

Genomic Signal Processing

Questions for exam

1. Lecture 1

- (a) Briefly explain the terms: DNA, chromosome, gene, nucleic acid, protein, aminoacid, and form a single phrase with all words.
- (b) How is realized the transfer of information from DNA to proteins?
- (c) Draw the diagram of organization and processing steps in a micro-array experiment (page 54).
- (d) What are the disadvantages of segmentation of spots by fixed circles on micro-array images?

2. Lecture 2

- (a) Explain intuitively the criterion used in Fisher discrimination

$$J(\underline{a}) = \frac{(\underline{a}^T \underline{m}_0 - \underline{a}^T \underline{m}_1)^2}{\sigma_0^2 + \sigma_1^2} \quad (1)$$

- (b) Problem: The algorithm at page 13 is given to you, together with a very small set of data (say 5 two-dimensional vectors, and the corresponding class labels). You are asked to find \underline{a}^* , and to guess the label of a 6'th vector (in your answer you should perform all simple calculations, but you may leave un-solved the inversion of the matrix).
- (c) Define the three sums of squares: TSS, BSS, WSS. Show that with these variables you can state a property of decomposing the total variance in the data into two meaningful components.
- (d) How can the ratio BSS/WSS be used for ranking the discriminative power of different genes in a two-class classification problem. Explain what is the intuitive explanation of the ratio.

3. Lecture 3

- (a) State the MDL principle in its general form
- (b) Apply the two-part code MDL for a ternary predictor of the form $g_\ell = f(g_1, g_2)$, to compare the best predictors of various orders: $g_\ell = f(g_1)$, $g_\ell = f(g_2)$, $g_\ell = f(g_1, g_2)$, when a dataset of 15 measurements is given.
- (c) The same problem as before, but for a hard Boolean predictor.

4. Lecture 4

- (a) Describe the normalization of the maximum likelihood for a Bernoulli model, and write the resulting NML codelength.

- (b) Describe the normalization of the maximum likelihood for the boolean regression model, $y = f(X) \oplus \varepsilon$, and write the resulting NML code-length (no need to write down the fast computation formulae).
- (c) Present the (ALL versus AML) classification problem for leukemia data-set, when measurements of 1000 gene expressions from 72 patients are available. What will be the procedure for finding the cross-validation errors, when using the classifier of your choice, for the training/test ratio 2:1.

5. Lecture 5

- (a) Describe the difference between filter and wrapper feature selection.
- (b) Describe the cross-validation methodology in Figure 1 (Fig. 1 will be printed on exam sheet).
- (c) the same for Figure 2.
- (d) Describe the feature selection by forward search (hill climbing, best first search), or by backward methods.
- (e) Consider the following optimization problem:

$$\begin{aligned}
 \min_{\boldsymbol{\theta}, b} \quad & \|\boldsymbol{\theta}\|_1 + C \sum_{t=1}^n \xi(t) \\
 \text{s.t.} \quad & Y(t)(\boldsymbol{\theta}^T \mathbf{X}(t) + b) \geq 1 - \xi(t), \\
 & \xi(t) \geq 0, \quad 1 \leq t \leq n.
 \end{aligned} \tag{2}$$

Explain how this problem can be used for finding the best features and the best classifier parameters in a classification problem.

6. Lecture 6

- (a) Define the sum-of-exponential model and describe the problem of parameter estimation and model order selection.
- (b) Given the plots at page 7, interpret qualitatively the performance of models of various orders. What will be a criterion able to select automatically the best order (Answer: formula at page 9).
- (c) Describe the data available and goals of clustering in the application of cerebellum development using time series of gene expressions. What can be the interpretation of clusters obtained by n_{NML} and by further refining according to the ranges of time constants (Table at page 27 -but marked 13(?))?

7. Lecture 7

- (a) What simplified model was postulated for an interacting network of proteins and genes? Why the approximation of derivative by the ratio of difference is not satisfactory for solving the differential equation system when using gene expression measurements?
- (b) Start from the trajectory of the system in the form $x(t) = e^{Mt}x(0)$ to obtain a model of sum-of-exponentials for the gene expression time series. Comment on the equidistance of time measurements.
- (c) Given the network at page 18, write down the system matrix M and define the connectivity matrix S . Explain why S may be more relevant biologically than the matrix M .

8. Lecture 8

- (a) Describe the experimental setting for lysate array measurement of concentrations.
- (b) Transform the nonlinear optimization problem

$$\{\hat{q}, \hat{\beta}_1\} = \arg \min_{q, \beta_1} \sum_{i=1}^n (y_i - \beta_1(q_{s_i} - d_i))^2 \quad (3)$$

into a linear one, and write the solution in closed form.

- (c) Specify four model classes suitable for modeling the protein concentration in lysate array experiments (noise model and calibration curve model).
- (d) Describe a relaxation algorithm for estimating the concentrations and the parameters of the polynomial model of the calibration curve in a lysate array experiment.

9. Lecture 9

- (a) Enumerate the biological reasons for the DNA sequence to be compressible. Define and give an (imaginary) example of a palindrome matching.
- (b) Enumerate several reasons why the DNA sequence is difficult to compress.
- (c) Summarize the technique used in Biocompress method and explain the performance of the method observed over DNA sequences (pages 15 and 44).
- (d) Present a simple statistical model for DNA matching and the resulting NML model obtained for a frame of length n (no need to show how to speed up the computation of the normalization constant).
- (e) How can one accelerate the search for the best match in the past, and what is the impact over the compression efficiency.
- (f) What are the three competing methods for compressing the current frame in GeNML algorithm?