Brendan Neal 1001160236|nealbre1 2018-11-12

CSC411 Homework 5

Question 1: Gaussian Discriminant Analysis

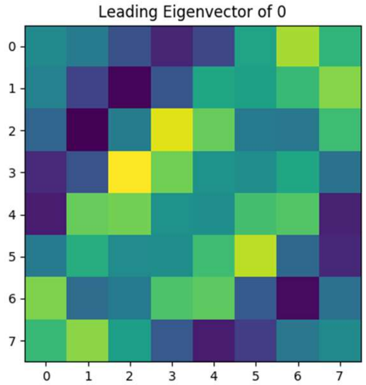
Part A, B: Average Conditional Log Likelihood, Accuracy

Output of program:

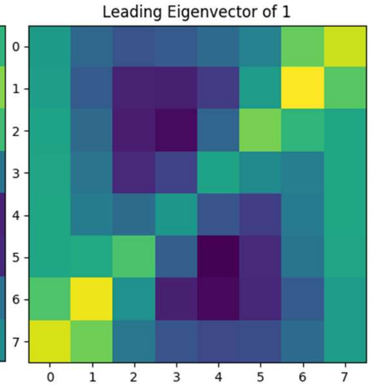
```
Part A: Average Conditional Log-Likelihoods

Train: -0.12462443666863048

Test: -0.19667320325525584


Part B: Accuracy

Train Accuracy: 0.9814285714285714

Test Accuracy: 0.97275
```

Part C: Visualizing Leading Covariance Eigenvectors

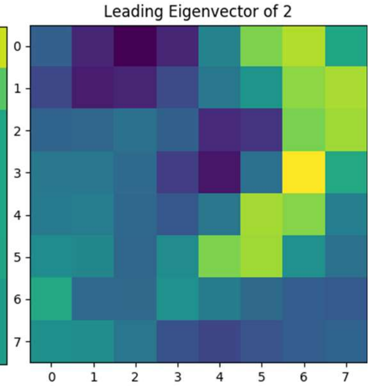
Leading Eigenvector of 0



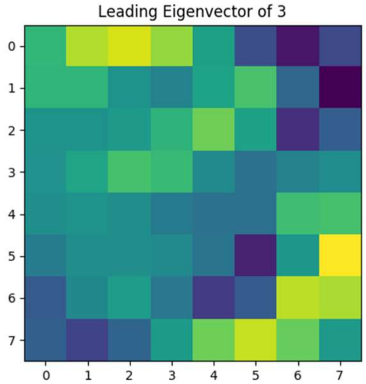
Leading Eigenvector of 1



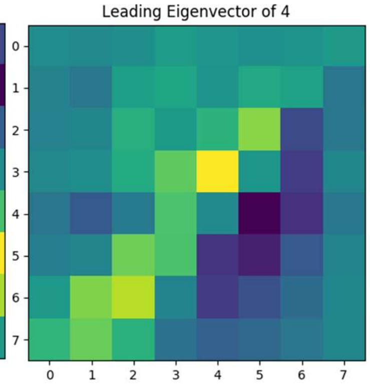
Leading Eigenvector of 2



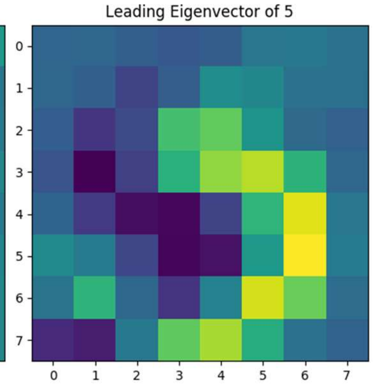
Leading Eigenvector of 3



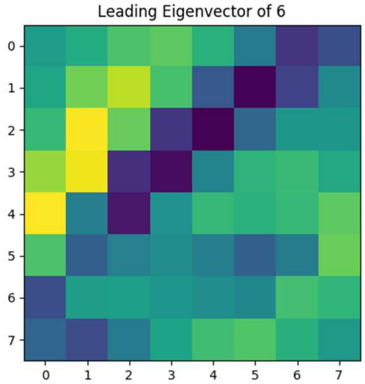
Leading Eigenvector of 4



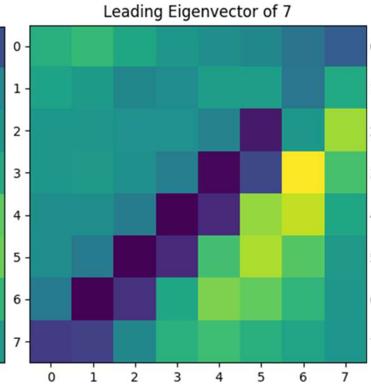
Leading Eigenvector of 5



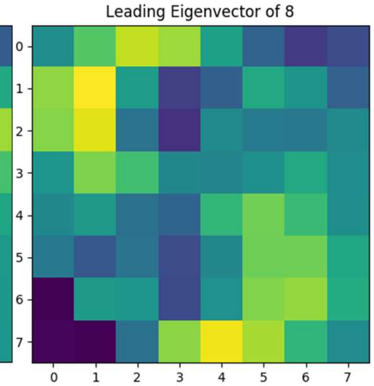
Leading Eigenvector of 6



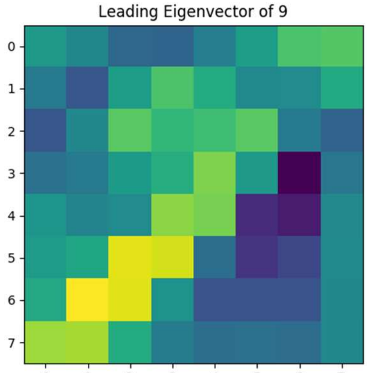
Leading Eigenvector of 7



Leading Eigenvector of 8



Leading Eigenvector of 9



Question 2

Part A: Posterior Distribution

Bayes rule:

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

Dirichlet as prior:

$$p(\theta) = \frac{\Gamma(\sum_{i=1}^{K} a_i)}{\prod_{i=1}^{K} \Gamma(a_i)} \prod_{i=1}^{K} \theta_i^{a_i - 1}$$

Likelihood of data $x_{1\ldots N}$ given θ

$$p(D|\theta) = \prod_{i=1}^{N} \prod_{j=1}^{K} \theta_j^{x_j^{(i)}}$$

Note that x is a one-hot vector. This means that all elements of θ except for the element corresponding to the correct class will "cancel out" of the product. If class j has N_j occurences, then the total power of θ_j will be N_j

$$p(D|\theta) = \prod_{j=1}^{K} \theta_j^{N_j}$$

For the purposes of this question, we can reason about $p(\theta|D)$ in proportional terms:

$$p(\theta|D) \ \alpha \ p(D|\theta)p(\theta)$$

$$p(\theta) \ \alpha \prod_{i=1}^{K} \theta_i^{a_i - 1}$$

$$\Rightarrow p(\theta|D) \ \alpha \prod_{j=1}^{K} \theta_j^{N_j} \prod_{i=1}^{K} \theta_i^{a_i - 1}$$

Merging the two products together:

$$\Rightarrow \ p(\theta|D) \ \ \alpha \ \prod_{i=1}^{K} \theta_i^{N_i + a_i - 1}$$

Note that the above proportionality means that the conditional probability also obeys a Dirichlet distribution

$$\Rightarrow p(\theta|D) \sim Dirichlet(N_1 + a_1, \ldots, N_K + a_k)$$

Posterior predictive probability

Let

$$x_k \in \mathbb{R}^K, \qquad x_k[i] = \begin{cases} 1 \ if \ i = k \\ 0 \ otherwise \end{cases}$$

Probability of x_k must be integrated over all possible θ.

$$p(x_k|D) = \int p(x_k[k] = 1)p(\theta|D)d\theta$$

Probability of $x_k[k] = 1$ is equal to θ_k

$$\Rightarrow p(x_k|D) = \int \theta_k p(\theta|D)d\theta$$

Integration over θ means the above is an expectation of θ_k

$$\Rightarrow p(x_k|D) = E_{Dirichlet(N_1+a_1,\dots,N_K+a_k)}(\theta_k)$$

Recalling expectation of Dirichlet distribution

$$\Rightarrow p(x_k|D) = \frac{N_k + a_k}{\sum_{i=1}^{K} N_i + a_i}$$

N is the sum of all N_i

$$\Rightarrow p(x_k|D) = \frac{N_k + a_k}{N + \sum_{i=1}^{K} a_i}$$

Part B: MAP of $\boldsymbol{\theta}$

Recall proportionality of $p(\theta|D)$:

$$p(\theta|D)\ \alpha \prod_{i=1}^{K} \theta_i^{N_i+a_i-1}$$

Since $p(\theta|D)$ is directly proportional to the product of powers of theta, we need only maximize the product

$$argmax_\theta p(\theta|D) = argmax_\theta \prod_{i=1}^{K} \theta_i^{N_i+a_i-1}$$

Logarithm function is monotonic, so maximizing the logarithm is also equivalent to maximizing the input of the logarithm

$$argmax_\theta \prod_{i=1}^{K} \theta_i^{N_i+a_i-1} = argmax_\theta \log\left(\prod_{i=1}^{K} \theta_i^{N_i+a_i-1}\right)$$

Applying the logarithm to the product

$$\log\left(\prod_{i=1}^{K} \theta_i^{N_i+a_i-1}\right) = \sum_{i=1}^{K} (N_i + a_i - 1) \log \theta_i$$

Note that $\sum \theta_j = 1$, so Lagrange multipliers must be used for maximization

$$\sum_{i=1}^{K} \theta_i = 1 \Rightarrow argmax_\theta \left(\sum_{i=1}^{K} (N_i + a_i - 1) \log \theta_i\right)$$

$$= argmax_\theta \left(\sum_{i=1}^{K} (N_i + a_i - 1) \log \theta_i - \lambda\left(\sum_{i=1}^{K} \theta_i - 1\right)\right)$$

$$= argmax_\theta \left(\sum_{i=1}^{K} \big((N_i + a_i - 1) \log \theta_i - \lambda\theta_i\big) - \lambda\right)$$

Taking derivatives of the Lagrangian equation:

$$\frac{d}{d\theta_j}\left(\sum_{i=1}^{K}\big((N_i + a_i - 1)\log\theta_i - \lambda\theta_i\big) - \lambda\right)$$

$$= \frac{d}{d\theta_j}\left(\sum_{i\neq j}^{K}\big((N_i + a_i - 1)\log\theta_i - \lambda\theta_i\big)\right) + \frac{d}{d\theta_j}\big((N_j + a_j - 1)\log\theta_j - \lambda\theta_j\big)$$

$$= \frac{N_j + a_j - 1}{\theta_j} - \lambda$$

$$\frac{d}{d\lambda}\left(\sum_{i=1}^{K}\big((N_i + a_i - 1)\log\theta_i - \lambda\theta_i\big) - \lambda\right) = -\sum_{i=1}^{K}\theta_i - 1$$

Setting derivatives to zero to find maximum

$$\left(0 = \frac{N_j + a_j - 1}{\theta_j} - \lambda\right) \Rightarrow \left(\lambda = \frac{N_j + a_j - 1}{\theta_j}\right) \Rightarrow \left(\theta_j = \frac{N_j + a_j - 1}{\lambda}\right)$$

$$\left(0 = -\sum_{i=1}^{K}\theta_i - 1\right) \Rightarrow \left(\sum_{i=1}^{K}\theta_i = 1\right)$$

Combining these equalities, we get:

$$\left(\sum_{i=1}^{K}\frac{N_i + a_i - 1}{\lambda} = 1\right) \Rightarrow \left(\sum_{i=1}^{K}(N_i + a_i - 1) = \lambda\right) \Rightarrow \left(\lambda = N + \sum_{i=1}^{K} a_i - K\right)$$

Substituting λ back into the definition of θ_j

$$\theta_j = \frac{N_j + a_j - 1}{N + \sum_{i=1}^{K} a_i - K}$$