Bayesian Mod for Machine Learning.

Haoyang Chen hc2812 HW2

Prob 1. Denote $X = \{x_1, \cdots, x_N\}$, $Z = \{z_1, \cdots, z_N\}$

(a) $\boxed{\text{For E-step}}$ $q_t(Z) = p(Z|X, W_{t-1}) = \prod_{i=1}^{N} p(z_i|x_i, W_{t-1})$

$p(z_i|x_i, W_{t-1}) \propto p(x_i|z_i, W_{t-1}) \cdot p(z_i) = N(W_{t-1} z_i, \sigma^2 I) \cdot N(0, I) = N(z_i|\mu_i, \Sigma)$

where $\Sigma = (I + \frac{1}{\sigma^2} W_{t-1}^T W_{t-1})^{-1}$, $\mu_i = \frac{\Sigma \cdot W_{t-1}^T x_i}{\sigma^2}$

$L_t(W) = \int q_t(Z) \ln p(X, W, Z) dZ - \int q_t(Z) \ln q_t(Z) dZ$

$= E_{q_t(Z)}[\ln p(X, W, Z)] + const$ (Since $\int q_t(Z) \ln q_t(Z) dZ$ does not contain W, it's denoted as const)

$= E_{q_t(Z)}[\ln p(W) + \sum_{i=1}^{N} \ln p(z_i) + \sum_{i=1}^{N} \ln p(x_i|z_i, W)] + const$

$= \ln p(W) + \sum_{i=1}^{N} E_{q_t(z_i)}[\ln p(x_i|z_i, W)] + const$ (Since $E_{q_t(Z)}[\sum_{i}^{N} \ln p(z_i)]$ free from W, it's added into const. term)

$= -\frac{\lambda}{2} tr(W^T W) + \sum_{i=1}^{N} E_{q_t(z_i)}[-\frac{1}{2\sigma^2}(x_i - W z_i)^T (x_i - W z_i)] + const.$

$= -\frac{\lambda}{2} tr(W^T W) + \sum_{i=1}^{N} -\frac{1}{2\sigma^2} \{-2tr(\mu_i x_i^T W) + tr(W^T W (\mu_i \mu_i^T + \Sigma))\} + const.$

(Since $E_{q_t(z_i)}[z_i] = \mu_i \Rightarrow E_{q_t(z_i)}[z_i z_i^T] = \mu_i \mu_i^T + \Sigma$

$\Rightarrow E_{q_t(z_i)}[z_i^T W^T W z_i] = E_{q_t(z_i)}[tr(z_i^T W^T W z_i)] = tr(W^T W \cdot E_{q_t(z_i)}[z_i z_i^T])$)

(b) $\boxed{\text{For M-step}}$ $W_t = \arg\max_W L_t(W)$

$$\frac{\partial L_t(W)}{\partial W} = \frac{1}{\sigma^2} \sum_{i=1}^{N} x_i \mu_i^T - W\{\frac{1}{\sigma^2} \sum_{i=1}^{N} (\mu_i \mu_i^T + \Sigma) + \lambda I\} = 0$$

$$\therefore W_t = (\sum_{i=1}^{N} x_i \mu_i^T)(\sum_{i=1}^{N} \mu_i \mu_i^T + N \cdot \Sigma + \lambda \sigma^2 \cdot I)^{-1}$$

(c) $\ln p(X, W_t) = \int q_t(Z) \ln \frac{p(X, W_t, Z)}{q_t(Z)} dZ + \int q_t(Z) \ln \frac{q_t(Z)}{p(Z|X, W_t)} dZ = E_{q_t(Z)}[\ln p(X, Z, W_t) - \ln p(Z|X, W_t)]$

$= \ln p(W_t) + \sum_{i=1}^{N} E_{q_t(z_i)}[\ln p(x_i|z_i, W_t) + \ln p(z_i) - \ln q_t(z_i)]$

$= -\frac{\lambda}{2} tr(W_t^T W_t) + \frac{dk}{2} \ln \frac{\lambda}{2\pi} + \sum_{i=1}^{N}[-\frac{d}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} x_i^T x_i + \frac{1}{\sigma^2} tr(\mu_i x_i^T W_t) - \frac{1}{2\sigma^2} tr(W_t^T W_t (\mu_i \mu_i^T + \Sigma))]$

$+ \sum_{i=1}^{N}[-\frac{k}{2} \ln(2\pi) - \frac{1}{2} tr(\mu_i^T \mu_i + \Sigma)] + \sum_{i=1}^{N}[\frac{1}{2} \ln \det(2\pi\Sigma)]$

The EM algorithm pseudo-code:

1. Initialize W_0 as a $d \times k$ matrix with zeros

2. For $t = 1, \cdots, T$, do:

   (i) E-Step: Calculate $\Sigma = (I + \frac{1}{\sigma^2} W_{t-1}^T W_{t-1})^{-1}$, $\mu_i = \frac{\Sigma \cdot W_{t-1}^T x_i}{\sigma^2}$, $i = 1, \cdots, N$

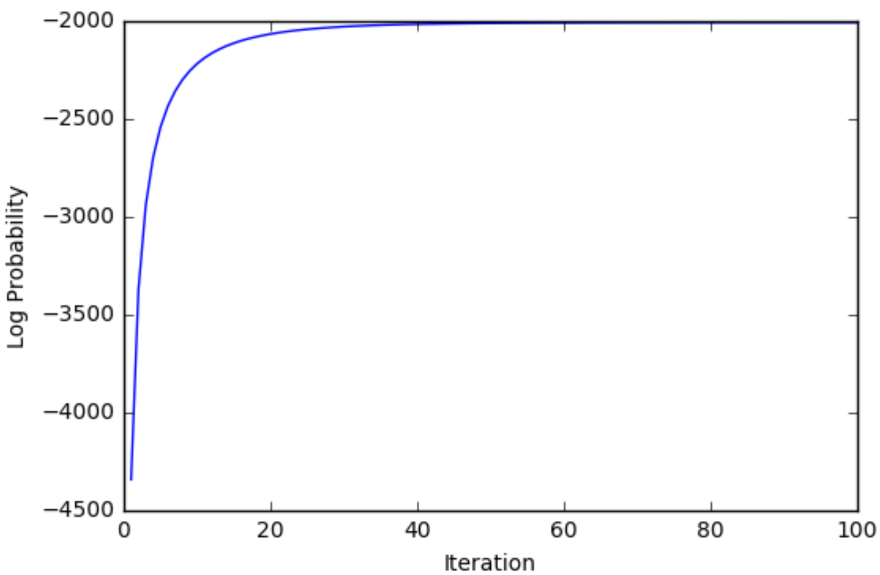
   (ii) M-Step. Calculate $W_t = (\sum_{i=1}^{N} x_i \mu_i^T)(\sum_{i=1}^{N} \mu_i \mu_i^T + N \cdot \Sigma + \lambda \sigma^2 \cdot I)^{-1}$

   (iii) Calculate $\ln p(X, W_t)$, if $\ln p(X, W_t) < \ln p(X, W_{t-1})$, stop.

2.

a). Shown in the code

b). Draw the plot:



c). The confusion matrix is shown below:

	Actual = 0	Actual = 1
Predicted = 0	930	52
Predicted = 1	77	932

Accuracy = 0.9352

d). 3 misclassified images

Image Index: 46

True label: 0

Predicted label: 1

Prob: 0.7819897

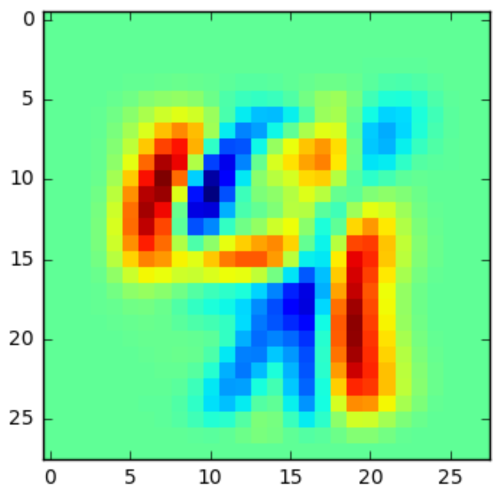


Image Index: 156

True label: 0

Predicted label: 1

Prob: 0.9310753

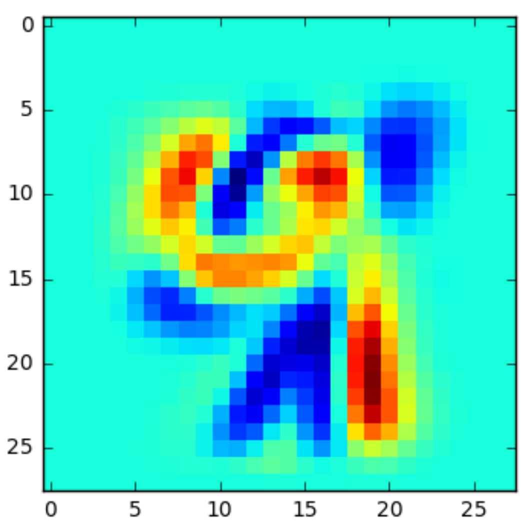
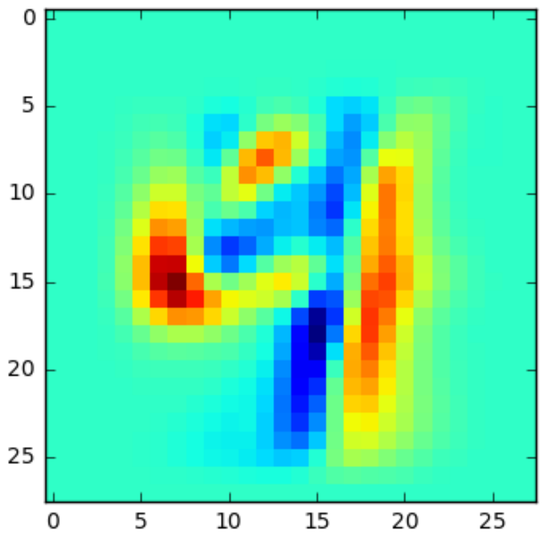


Image Index: 259
True label: 0
Predicted label: 1
Prob: 0.8018295



e). 3 most ambiguous predictions
Image Index: 586
True label: 0
Predicted label: 1
Prob: 0.5002605

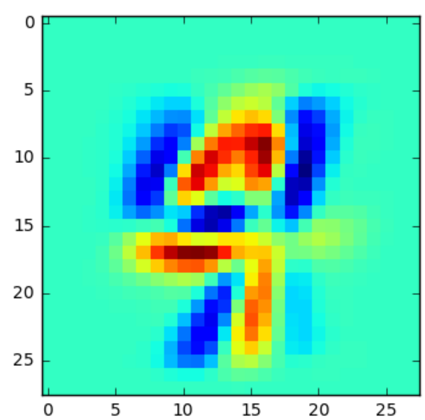


Image Index: 340
True label: 0
Predicted label: 1
Prob: 0.5046497

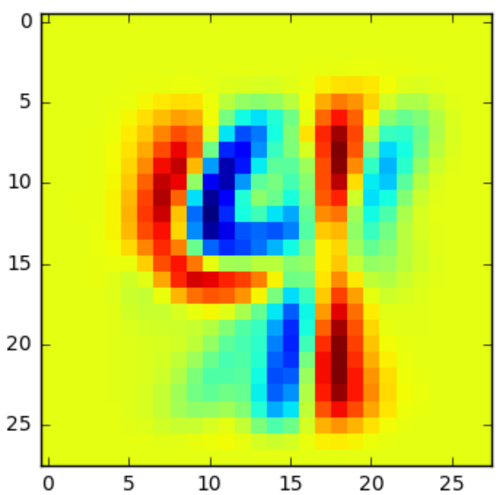
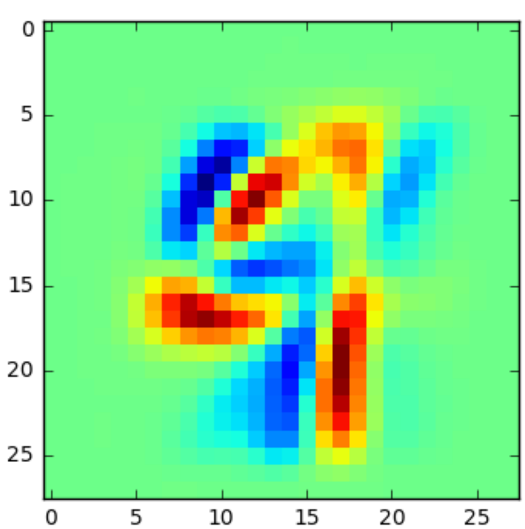


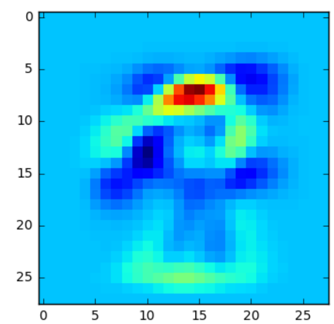
Image Index: 210
True label: 0
Predicted label: 1
Prob: 0.5061759



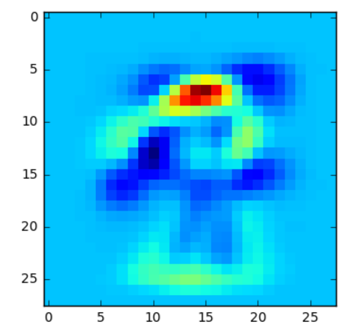
f). Reconstruct W

The weights images look like the number images. The difference between images becomes smaller when t gets larger. It means when t become larger, the convergence rate of weights would be slower. The plot in problem b also shows this fact.

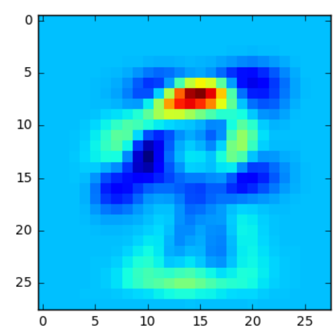
1



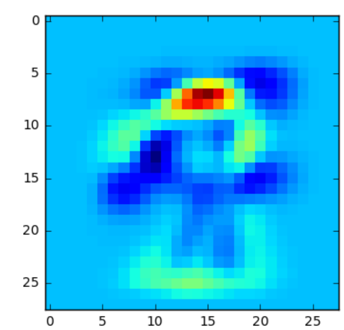
5



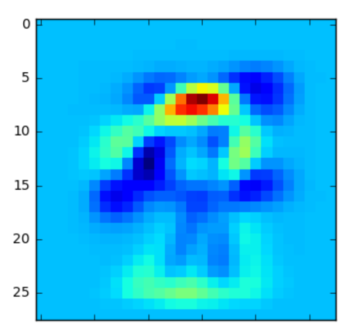
10



25



50



100