CSC311 Assignment3

Peiqing Yu 2020/11/21 1003769004

Q1(a).

```
Computing the log-sum-exp of a=[-100000 -100000 -100000]
    Unstable: -inf
    Stable: -99998.90138771133
Computing the log-sum-exp of b=[100000 100000 100000]
    Unstable: inf
    Stable: 100001.09861228867
```

Q1(b).

Proof: Let $s = \max_{j=0} \{a_j\}$, let $y = \log\left(\sum_{i=0}^{k} \exp(a_i)\right)$

Then $$\text{LHS} = \log\left(\sum_{i=0}^{k} \exp(a_i)\right) = y$$

$$\sum_{i=0}^{k} \exp(a_i) = e^{y}$$

$$e^{-s} \sum_{i=0}^{k} \exp(a_i) = e^{y} \cdot e^{-s}$$

$$\sum_{i=0}^{k} \exp(a_i - s) = e^{y-s}$$

$$y - s = \log\left(\sum_{i=0}^{k} \exp(a_i - s)\right)$$

$$y = \log\left(\sum_{i=0}^{k} \exp(a_i - s)\right) + s$$

$$= \text{RHS} \quad \blacksquare$$

This numerical stable version is more robust to underflow or overflow, because we shift the values in the exponent by the maximum value of log(p(x, i)). Therefore, if we have a[i] very large, we can use the maximum value to adjust it and have the largest value be exp(0) which prevents overflow. For each value of a is small, we can also use the maximum value to adjust it and prevent underflow.

Q2(a).

```
The average conditional log-likelihood on train set is -0.12462443666863023.
The average conditional log-likelihood on test set is -0.1966732032552555.
```

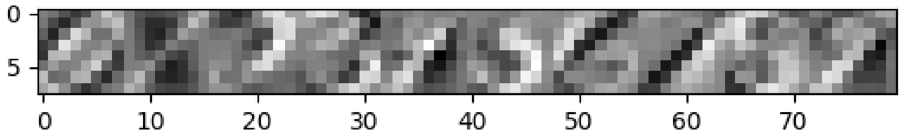
Q2(b).

```
The accuracy on train set is 0.9814285714285714.
The accuracy on test set is 0.97275.
```

The code to compute the accuracy is:

```
# find the accuracy on train and test datasets
train_acc = np.mean(classify_data(train_data, means, covariances) == train_labels)
test_acc = np.mean(classify_data(test_data, means, covariances) == test_labels)
```

Q2(c). The leading eigenvectors for each covariance matrix plotted side by side are:



Q3(a).

$$p(\theta|D) = \frac{P(D|\theta)\,P(\theta)}{P(D)} \propto p(D|\theta)P(\theta)$$

$$\propto \left(\prod_{i=1}^{n}\prod_{k=1}^{K} \theta_k^{x_k^{(i)}}\right)\left(\theta_1^{a_1-1}\,\theta_2^{a_2-1}\cdots\,\theta_K^{a_K-1}\right)$$

$$\propto \left(\theta_1^{x_1^{(1)}}\,\theta_1^{x_1^{(2)}}\cdots\,\theta_1^{x_1^{(n)}}\,\theta_1^{a_1-1}\right)\cdots\cdots\left(\theta_K^{x_K^{(1)}}\,\theta_K^{x_K^{(2)}}\cdots\,\theta_K^{x_K^{(n)}}\,\theta_K^{a_K-1}\right)$$

$$\propto \theta_1^{\sum_{i=1}^{n} x_1^{(i)}+a_1-1} \;\cdots\cdots\; \theta_K^{\sum_{i=1}^{n} x_K^{(i)}+a_K-1}$$

By examing this, we can identify the posterior distribution $P(\theta|D)$ follows dirichlet distribution with parameters $b_1, b_2, \cdots, b_K$ where $b_k = \sum_{i=1}^{n} x_k^{(i)} + a_k = N_k + a_k$

Q3(b).

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\; p(\theta|D)$$

$$= \underset{\theta}{\operatorname{argmax}}\; p(\theta)p(D|\theta)$$

$$= \underset{\theta}{\operatorname{argmax}}\; \log(P(\theta)p(D|\theta))$$

$$l(\theta) = \log(P(\theta)p(D|\theta)) = \text{constant} + (N_1+a_1-1)\log(\theta_1) + \cdots + (N_K+a_K-1)\log(\theta_K)$$

since $\sum_{i=1}^{K}\theta_i = 1$, then $\theta_K = 1-\sum_{i=1}^{K-1}\theta_i$

$$l(\theta) = \text{constant} + \left(\sum_{i=1}^{K-1}(N_i+a_i-1)\log(\theta_i)\right) + (N_K+a_K-1)\log\left(1-\sum_{i=1}^{K-1}\theta_i\right)$$

$$\frac{\partial l(\theta)}{\partial\theta_i} = \frac{N_i+a_i-1}{\theta_i} - \frac{N_K+a_K-1}{1-\sum_{i=1}^{K-1}\theta_i}$$

$$\frac{\partial l(\theta)}{\partial\theta_i} = \frac{N_i+a_i-1}{\theta_i} - \frac{N_K+a_K-1}{\theta_K} = 0$$

I

$$\frac{\theta_i}{\theta_k} = \frac{N_i + a_i - 1}{N_k + a_k - 1}$$

We assure that $\frac{\hat{\theta_i}}{\hat{\theta_k}}$ is maximized by checking l''

since we have $\frac{\hat{\theta_1}}{\hat{\theta_k}} + \frac{\hat{\theta_2}}{\hat{\theta_k}} + \cdots + \frac{\hat{\theta_k}}{\hat{\theta_k}} = \frac{\sum_{i=1}^{k} \hat{\theta_i}}{\hat{\theta_k}} = \frac{1}{\hat{\theta_k}}$

$$\frac{N_1 + a_1 - 1}{N_k + a_k - 1} + \frac{N_2 + a_2 - 1}{N_k + a_k - 1} + \cdots + \frac{N_k + a_k - 1}{N_k + a_k - 1} = \frac{1}{\hat{\theta_k}}$$

$$\frac{\left(\sum_{i=1}^{k} N_i + a_i\right) - k}{N_k + a_k - 1} = \frac{1}{\hat{\theta_k}}$$

$$\hat{\theta_k} = \frac{N_k + a_k - 1}{\left(\sum_{i=1}^{k} N_i - a_i\right) - k}$$

Thus the j th entry of $\hat{\theta}_{MAP}$ is $\frac{N_j + a_j - 1}{\left(\sum_{i=1}^{k} N_i + a_i\right) - k}$, where $j = 1, 2, \cdots, k$

Q3(c).

We know $P(x^{(N+1)} | D) = \int P(x^{(N+1)} | \theta) P(\theta | D) \, d\theta$;

suppose $x^{(N+1)}{}_k = 1$.

Then $\int P(x^{(N+1)} | \theta) P(\theta | D) \, d\theta$

$= \int \theta_k^{x_k^{(N+1)}} P(\theta | D) \, d\theta$

$= \int \theta_k P(\theta | D) \, d\theta$.

$= E(\theta_k | D)$ # which follows distribution of Q3(a)

$= \frac{N_k + a_k}{\sum_{i=1}^{k} N_i + a_i}$