CSC411 Fall 2018: Homework 5

Niloufar Afsariardchi

November 15, 2018

1 Problem 1

a) Based on Bayes rule:

$$
\begin{aligned}
P(y=k|\mathbf{x},\mu,\Sigma) &= \frac{P(\mathbf{x}|y=k,\mu,\Sigma)P(y=k)}{P(\mathbf{x}|\mu,\sigma)} && (1)\\
&= \frac{P(\mathbf{x}|y=k,\mu,\Sigma)P(y=k)}{\sum_t P(\mathbf{x}|y=t,\mu,\sigma)P(y=t)} && (2)\\
\log P(y=k|\mathbf{x},\mu,\Sigma) &= \log P(\mathbf{x}|y=k,\mu,\Sigma) + \log P(y=k) && (3)\\
&\quad - \log \sum_t P(\mathbf{x}|y=t\mu,\sigma)P(y=t) && (4)\\
\log P(y=k|\mathbf{x},\mu,\Sigma) &= \log P(\mathbf{x}|y=k,\mu,\Sigma) + \log 0.1 && (5)\\
&\quad - \log \sum_t 0.1 P(\mathbf{x}|y=t,\mu,\sigma) && (6)
\end{aligned}
$$

where the first term of RHS is given in the equation (1) of the handout. The LHS of above equation is conditional log-likelihood or log posterior distribution.

These are the results I got for average conditional log-likelihood:

conditional log-likelihood for training set: -0.125

conditional log-likelihood for test set: -0.197

conditional likelihood for training set: 0.88

conditional likelihood for test set: 0.82

This shows that the test set has the higher inherent uncertainty and perhaps is has more noise that the training set.

b) Here is the accuracy of the algorithm obtained by assigning each datatum to the digit with highest conditional log likelihood:

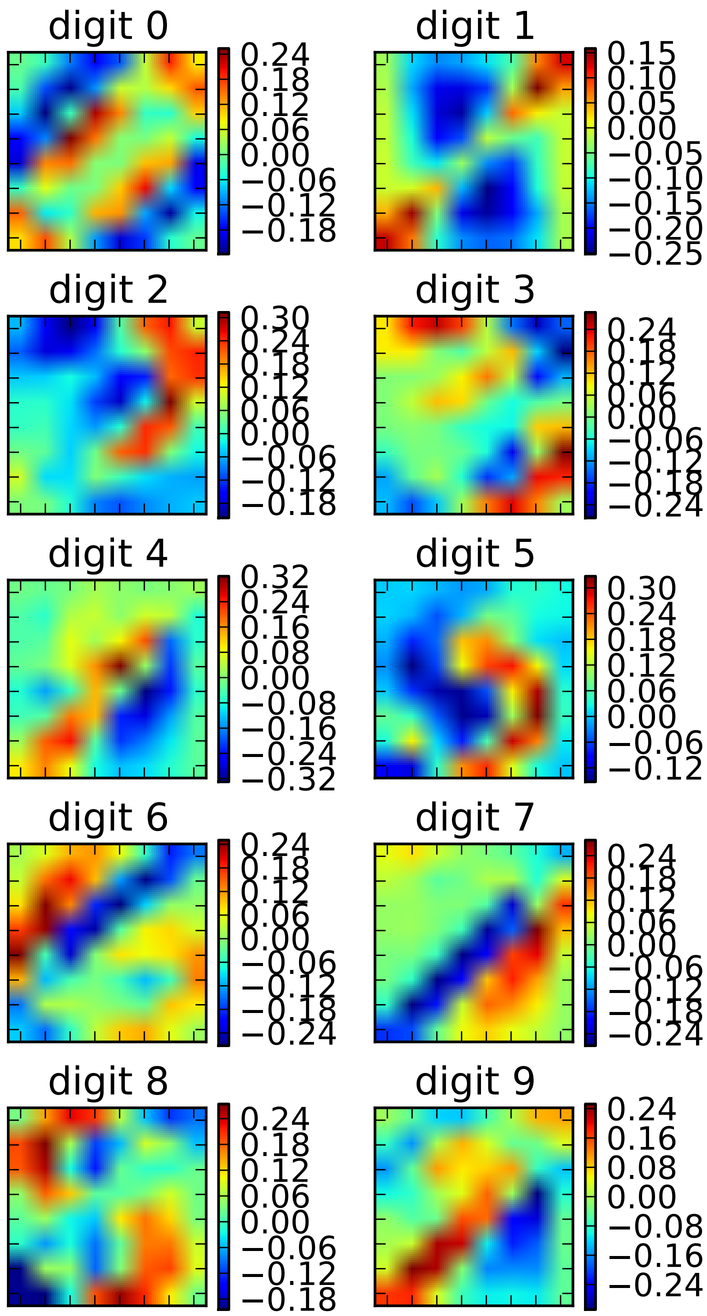
accuracy of training set: 0.98

accuracy of test set: 0.97

c) See attached Figure 1 below. There are some patterns in the eigenvectors, however they are not very clear. It is not expected for eigenvectors to reflect the digits in the exact form either.

2 Problem 2

Continue to next page


digit 0
0.24
0.18
0.12
0.06
0.00
−0.06
−0.12
−0.18
digit 1
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.25
digit 2
0.30
0.24
0.18
0.12
0.06
0.00
−0.06
−0.12
−0.18
digit 3
0.24
0.18
0.12
0.06
0.00
−0.06
−0.12
−0.18
−0.24
digit 4
0.32
0.24
0.16
0.08
0.00
−0.08
−0.16
−0.24
−0.32
digit 5
0.30
0.24
0.18
0.12
0.06
0.00
−0.06
−0.12
digit 6
0.24
0.18
0.12
0.06
0.00
−0.06
−0.12
−0.18
−0.24
digit 7
0.24
0.18
0.12
0.06
0.00
−0.06
−0.12
−0.18
−0.24
digit 8
0.24
0.18
0.12
0.06
0.00
−0.06
−0.12
−0.18
digit 9
0.24
0.16
0.08
0.00
−0.08
−0.16
−0.24

Question 2

(a) $P(\theta|D) = \dfrac{P(D|\theta)P(\theta)}{P(D)}$

assuming i.i.d xs

$$= \frac{\prod_{i=1}^{m} P(x^{(i)}|\theta) P(\theta)}{\prod_{i=1}^{m} P(x^{(i)})}$$

$$\propto \frac{\prod_{i=1}^{m} \prod_{k=1}^{K} \theta_k^{x_k^{(i)} - a_k - 1}}{\prod_{i=1}^{m} \int_\theta P(x^{(i)}|\theta') P(\theta') \, d\theta'}$$

Let's assume $k^{(i)}$ is the index of the category for which $x_k^{(i)} = 1$

$$= \frac{\prod_{k=1}^{K} \theta_k^{N_k - a_k - 1}}{\prod_{i=1}^{m} \int \theta_{k^{(i)}} Dir(\theta) \, d\theta}$$

$$= \frac{\prod_{k=1}^{K} \theta_k^{N_k - a_k - 1}}{\prod_{i=1}^{m} E[\theta_{k^{(i)}}]}$$

$$= \frac{\prod_{k=1}^{K} \theta_k^{N_k - a_k - 1}}{\prod_{i=1}^{m} \frac{a_{k^{(i)}}}{\sum_{k'} a_{k'}}}$$

$$= \frac{(\sum_{k'} a_{k'})^K \prod_{k=1}^{K} \theta_k^{N_k - a_k - 1}}{\prod_{k=1}^{K} a_k^{N_k}}$$

constant

$$= \left(\sum_{k'} a_{k'}\right)^K \prod_{k=1}^{K} \left(\frac{\theta_k}{a_k}\right)^{N_k} \theta_k^{-a_k - 1}$$

$$\propto \prod_{k=1}^{K} \left(\frac{\theta_k}{a_k}\right)^{N_k} \theta_k^{-a_k - 1}$$

which is Dirichlet

$P(\theta|D) \sim \text{Dirichlet}(a_1 - N_1, a_2 - N_2, \ldots, a_K - N_K)$

Posterior predictive distribution is

$$P(D'|D) = P(x_{k'}=1 \mid D)$$ that the next observation belongs to k'-th category

$$= \int P(\theta|D)\, P(x_{k'}=1|\theta)\, d\theta$$

$$\propto \int \left(\prod_{x=1}^{K} \left(\frac{\theta_k}{a_k}\right)^{N} \theta_x^{-a_x-1} \right) \cdot \theta_{x'}\, d\theta$$

We know $\int P(\theta_x|D)\, d\theta_x = 1$

$$= \int P(\theta_{x'}|D)\, \theta_{x'}\, d\theta_{x'}$$

$$= \int Dir(a_{x} - N_{k})\, \theta_{x'}\, d\theta_{x'}$$

$$= \underset{P(\theta|D)}{E}(\theta_{x'})$$

$$= \frac{a_{x'} - N_{x'}}{-N + \sum_{x=1}^{K} a_k}$$

(b) $$\hat{\theta} = \underset{\theta}{\text{argmax}} \log P(\theta|D)$$

$$= \underset{\theta}{\text{argmax}} \log \text{Dirichlet}(a_1+N_1, \ldots, a_K+N_K)$$

$$= \underset{\theta}{\text{argmax}} \log \prod_{x=1}^{K} \theta_k^{N_x - a_k - 1} = \underset{\theta}{\text{argmax}}\ J$$

$$J = \sum_{k=1}^{K} (N_x - a_x - 1) \log \theta_k$$

Optimization:

$$\underset{\theta}{\text{maximize}}\ J(\theta)$$

assuming $\sum_{x=1}^{K} \theta_x = 1$

maximize $O = J(\theta) - \lambda(1 - \sum_{k=1}^{K} \theta_k)$

① $\frac{\partial O}{\partial \theta_i} = \frac{N_i - a_i - 1}{\theta_i} + \lambda = 0 \implies \theta_i = \frac{1 + a_i - N_i}{\lambda}$

② $\frac{\partial O}{\partial \lambda} = 1 - \sum_{k=1}^{K} \theta_k = 0$

from ②: $\sum_{i=1}^{K} \frac{1 + a_i - N_i}{\lambda} = 1$

$\implies \lambda = \sum_{i=1}^{K} 1 + a_i - N_i$

from ① $\implies \boxed{\theta_i = \frac{1 + a_i - N_i}{-N + K + \sum_{i=1}^{K} a_i}}$

which is a little bit different from predictive distribution