CSE 586A **Weijie Gan**

Problem Set IV

1. **Autoencoder (30%).** Train an autoencoder (AE) network (provided in Matlab) with aligned faces obtained from the PS3 question 1. Reconstruct the training data with different sizes of latent (hidden) layers to answer the following questions.

   No regularity in AE:

   1) Can you recover the original aligned faces with a full size of hidden layers? Plot the results.

   2) Set $1\%, 3\%, 10\%$ of hidden layer size (compare with original data dimension) to plot out the reconstruct face and report the reconstruction error.

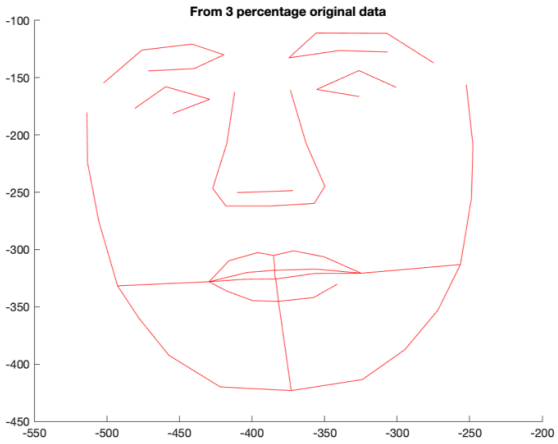
   With regularity in AE:

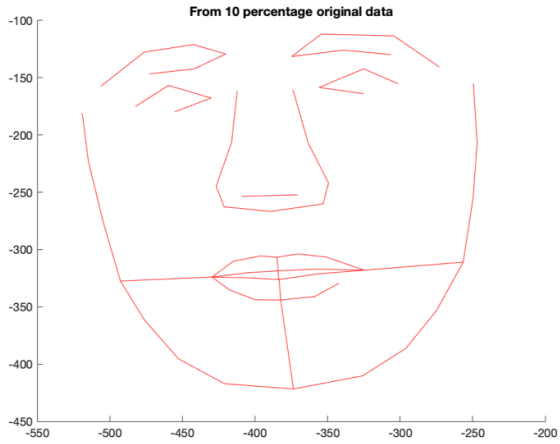
   3) Increase the weight w of L_2 regularity term in AE, plot out the reconstruction errors with different weights $w = \{0.1, 0.2, 0.3, 0.4, 0.5\}$.

1.

(a) From the experiment result, with a full size of hidden layers, the reconstruction error will be 0.0049. The error is so low that we can consider the reconstructed face as a original face. The plot shows the first reconstructed face.

(b) The following plot shows the first reconstructed face with different hidden layer size.

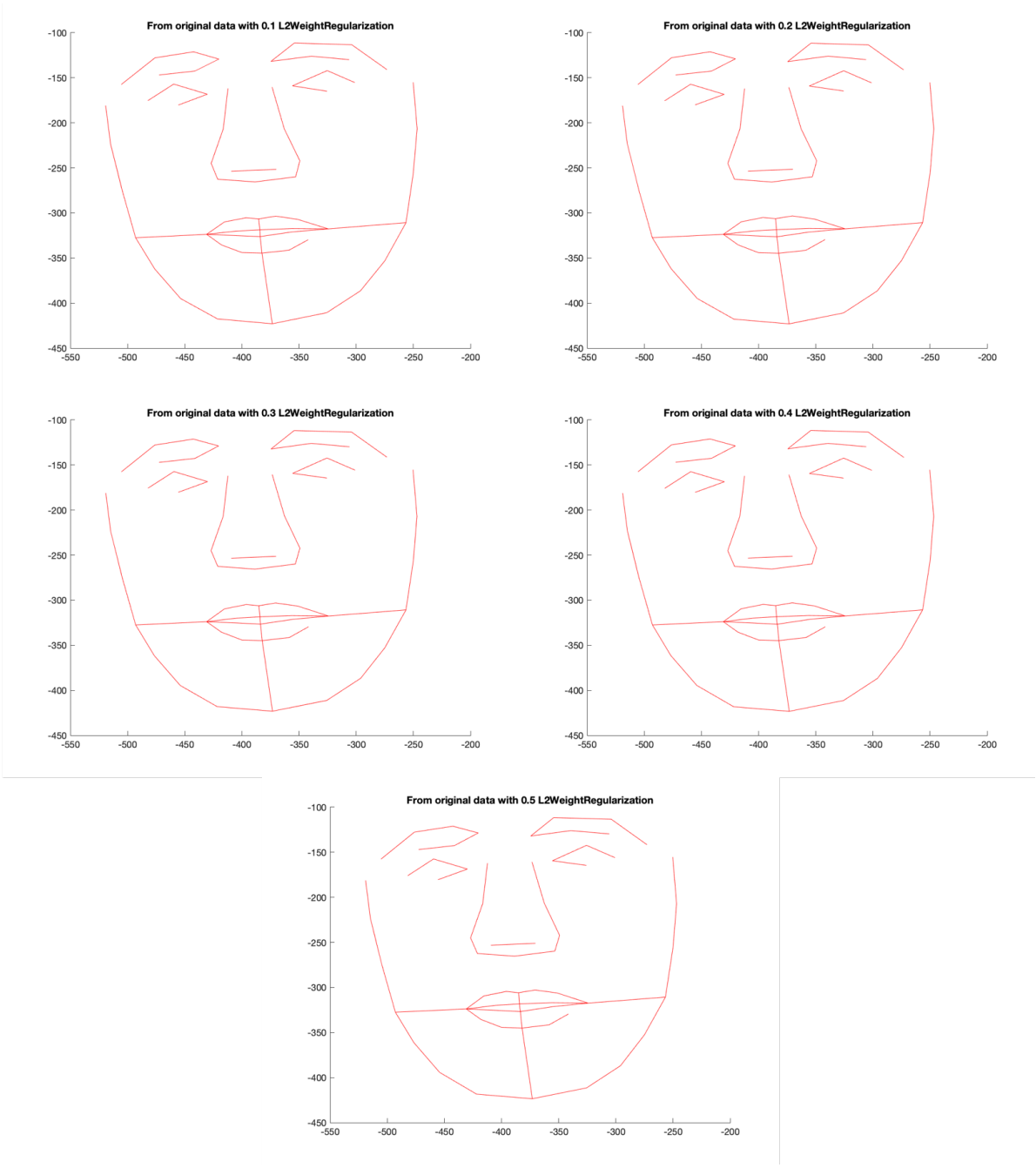




The following table shows the reconstruction errors.

Hidden layer size	Reconstruction error(mse)
full size	0.0049
1%	8.8065
4%	2.3797
10%	0.2360

(c) The following plot shows the first reconstructed face with different the weight w of L2 regularity term in Autoencoder.



The following table shows the reconstruction errors.

The weight w of L2 regularity term	Reconstruction error(mse)
0.1	0.1441
0.2	0.2755
0.3	0.4001
0.4	0.5060
0.5	0.6168

2. **Regression (70%).** Given data (X, Y) with $X \in \mathbb{R}^d$ and $Y \in \{0, 1\}$, our goal is to train a classifier that will predict an unknown class label $\tilde{y}$ from a new data point $\tilde{x}$. Consider the following model:

$$Y \sim \text{Ber}\left(\frac{1}{1+e^{-X^T\beta}}\right),$$
$$\beta \sim \text{N}(0, \sigma^2 I).$$

This is a **Bayesian logistic regression** model. Your goal is to derive and implement a MAP (maximum a posterior) Bayesian inference on β.

(a) Write down the formula for the unormalized posterior of $\beta \,|\, Y$, i.e.,

$$p(\beta \,|\, y; x, \sigma) \propto \prod_{i=1}^{n} p(y_i \,|\, \beta; x_i) p(\beta; \sigma)$$

(b) Show that this posterior is proportional to $\exp(-U(\beta))$, where

$$U(\beta) = \sum_{i=1}^{n} (1 - y_i) x_i^T \beta + \log(1 + e^{-x_i^T \beta}) + \frac{1}{2\sigma^2} \|\beta\|^2.$$

(c) Implement MAP to infer β.

(d) Use your code to analyze the `iris` data (provided in txt file), looking only at two species, *versicolor* and *virginica*. The species labels are your Y data, and the four features, petal length and width, sepal length and width, are your X data. Also, add a constant term, i.e., a column of 1's to your X matrix. Use the first 30 rows for each species as training data and leave out the last 20 rows for each species as test data (for a total of 60 training and 40 testing). Use the estimated β to get a prediction, $\tilde{y}$, of the class labels for the test data.

(e) Compare this to the true class labels, y, and see how well you did by estimating the average error rate, $\text{E}[|y - \tilde{y}|]$ (a.k.a. the zero-one loss). What values of σ, ϵ, and L did you use?

2.

(a) In Bayesian Logistic regression model, we can suppose Y as a Bernoulli Distribution, as shown as:

$$\begin{aligned} Y &= \prod_{i=1}^{n} p(y_i|\beta; x_i) = \text{Ber}\left(\frac{1}{1+e^{(-X^T\beta)}}\right) \\ &= \prod_{i=1}^{n} \left(\frac{1}{1+e^{(-x_i^T\beta)}}\right)^{y_i} \left(1 - \frac{1}{1+e^{(-x_i^T\beta)}}\right)^{1-y_i} \end{aligned} \tag{1}$$

Then, the we consider the prior function about β as a Gaussian Distribution:

$$\begin{aligned} \beta = p(\beta; \sigma) &= \text{N}(0, \sigma^2 I) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\beta^2}{2\sigma^2}\right) \end{aligned} \tag{2}$$

So, the formulation of posterior will be:

$$\begin{aligned} p(\beta|y; x; \sigma) &= \prod_{i=1}^{n} p(y_i|\beta; x_i) p(\beta; \sigma) \\ &= \prod_{i=1}^{n} \left(\frac{1}{1+e^{(-x_i^T\beta)}}\right)^{y_i} \left(1 - \frac{1}{1+e^{(-x_i^T\beta)}}\right)^{1-y_i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\beta^2}{2\sigma^2}\right) \end{aligned} \tag{3}$$

(b) Based on the above equation of posterior, its log-equation will be:

$$
\begin{aligned}
p(\beta|y;x;\sigma) &= \prod_{i=1}^{n}\left(\frac{1}{1+e^{(-x_i^T\beta)}}\right)^{y_i}\left(1-\frac{1}{1+e^{(-x_i^T\beta)}}\right)^{1-y_i}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\beta^2}{2\sigma^2}\right)\\
\sqrt{2\pi\sigma^2}\ p(\beta|y;x;\sigma) &= \sum_{i=1}^{n}\left(\frac{1}{1+e^{(-x_i^T\beta)}}\right)^{y_i}\left(1-\frac{1}{1+e^{(-x_i^T\beta)}}\right)^{1-y_i}\exp\left(-\frac{\beta^2}{2\sigma^2}\right)\\
log(\sqrt{2\pi\sigma^2}\ p(\beta|y;x;\sigma)) &= \sum_{i=1}^{n}\ log\left(\left(\frac{1}{1+e^{(-x_i^T\beta)}}\right)^{y_i}\left(1-\frac{1}{1+e^{(-x_i^T\beta)}}\right)^{1-y_i}\exp\left(-\frac{\beta^2}{2\sigma^2}\right)\right)\\
log(\sqrt{2\pi\sigma^2}\ p(\beta|y;x;\sigma)) &= \sum_{i=1}^{n}\ y_i log\left(\frac{1}{1+e^{(-x_i^T\beta)}}\right)+(1-y_i)log\left(1-\frac{1}{1+e^{(-x_i^T\beta)}}\right)-\frac{\beta^2}{2\sigma^2}\\
log(\sqrt{2\pi\sigma^2}\ p(\beta|y;x;\sigma)) &= \sum_{i=1}^{n}\ -y_i log\big(1+e^{(-x_i^T\beta)}\big)+(1-y_i)log\left(\frac{e^{(-x_i^T\beta)}}{1+e^{(-x_i^T\beta)}}\right)-\frac{\beta^2}{2\sigma^2}\\
log(\sqrt{2\pi\sigma^2}\ p(\beta|y;x;\sigma)) &= \sum_{i=1}^{n}\ -y_i log\big(1+e^{(-x_i^T\beta)}\big)+(1-y_i)log\big(e^{(-x_i^T\beta)}\big)-\\
&\qquad (1-y_i)log\big(1+e^{(-x_i^T\beta)}\big)-\frac{\beta^2}{2\sigma^2}\\
log(\sqrt{2\pi\sigma^2}\ p(\beta|y;x;\sigma)) &= \sum_{i=1}^{n}\ -(1-y_i)(x_i^T\beta)-log\big(1+e^{(-x_i^T\beta)}\big)-\frac{\beta^2}{2\sigma^2}\\
log(\sqrt{2\pi\sigma^2}\ p(\beta|y;x;\sigma)) &= -U(\beta)\\
\sqrt{2\pi\sigma^2}\ p(\beta|y;x;\sigma) &= \exp(-U(\beta))
\end{aligned}
\tag{4}
$$

(c) From above equation, we found the formulation of posterior should be:

$$
log(\sqrt{2\pi\sigma^2}\ p(\beta|y;x;\sigma)) = \sum_{i=1}^{n}\ -(1-y_i)(x_i^T\beta)-log\big(1+e^{(-x_i^T\beta)}\big)-\frac{\beta^2}{2\sigma^2} \tag{5}
$$

To maximize the posterior(MAP):

$$
\begin{aligned}
\max\ p(\beta|y;x;\sigma) &= \max\ log(\sqrt{2\pi\sigma^2}\ p(\beta|y;x;\sigma))\\
&= \max\ \sum_{i=1}^{n}\ -(1-y_i)(x_i^T\beta)-log\big(1+e^{(-x_i^T\beta)}\big)-\frac{\beta^2}{2\sigma^2}\\
&= \min\ \sum_{i=1}^{n}\ (1-y_i)(x_i^T\beta)+log\big(1+e^{(-x_i^T\beta)}\big)+\frac{\beta^2}{2\sigma^2}
\end{aligned}
\tag{6}
$$

Solving map by gradient descent, we need to compute first-order derivative first as:

$$
\frac{\partial U(\beta)}{\partial\beta} = \sum_{i=1}^{n}\ (1-y_i)(x_i)+\frac{e^{(-x_i^T\beta)}}{1+e^{(-x_i^T\beta)}}+\frac{1}{\sigma^2}\beta \tag{7}
$$

The implement can be seen in my Matlab code.

(d) After gradient descent discussed above and obtaining value of β, I predict the label using equation:

$$\hat{y}_i = \begin{cases} 0, \text{ if } 1/(1+\exp(-x_i^T\beta)) >= 0.5 \\ 1, \text{ if } 1/(1+\exp(-x_i^T\beta)) < 0.5 \end{cases}$$

The concrete process can be seen in my Matlab code.

(e) With $\sigma = 200$ and $\xi = 0.0015$, we can get a logistic regression model with 0.5 average error rate using zero-one loss, $E||y - \hat{y}||$.

I understand the error is not good enough, but I try my code in other dataset and it works perfectly.