$$

First we need to calculate the posterior of z:

$$p(z | X, W) \propto p(X | W, z) p(z) \propto$$

$$\propto \prod_{i=1}^{N} (2\pi)^{-d/2} \cdot |\sigma^2 I|^{-1/2} \exp\left\{-\frac{1}{2}(x_i - Wz_i)^T (\sigma^2 I)(x_i - Wz_i)\right\} \cdot \prod_{i=1}^{N} (2\pi)^{-d/2} \exp\left\{-\frac{1}{2} z_i^T z_i\right\}$$

$$\propto 2\pi^{-dN/2} \ |\sigma^2 I|^{-1/2} \exp\left\{\sum_{i=1}^{N} -\frac{1}{2} \frac{(x_i - Wz_i)^T (x_i - Wz_i)}{\sigma^2}\right\} \cdot 2\pi^{-\frac{Nd}{2}} \exp\left\{\sum_{i=1}^{N} -\frac{1}{2} z_i^T z_i\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{(x_i - Wz_i)^T(x_i - Wz_i)}{\sigma^2} + z_i^T z_i\right)\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i^T x_i - x_i^T W z_i - z_i^T W^T x_i + z_i^T W^T W z_i}{\sigma^2} + z_i^T z_i\right)\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\frac{1}{\sigma^2}\left(x_i^T x_i - 2 z_i^T W^T x_i + z_i^T\left(W^T W + \sigma^2 I\right) z_i\right)\right\}$$

Next we have to complete the square to get the form:

$$-\frac{1}{2}\sum_{i=1}^{N}\left[(A - z_i)^T \frac{B}{\sigma^2}(A - z_i) + \frac{C}{\sigma^2}\right]$$

$$B = W^T W + \sigma^2 I$$

$$-2BA = -2W^T x_i$$

$$A = (W^T W + \sigma^2 I)^{-1}(W^T x_i)$$

$$A^T B A + C = x_i^T x_i$$

$$C = x_i^T x_i - (W^T x_i)^T (W^T W + \sigma^2 I)^{-1} (W^T x_i)$$

$$= x_i^T x_i - (x_i^T W)(W^T W + \sigma^2 I)^{-1}(W^T x_i)$$

Next we set $q(z_i)$ to be equal to the posterior

$$\prod_{i=1}^{N} q_t(z_i) \propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\left[(A-z_i)^T \Sigma^{-1}\beta\right)^{-1}(A-z_i)\right]$$

Next we need to do the M step, updating *W* using the expectation:

$$\ln p(X|W) = \underbrace{\sum_{i=1}^{N}\int q(z_i)\ln\frac{p(x_i,W,z_i)}{q(z_i)}dz_i}_{\mathcal{L}(W)} + \sum_{i=1}^{N}\int q(z_i)\ln\frac{q(z_i)}{p(z_i|x_i,W)}$$

$$\mathcal{L}(W) = \sum_{i=1}^{N}\int q(z_i)\ln p(x_i,W,z_i)dz_i + \text{const w.r.t. } W$$

$$= \sum_{i=1}^{N} E_{q_i}\left[\ln p(x_i|W,z_i)\cdot p(W)\cdot p(z_i)\right] + \text{const w.r.t. } W$$

$$= \sum_{i=1}^{N} E_{q_i}\left[\ln p(x_i|W,z_i)\right] + E\left[\ln p(W)\right] + \text{const w.r.t. } W$$

$$= \sum_{i=1}^{N} E_{q_i}\left[-\frac{1}{2\sigma^2}(x_i - Wz_i)^T(x_i - Wz_i)\right] - \frac{\lambda}{2}\mathrm{Tr}(W^TW) + \text{Const}$$

$$= \sum_{i=1}^{N} E_{q_i}\left[-\frac{1}{2\sigma^2}\left(x_i^Tx_i - 2z_i^TW^Tx_i + z_i^TW^TWz_i\right)\right] - \frac{\lambda}{2}\mathrm{Tr}(W^TW) + \text{const}$$

$$= \sum_{i=1}^{N} -\frac{1}{2\sigma^2}\left(-2E[z_i]^TW^Tx_i + E[z_i^TW^TWz_i]\right) - \frac{\lambda}{2}\mathrm{Tr}(W^TW) + \text{const w.r.t. } W$$

$$\nabla_W \mathcal{L}(W) = \frac{1}{\sigma^2} \sum_{i=1}^{N} x_i E[z_i]^T - \frac{1}{2\sigma^2} \sum_{i=1}^{N} E[W \cdot 2 z_i z_i^T] - \lambda W$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{N} x_i E[z_i]^T - \frac{1}{\sigma^2} \sum_{i=1}^{N} W E[z_i z_i^T] - \lambda W$$

$$E[z_i] = \left(W_{(t-1)}^T W_{(t-1)} + \sigma^2 I\right)^{-1} \left(W_{(t-1)}^T x_i\right) = A$$

$$E[z_i z_i^T] = \sigma^2 B^{-1} + AA^T$$

Setting the derivative to zero:

$$\frac{1}{\sigma^2} \sum_{i=1}^{N} x_i A^T - \frac{1}{\sigma^2} \sum_{i=1}^{N} W\left(\sigma^2 B^{-1} + AA^T\right) - \lambda W = 0$$

$$W_{(t)} = \left(\sum_{i=1}^{N} x_i A^T\right) \left(\sum_{i=1}^{N} \left(\sigma^2 B^{-1} + AA^T\right) - \sigma^2 \lambda I\right)^{-1}$$

With

$$B = \left(W_{(t-1)}^T W_{(t-1)} + \sigma^2 I\right)$$

$$A = \left(W_{(t-1)}^T W_{(t-1)} + \sigma^2 I\right)^{-1} \left(W_{(t-1)}^T x_i\right)$$

Algorithm:
First we need to initialize *W* (example with zeros). Next, for each iteration:

- Calculate the posterior of the hidden variable P(z|W,X), using W from previous iteration
- Set q(z) to the posterior
- Update *W* by maximizing L_W using q(z) from previous step
- The updated *W* will be used in the next iteration
- You can check the convergence by calculating the log likelihood

2. Bayesian classifier

In this problem, you will implement an EM algorithm for the probit regression model using the same digits data set from the last homework. A detailed discussin of EM for probit regression can be found in the class notes. For this problem do the following:

a) Implement the EM algorithm for probit regression described in the class notes. Use the parameter setting $\sigma = 1.5$ and $\lambda = 1$. Run your algorithm on the data set provided for $T = 100$ iterations.

b) Plot $\ln p(\vec{y}, w_t|X)$ as a function of t.

c) Make predictions for all data in the testing set by assigning the most probable label to each feature vector. In a 2×2 table, list the total number of 4's classified as 4's, 9's classified as 9's, 4's classified as 9's, and 9's classified as 4's (i.e., a confusion matrix). Use the provided ground truth for this evaluation.

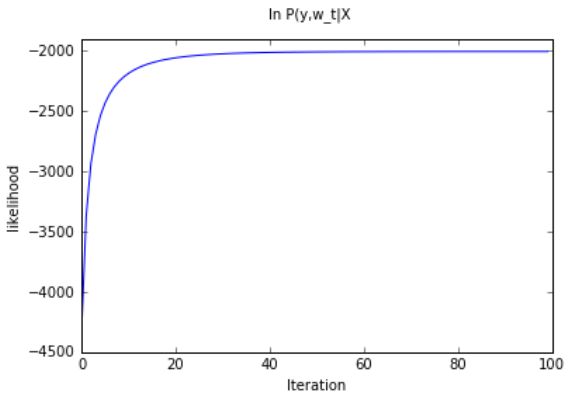
d) Pick three misclassified digits and reconstruct the images as described in the readme file. Show these three images and their predictive probabilities.

e) Pick the three most ambiguous predictions, i.e., the digits whose predictive probabilities are the closest to 0.5. Reconstruct the three images as described in the readme file and show them and their predictive probabilities.

f) Treat the vector w_t as if it were a digit and reconstruct it as an image for $t = 1, 5, 10, 25, 50, 100$. Show these images and comment on what you observe.

a, Code

b,

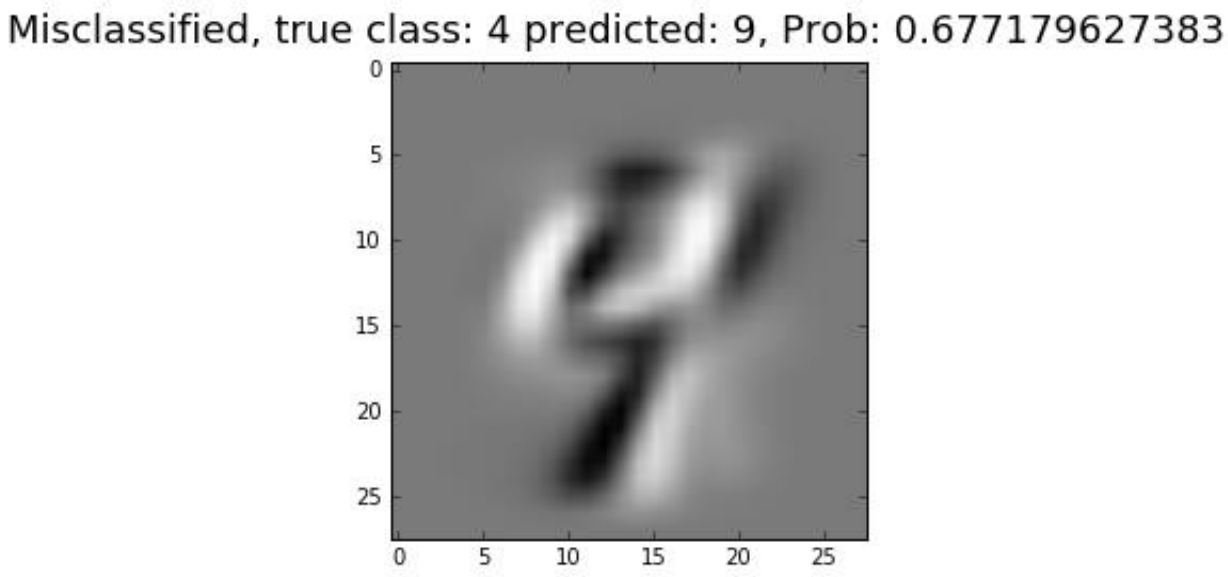


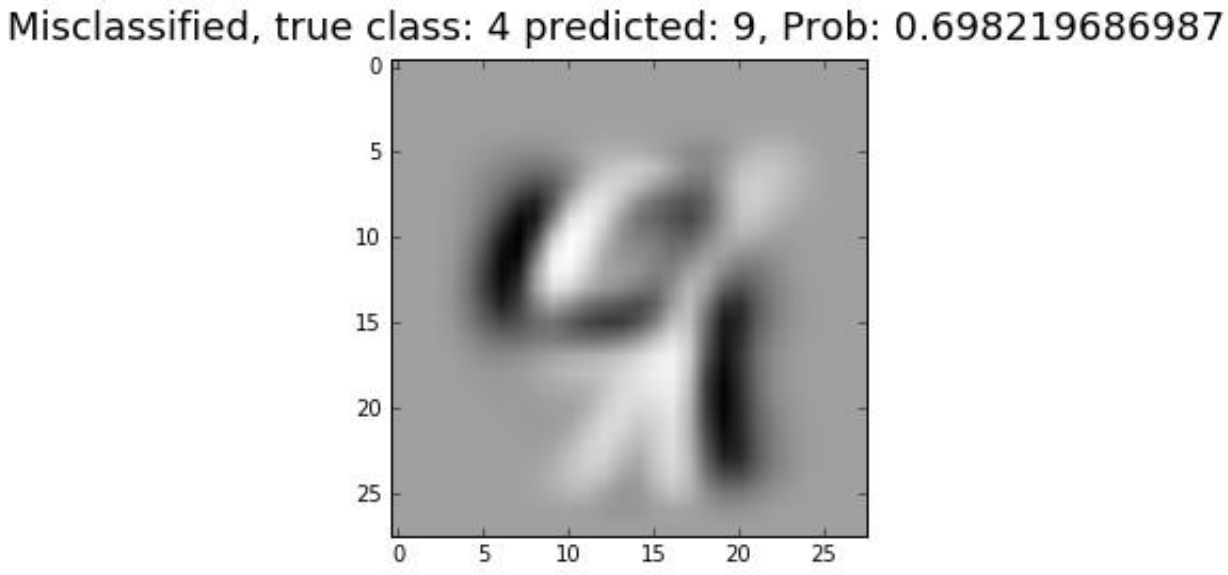
c, Confusion matrix:

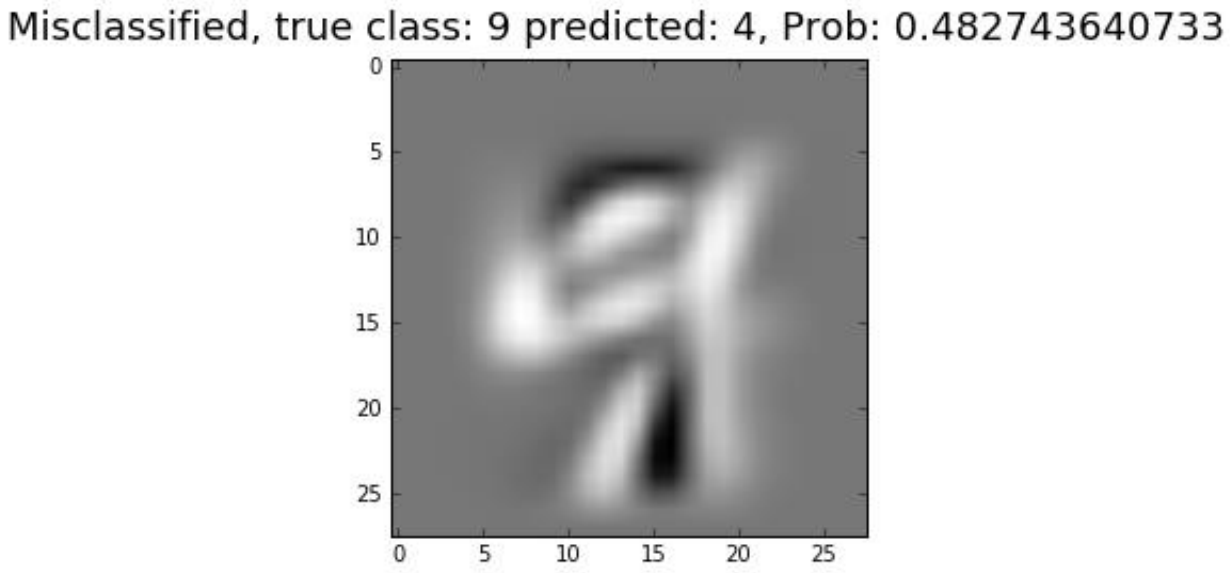
True Label \ Predicted label	0	1
0	930	52
1	77	932

Accuracy: 0.935208437971

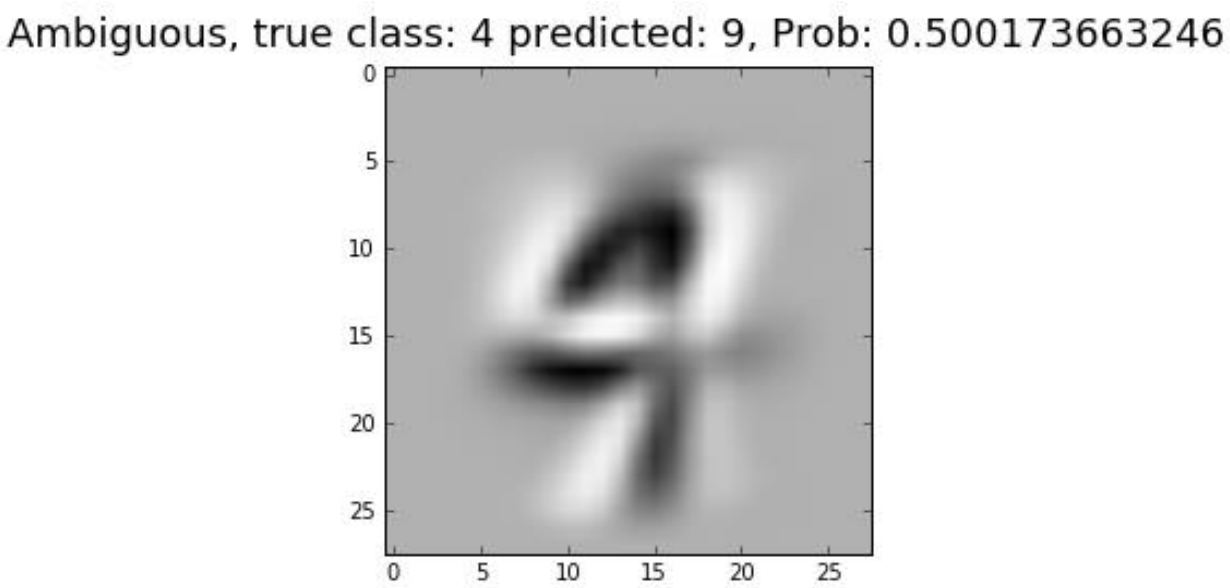
d, Misclassified images

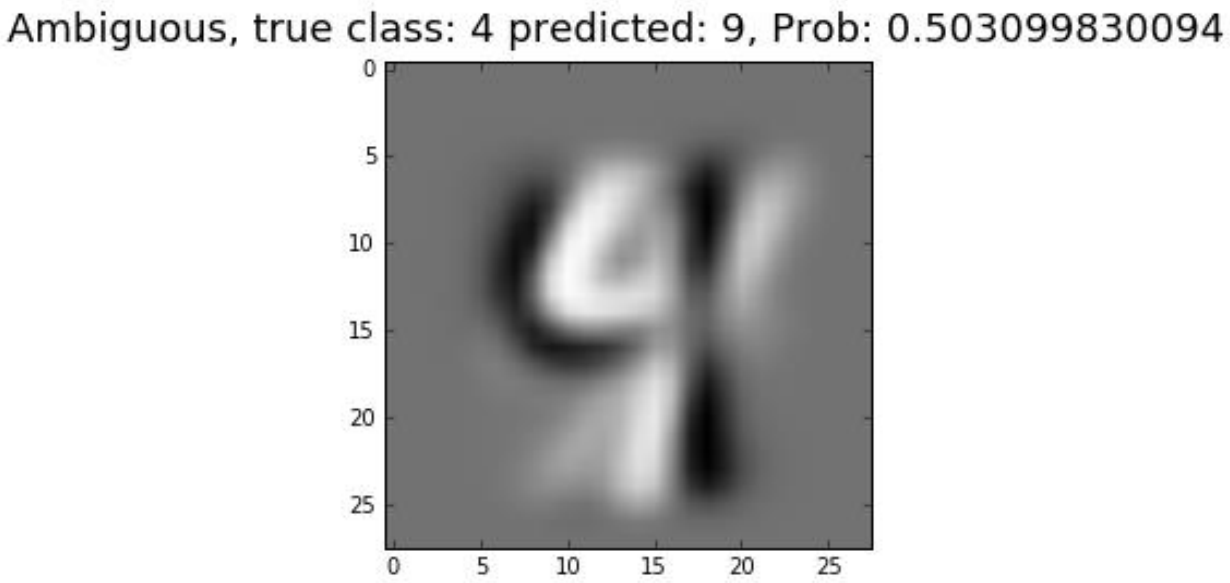


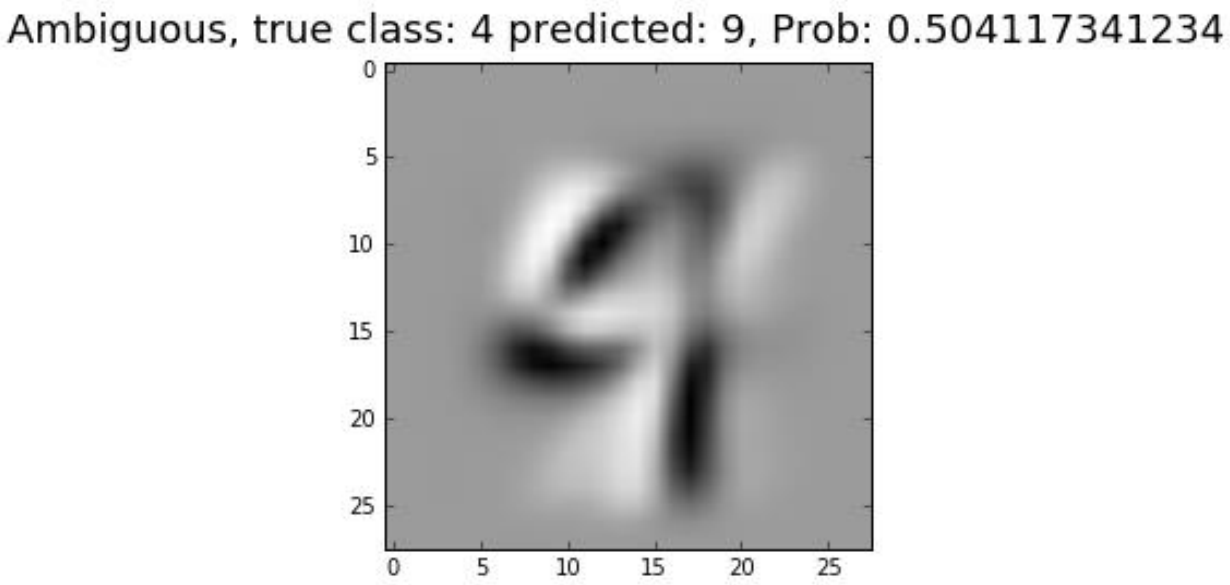




e, Ambiguous images:



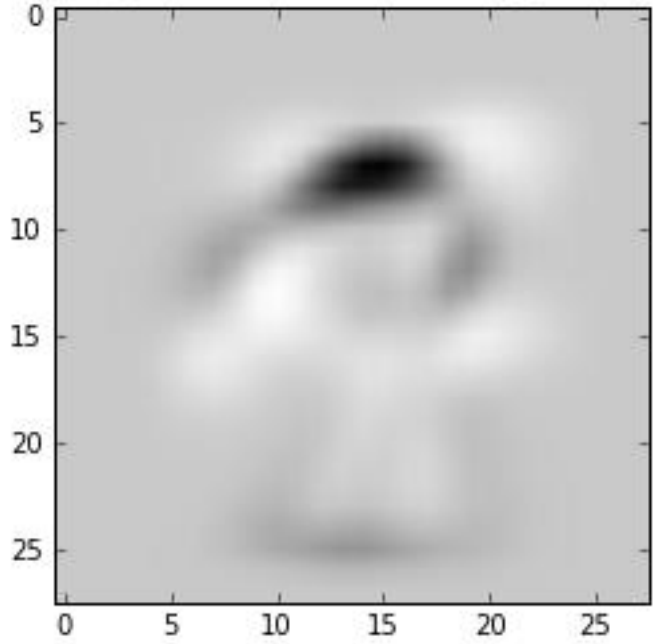




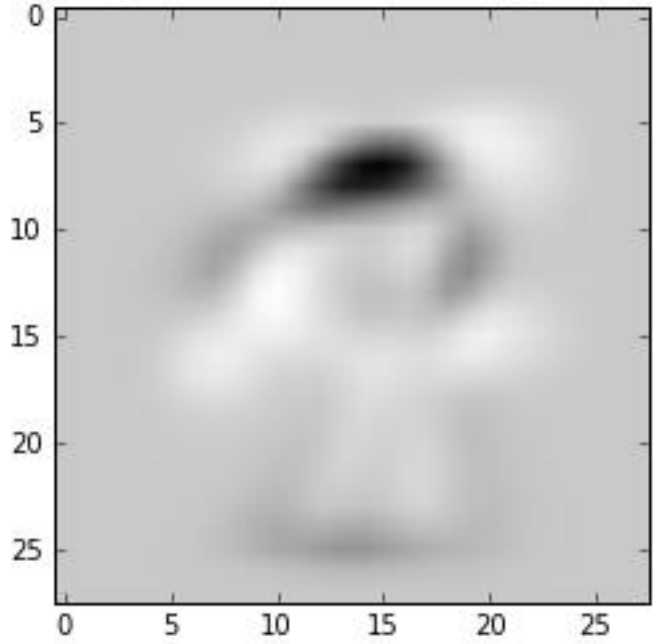
f, weight vectors as image:

We can see that the weight vectors are very similar and even after the first iteration, it looks close to the converged one. We can also see that the dark area in the top side is where the most difference is between a number 4 and a number 9. Within each iteration there is only a slight change in the value of the weigth.

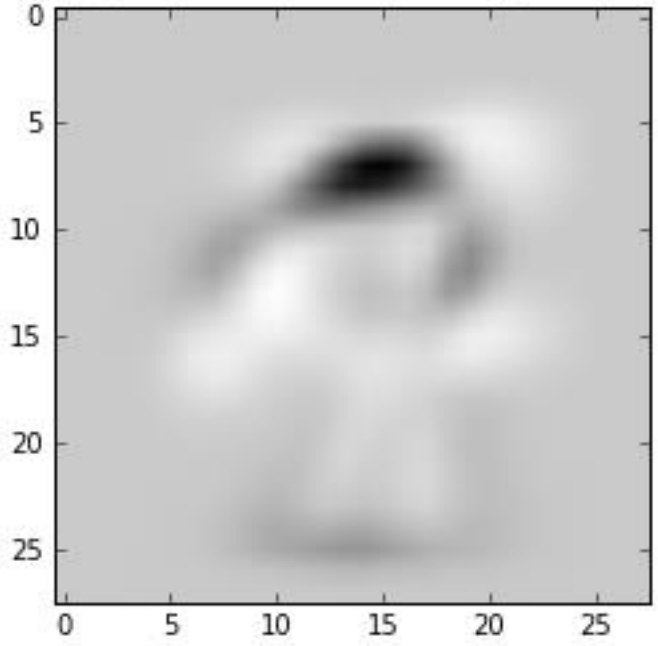
Omega_t treated as image, t = 1



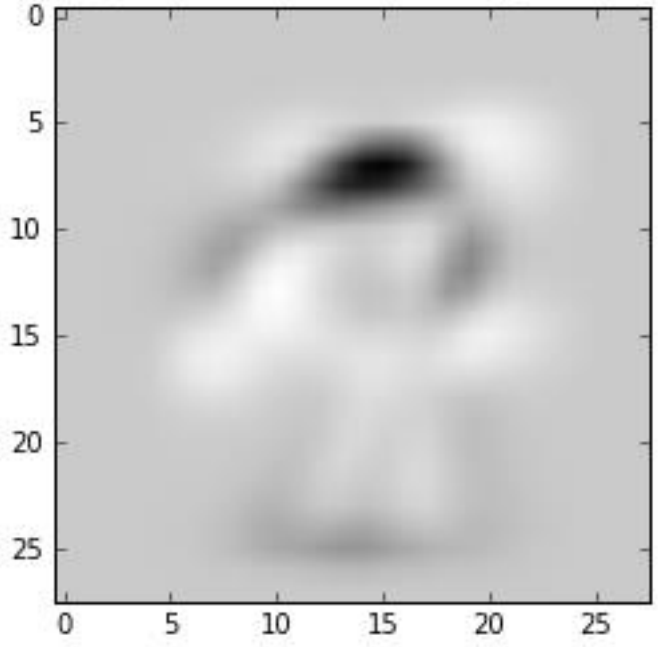
Omega_t treated as image, t = 5



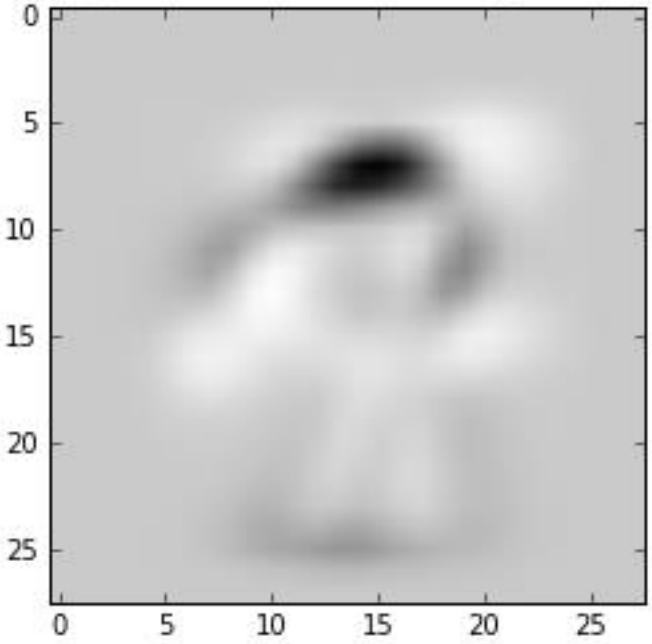
Omega_t treated as image, t = 10



Omega_t treated as image, t = 25



Omega_t treated as image, t = 50



Omega_t treated as image, t = 100