CSC411H1 L0101, Fall 2018
Assignment 5

Name: JingXi, Hao
Student Number: 1000654188

Due Date: 14th November, 2018 11:59pm

1. **Gaussian Discriminant Analysis.** For this question you will build classifiers to label images of handwritten digits. Each image is 8 by 8 pixels and is represented as a vector of dimension 64 by listing all the pixel values in raster scan order. The images are grayscale and the pixel values are between 0 and 1. The labels y are 0, 1, 2, ..., 9 corresponding to which character was written in the image. There are 700 training cases and 400 test cases for each digit; they can be found in *a2digits.zip*.

   Starter code is provided to help you load the data (*data.py*). A skeleton (*q1.py*) is also provided for each question that you should use to structure your code.

   Using maximum likelihood, fit a set of 10 class-conditional Gaussians with a separate, full covariance matrix for each class. Remember that the conditional multivariate Gaussian probability density is given by,

$$p(\mathbf{x}|y = k, \mu, \Sigma_k) = (2\pi)^{(-d/2)}|\Sigma_k|^{-1/2} exp\{-\frac{1}{2}(\mathbf{x} - \mu_k)^T \Sigma_k (\mathbf{x} - \mu_k)\} \quad (1)$$

   You should take $p(y = k) = \frac{1}{10}$. You will compute parameters μ_{kj} and Σ_k for $k \in (0 \dots 9)$, $j \in (1 \dots 64)$. You should implement the covariance computation yourself (i.e. without the aid of '*np.cov*'). *Hint: To ensure numerical stability you may have to add a small multiple of the identity to each covariance matrix. For this assignment you should add 0:01* $\boldsymbol{I}$ *to each matrix.*

   (a) Using the parameters you fit on the training set and Bayes rule, compute the average conditional log-likelihood, i.e. $\frac{1}{N}\Sigma_{i=1}^{N} logp(y^{(i)}|\mathbf{x}^{(i)}, \theta)$ on both the train and test set and report it.

   Solution:

   Please see the detailed code implementation for this question in the file, ***q1.py***.

   Then, we show the average conditional log-likelihood computed on both the train and test set below.

```
The average conditional log-likelihood on train set is -0.12462443666862928.

The average conditional log-likelihood on test set is -0.1966732032552546.
```

(b) Select the most likely posterior class for each training and test data point as your prediction, and report your accuracy on the train and test set.

Solution:

Please see the detailed code implementation for this question in the file, ***q1.py***.

Then, we report the accuracy on the train and test set below.

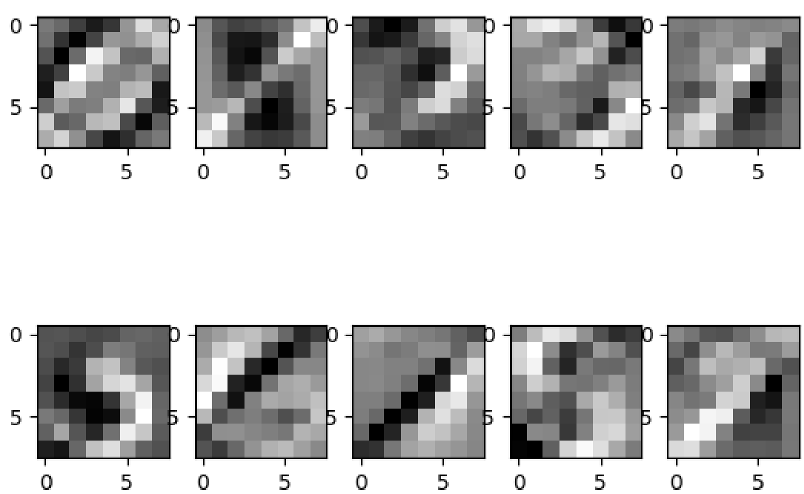
```
The accuracy on train set is 0.9814285714285714.

The accuracy on test set is 0.97275.
```

(c) Compute the leading eigenvectors (largest eigenvalue) for each class covariance matrix (can use *np.linalg.eig*) and plot them side by side as 8 by 8 images.

Solution:

Please see the detailed code implementation for this question in the file, ***q1.py***.

Then, we show the plot below, with plotting the image for each digit side by side in the dimensions of 8 by 8. Note that the order of the image is from top to bottom and from left to right.



2. **Categorial Distribution.** Let's consider fitting the categorical distribution, which is a discrete distribution over K outcomes, which we'll number 1 through K. The probability of each category is

explicitly represented with parameter θ_k. For it to be a valid probability distribution, we clearly need $\theta_k \geq 0$ and $\Sigma_k \theta_k = 1$. We'll represent each observation x as a 1-of-K encoding, i.e, a vector where one of the entries is 1 and the rest are 0. Under this model, the probability of an observation can be written in the following form:

$$p(\mathrm{x}; \boldsymbol{\theta}) = \prod_{k=1}^{K} {\theta_k}^{x_k} \tag{2}$$

Denote the count for outcome k as N_k, and the total number of observations as N. In the previous assignment, you showed that the maximum likelihood estimate for the counts was:

$$\hat{\theta_k} = \frac{N_k}{N} \tag{3}$$

Now let's derive the Bayesian parameter estimate.

(a) For the prior, we'll use the Dirichlet distribution, which is defined over the set of probability vectors (i.e. vectors that are non-negative and whose entries sum to 1). Its PDF is as follows:

$$p(\boldsymbol{\theta}) \propto {\theta_1}^{a_1 - 1} \cdots {\theta_K}^{a_K - 1} \tag{4}$$

A useful fact is that if $\boldsymbol{\theta} \sim Dirichlet(a_1, \ \ldots, \ a_K)$, then

$$\mathbb{E}[\theta_k] = \frac{a_k}{\Sigma_{k'} a_{k'}} \tag{5}$$

Determine the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$, where $\mathcal{D}$ is the set of observations. From that, determine the posterior predictive probability that the next outcome will be k.

Solution:

First, we determine the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$, where $\mathcal{D}$ is the set of observations. By applying the Bayes' rule, we obtain that $p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}$. Since we try to determine the mostly likely $\boldsymbol{\theta}$, thus, we need not to compute the denominator. Hence, we have that

$$\begin{aligned}
p(\boldsymbol{\theta}|\mathcal{D}) &= \tfrac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})} \\
&\propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\
&= \textstyle\prod_{i=1}^{N} p(\mathrm{x}^{(i)}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \qquad \text{\# Since } \mathcal{D} \text{ is the set of observations} \\
&= \textstyle\prod_{i=1}^{N} \prod_{k=1}^{K} {\theta_k}^{\mathrm{x}_k^{(i)}} p(\boldsymbol{\theta}) \qquad \text{\# Substitute } p(\mathrm{x}; \boldsymbol{\theta}) = \textstyle\prod_{k=1}^{K} {\theta_k}^{x_k} \\
&= \textstyle\prod_{k=1}^{K} \prod_{i=1}^{N} {\theta_k}^{\mathrm{x}_k^{(i)}} p(\boldsymbol{\theta}) \\
&= \textstyle\prod_{k=1}^{K} {\theta_k}^{N_k} p(\boldsymbol{\theta}) \\
&\propto (\textstyle\prod_{k=1}^{K} {\theta_k}^{N_k})({\theta_1}^{a_1 - 1} \cdots {\theta_K}^{a_K - 1}) \qquad \text{\# Substitute } p(\boldsymbol{\theta}) \propto {\theta_1}^{a_1 - 1} \cdots {\theta_K}^{a_K - 1} \\
&= (\textstyle\prod_{k=1}^{K} {\theta_k}^{N_k})(\prod_{k=1}^{K} {\theta_k}^{a_k - 1}) \\
&= \textstyle\prod_{k=1}^{K} {\theta_k}^{N_k} {\theta_k}^{a_k - 1} \\
&= \textstyle\prod_{k=1}^{K} {\theta_k}^{N_k + a_k - 1}
\end{aligned}$$

Therefore, we obtain that $\boldsymbol{\theta}$ follows the Dirichlet distribution, $\boldsymbol{\theta} \sim Dirichlet(N_1 + a_1, \ \ldots, \ N_K + a_K)$. Since the posterior predictive probability that the next outcome will be k is the expectation value of θ_k, therefore, the posterior predictive probability that the next outcome will be k is $\mathbb{E}(\theta_k) = \frac{N_k+a_k}{\Sigma_{k'} N_{k'}+a_{k'}}$.

(b) Still assuming the Dirichlet prior distribution, determine the MAP estimate of the parameter vector $\boldsymbol{\theta}$. For this question, you may assume each $a_k > 1$.

Solution:

For this question, we need to determine the MAP estimate of the parameter of vector $\boldsymbol{\theta}$. This means that we need to maximize $p(\boldsymbol{\theta}|\mathcal{D})$ subject to $\Sigma_k \theta_k = 1$, which means that we need to maximize $logp(\boldsymbol{\theta}|\mathcal{D})$ subject to $\Sigma_k \theta_k = 1$. Since we try to determine the mostly likely $\boldsymbol{\theta}$, therefore, we need not to compute the denominator for $p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}$. Therefore, we can write this as $p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Then, we need to maximize $log[p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})]$ subject to $\Sigma_k \theta_k = 1$. Let $f = log[p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})]$ and $g = \Sigma_k \theta_k - 1$. Thus, there exists a Lagrange multiplier λ such that $L(\boldsymbol{\theta}, \mathrm{x}, \lambda) = f - \lambda g = log[p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})] - \lambda(\Sigma_k \theta_k - 1)$, where $\frac{\partial L}{\partial \theta_k} = 0$. Therefore, we take the partial derivative of $L(\boldsymbol{\theta}, \mathrm{x}, \lambda)$ with respect to θ_k. Then, we have that

$$
\begin{aligned}
\frac{\partial L}{\partial \theta_k} &= \frac{\partial [log[p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})] - \lambda(\Sigma_k \theta_k - 1)]}{\partial \theta_k} \\
&= \frac{\partial [log(\prod_{k=1}^{K} {\theta_k}^{N_k+a_k-1}) - \lambda(\Sigma_k \theta_k - 1)]}{\partial \theta_k} \qquad \text{\# From 2(a), we have that } p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = \prod_{k=1}^{K} {\theta_k}^{N_k+a_k-1} \\
&= \frac{\partial [\Sigma_{k=1}^{K}(N_k+a_k-1)log(\theta_k) - \lambda(\Sigma_k \theta_k - 1)]}{\partial \theta_k} \\
&= \frac{N_k+a_k-1}{\theta_k} - \lambda
\end{aligned}
$$

Let the partial derivative equals to 0. Then, we have that

$$\frac{\partial L}{\partial \theta_k} = 0$$

$$\frac{N_k + a_k - 1}{\theta_k} - \lambda = 0$$

$$\frac{N_k + a_k - 1}{\theta_k} = \lambda$$

$$\theta_k = \frac{N_k + a_k - 1}{\lambda}$$

Since $\Sigma_k \theta_k = 1$, then by substitution of θ_k found above, we obtain that

$$\Sigma_k \theta_k = 1$$

$$\Sigma_{k=1}^{K} \frac{N_k + a_k - 1}{\lambda} = 1$$

$$\frac{1}{\lambda} \Sigma_{k=1}^{K} N_k + a_k - 1 = 1$$

$$\lambda = \Sigma_{k=1}^{K} N_k + a_k - 1$$

Then, substitute $\lambda = \Sigma_{k=1}^{K} N_k + a_k - 1$ into the equatio of $\theta_k = \frac{N_k + a_k - 1}{\lambda}$. We have that

$$\theta_k = \frac{N_k + a_k - 1}{\Sigma_{k'=1}^{K} N_{k'} + a_{k'} - 1} = \frac{N_k + a_k - 1}{N - K + \Sigma_{k'=1}^{K} a_{k'}}$$

.

Hence, we are able to find every entry of the vector $\boldsymbol{\theta}$, which indicates that we are able to find an estimation of vector $\boldsymbol{\theta}$ that maximizes $p(\boldsymbol{\theta}|\mathcal{D})$ subject to $\Sigma_k \theta_k = 1$.