CS 8850: Advanced Machine Learning **Fall 2017**

Homework 3: Hypotheses Testing

INSTRUCTOR: DANIEL L. PIMENTEL-ALARCÓN DUE 9/20/2017

Problem 3.1 (Healthy vs. Diabetic)**.** The blood glucose level (in mg/dL) of a healthy person can be modeled as $\mathcal{N}(95, \sigma^2)$, while that of a diabetic can be modeled as $\mathcal{N}(140, \sigma^2)$. Given a new patient with glucose level x, you want to decide between two hypotheses:

$$\begin{aligned} H_0: \ x &\sim \mathcal{N}(95, \sigma^2) && \Rightarrow \text{healthy}, \\ H_1: \ x &\sim \mathcal{N}(140, \sigma^2) && \Rightarrow \text{diabetic}. \end{aligned}$$

Derive the likelihood ratio test for this hypothesis problem. In your own words, what does this test suggest?

Problem 3.2 (Different variances)**.** Derive the likelihood ratio test for

$$\begin{aligned} H_0: \ x_1, \dots, x_{\mathrm{N}} &\overset{iid}{\sim} \mathcal{N}(0, \sigma_0^2), \\ H_1: \ x_1, \dots, x_{\mathrm{N}} &\overset{iid}{\sim} \mathcal{N}(0, \sigma_1^2). \end{aligned}$$

where $\sigma_0 < \sigma_1$ are known.

Problem 3.3 (Exponentials)**.** Let $\boldsymbol{x} \in \mathbb{R}^2$ be a random vector, and consider the following hypotheses:

$$\begin{aligned} H_0: \ \boldsymbol{x} &\sim \frac{1}{2\pi} e^{-\frac{\|\mathbf{x}\|_2^2}{2}}, \\ H_1: \ \boldsymbol{x} &\sim \frac{1}{2} e^{-\sqrt{2}\|\mathbf{x}\|_1}. \end{aligned}$$

(a) Compute the mean and covariance matrix under each hypothesis.

(b) Derive a test with minimum probability of error.

(c) Sketch the decision regions of this test.

Problem 3.4 (Biotech startup)**.** Imagine that you have been hired by a biotechnology start-up company to help them identify whether certain genes may be associated with a form of cancer. They are currently interested in a particular gene, because they have developed a very cost-effective screening procedure. The procedure generates a number for each person that is screened. A study conducted on a large group of people showed that the numbers produced by the procedure can be modeled as outcomes of a $\mathcal{N}(0, 1)$ random variable for healthy people, and $\mathcal{N}(1, 1)$ for cancer patients.

(a) State this as a hypotheses test problem.

(b) Derive a test with minimum probability of error.

(c) Sketch the decision regions of this test.

(d) Do you think the company should market the procedure as a good test for this cancer? Why or why not?

After further analysis, it turns out that if you repeat the screening procedure on the same person, then you get a different number each time. However, the values that are produced by repeating the screening procedure multiple times can be modeled as independent realizations of a $\mathcal{N}(0,1)$ or $\mathcal{N}(1,1)$ random variable, for healthy people and cancer patients, respectively.

(e) Construct a more robust test for cancer based on this observation.

(f) Sketch the decision regions of this test.

(g) State the pros and cons of this new test.

Problem 3.5 (Brain Regions)**.** Scientists are studying how the brain performs a certain information-processing task. Three regions of the brain are involved, denoted A, B and C. There is prior evidence that there are direct neural connections between regions A and B, and regions B and C. However, it is uncertain whether regions A and C are directly connected. So the scientists design an experiment to test this. They scan human subjects' brains while performing the information-processing task. The activity level in each region is a binary-valued variable, indicating whether the region is significantly active. They record many measurements of these variables, for repeated trials of the task and different human subjects. Let $(a_{\mathrm{i}}, b_{\mathrm{i}}, c_{\mathrm{i}})$ denote the activity level in each region at the i^{th} recording. We can model each triple as an independent realization of the same multivariate random variable (a, b, c). However, each triplet $(a_{\mathrm{i}}, b_{\mathrm{i}}, c_{\mathrm{i}})$ may be correlated. If there is no direct connection between regions A and C, then we conjecture that a and c will be conditionally independent given b.

(a) How would you use your data to check for whether a and c are conditionally independent given b?

(b) I have generated two datasets, `brain_data1.mat` and `brain_data2.mat`, which simulate two different information-processing tasks. Use these data to determine whether a and c are conditionally independent given b. Your answer may be different in the two cases.

(c) Implement your procedure in Matlab and discuss your conclusions.