CM224 HW 7 Solution

Ashish Kumar Singh (UID:105479019)

December 13, 2021

Problem 1. Viterbi Algorithm

Sub-Problem 1a. Most likely population ancestry path with haplotype 10111?

Answer 1a. Most likely population ancestry path is 11122

Solution 1a. Since either population is equivalent and $h_1 = 1$ we get,

$$C(1,1) = 0.5 * \xi(h_1, e_{11}) = 0.5 * e_{11} = 0.5 * 0.3 = 0.15$$

$$C(1,2) = 0.5 * \xi(h_1, e_{12}) = 0.5 * e_{12} = 0.5 * 0.2 = 0.1$$

For the second state $s = 2$, we have $h_2 = 0$ we get,

$$C(s+1, j) = max(C(s,i)\delta_{ij}\xi(h_{s+1}, e_{s+1,j}))$$

$$C(2,1) = max(C(1,i)\delta_{i1}\xi(h_2, e_{21}))$$

$$C(2,1) = max(C(1,1)\delta_{11}(1 - e_{21}), C(1,2)\delta_{21}(1 - e_{21}))$$

$$C(2,1) = max(0.15 * 0.5 * 0.3, 0.1 * 0.4 * 0.3) = 0.0225$$

$$C(2,2) = max(C(1,i)\delta_{i2}\xi(h_2, e_{22}))$$

$$C(2,2) = max(C(1,1)\delta_{12}(1 - e_{22}), C(1,2)\delta_{22}(1 - e_{22}))$$

$$C(2,2) = max(0.15 * 0.5 * 0.2, 0.1 * 0.6 * 0.2) = 0.015$$

For the third state $s = 3$, we have $h_3 = 1$ we get,

$$C(3,1) = max(C(2,i)\delta_{i1}\xi(h_2, e_{31}))$$

$$C(3,1) = max(C(2,1)\delta_{11}e_{31}, C(2,2)\delta_{21}e_{31})$$

$$C(3,1) = max(0.0225 * 0.5 * 0.3, 0.015 * 0.4 * 0.3) = 0.003375$$

$$C(3,2) = max(C(2,i)\delta_{i2}\xi(h_2, e_{32}))$$

$$C(3,2) = max(C(2,1)\delta_{12}e_{32}, C(2,2)\delta_{22}e_{32})$$

$$C(3,2) = max(0.0225 * 0.5 * 0.2, 0.015 * 0.6 * 0.2) = 0.00225$$

For the fourth state $s = 4$, we have $h_4 = 1$ we get,

$$C(4,1) = max(C(3,i)\delta_{i1}\xi(h_2, e_{41}))$$

$$C(4,1) = max(C(3,1)\delta_{11}e_{41}, C(3,2)\delta_{21}e_{41})$$

$$C(4,1) = max(0.003375 * 0.5 * 0.2, 0.00225 * 0.4 * 0.2) = 0.0003375$$

$$C(4,2) = max(C(3,i)\delta_{i2}\xi(h_2, e_{42}))$$

$$C(4,2) = max(C(3,1)\delta_{12}e_{42}, C(3,2)\delta_{22}e_{42})$$

$$C(4,2) = max(0.003375 * 0.5 * 0.3, 0.00225 * 0.6 * 0.3) = 0.00050625$$

For the fifth state $s = 5$, we have $h_5 = 1$ we get,

$$C(5,1) = max(C(4,i)\delta_{i1}\xi(h_2, e_{51}))$$

$$C(5,1) = max(C(4,1)\delta_{11}e_{51}, C(4,2)\delta_{21}e_{51})$$

$$C(5,1) = max(0.0003375 * 0.5 * 0.3, 0.00050625 * 0.4 * 0.3) = 0.00006075$$

$$C(5,2) = max(C(4,i)\delta_{i2}\xi(h_2, e_{52}))$$

$$C(5,2) = max(C(4,1)\delta_{12}e_{52}, C(4,2)\delta_{22}e_{52})$$

$$C(5,2) = max(0.0003375 * 0.5 * 0.25, 0.00050625 * 0.6 * 0.25) = 0.0000759375$$

On comparing above values, the most likely state is 11122

Problem 2. Similar to PCA

Answer 2. c) $\sum_{i=1}^{p} x_i^t \hat{U}\hat{U}^t x_i = \lambda_1 + \lambda_2$

Solution 2.

$$\sum_{i=1}^{p} x_i^t U U^t x_i = \sum_{j=1}^{2} u_j^t X X^t u_j$$

Let $v_1, ...v_n$ be the eigenvectors of XX^t

$$\sum_{i=1}^{p} x_i^t U U^t x_i = \sum_{j=1}^{2} \sum_{k=1}^{n} (u_j^t v_k)^2 \lambda_k$$

Assuming orthogonality, the above maximizes when $u_k = v_k$ and we get,

$$\sum_{i=1}^{p} x_i^t \hat{U}\hat{U}^t x_i = \sum_{k=1}^{2} \lambda_k$$

$$\sum_{i=1}^{p} x_i^t \hat{U}\hat{U}^t x_i = \lambda_1 + \lambda_2$$

Problem 3. Methylation data PCA coding

Sub-Problem 3a. What is the proportion of variance explained by the first 10 principal components?

Answer 3a. 56.8808%
Explained variance ratio by first 10 components is 56.8808%.
Data normalization is done to make it mean 0 and variance 1. code is attached.

Sub-Problem 3b. Find component with highest squared correlation to the batch vector?

Answer 3b. Fourth PC
Fourth PC gives the highest squared correlation of 0.84 to the batch vector. code is attached

Sub-Problem 3c. Find component with highest squared correlation to the batch vector obtained using PCA with sparsity procedure?

Answer 3c. First PC
First PC in the sparse PCA analysis gives the highest squared correlation of 0.49 to the batch vector. code is attached.