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

Homework 2: Review of Probability Theory

INSTRUCTOR: DANIEL L. PIMENTEL-ALARCÓN DUE 9/8/2017

Problem 2.1 (Diabetes testing)**.** With 9.3% of the U.S. population having diabetes, there is an increasing interest in studying this disease. Geneticists have determined that 95% of the people that develop type-2 diabetes have the following genes inactive:

- TCF7L2. Affects insulin secretion and glucose production.
- ABCC8. Helps regulate insulin.
- GLUT2. Helps move glucose into the pancreas.

(a) If you sequence your genome and find out that these genes are inactive, what is the probability that you develop type-2 diabetes?

(b) What other information would you need to know?

(c) Based on this information, when should you be concerned?

Problem 2.2 (Gaussian mixture)**.** Let $\boldsymbol{x} \in \mathbb{R}^{\mathrm{D}}$ be a random vector and $z \in \{1, \ldots, \mathrm{K}\}$ be a random variable. Suppose

$$\boldsymbol{x} \big| z = \mathrm{k} \;\sim\; \mathcal{N}(\boldsymbol{\mu}_{\mathrm{k}}, \boldsymbol{\Sigma}_{\mathrm{k}}).$$

Derive an expression for the marginal distribution of $\boldsymbol{x}$ alone.

Problem 2.3 (Snapchat's delays)**.** Suppose that you are sending pics to your girlfriend/boyfriend overseas. Each time you send a picture through the Internet it takes a certain amount of time to reach your gf/bf. Assume that you can measure the time delay. The delay won't be constant, since it depends on how much other traffic is in the Internet (in particular at the routers that handle your messages). You and your gf/bf measure the delays of several packet transmissions. It appears that there is a minimal time delay, say t_0 (msec). Based on your observations, it seems that larger delays are rarer than shorter ones. Propose a probabilistic model for the delays with a single free parameter θ. The value of θ should govern the expected delay characteristics. Let x denote a random variable that represents the delay. The observations you have made are assumed to be independent realizations of this random variable. Let $\mathsf{p}(\mathrm{x}|\theta)$ denote the probability density of x. Give an explicit form for $\mathsf{p}(\mathrm{x}|\theta)$ and explain the rationale of your model.

Problem 2.4 (Simulating random variables)**.** In this problem you will simulate random variables using Matlab, and study their distributions.

(a) Generate $\mathrm{N} = 1000$ i.i.d. Uniform$(0, 1)$ random variables $x_1, \ldots, x_{\mathrm{N}}$, and plot their histogram. Does it look fairly uniform?

(b) Let

$$y_{\mathrm{i}} \;=\; \begin{cases} 1 & \text{if } x_{\mathrm{i}} \leq \mathrm{p} \\ 0 & \text{otherwise.} \end{cases}$$

What is the distribution of y_{i}?

(c) Plot the histogram of the y_i's for $p = 1/4, 1/2, 3/4$. Do these histograms match the distribution of your answer from (b)?

(d) Let z_k be the sum of the k^{th} batch of n y_i's. What is the distribution of z_k?

(e) Plot the histogram of the z_k's with $n = 10$ and $p = 1/4, 1/2, 3/4$. Do these histograms match the distribution of your answer from (d)?

Problem 2.5 (Simulating random vectors)**.** In this problem you will simulate random **vectors** using Matlab, and study their distributions.

(a) Generate $N = 1000$ random vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^3$ with entries distributed i.i.d. Uniform$(0, 1)$ and plot this sample.

(b) Generate $N = 1000$ random vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N \in \mathbb{R}^3$ with entries distributed i.i.d. Normal$(0, 1)$ and plot this sample.

(c) What differences do you notice?

(d) What is the covariance matrix of $\boldsymbol{y}_i$?

(e) Compute the sample covariance matrix of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$. What is its rank?

Next you will generate a random sample lying in a subspace.

(f) Create a basis $\mathbf{U}$ of a 2-dimensional subspace of $\mathbb{R}^3$ (Hint: review Homework 1).

(g) Generate $N = 1000$ random vectors $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_N \in \mathbb{R}^2$ with entries distributed i.i.d. Uniform$(0, 1)$. These will be our coefficients. Construct our samples as $\boldsymbol{x}_i = \mathbf{U}\boldsymbol{c}_i$, and plot them.

(h) Generate $N = 1000$ random vectors $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_N \in \mathbb{R}^2$ with entries distributed i.i.d. Normal$(0, 1)$. These will be our coefficients. Construct our samples as $\boldsymbol{y}_i = \mathbf{U}\boldsymbol{c}_i$, and plot them.

(i) What differences do you notice?

(j) What is the covariance matrix of $\boldsymbol{y}_i$?

(k) Compute the sample covariance matrix of $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$. What is its rank? How do you interpret this?