CS 760: Machine Learning © Fall 2020

Homework 1: Review

INSTRUCTOR: DANIEL L. PIMENTEL-ALARCÓN DUE 09/24/2020

GO GREEN. AVOID PRINTING, OR PRINT 2-SIDED MULTIPAGE.

In this homework you will review some basic linear algebra, probability, and optimization concepts. Each sub problem is worth 10 points.

Problem 1.1. Show that $\mathbb{R}^{\mathrm{D}}$ is a subspace.

Problem 1.2. Subspaces spaces are, by definition, *closed* under linear combinations. For example, when you add or multiply elements of $\mathbb{R}^{\mathrm{D}}$, you end up with an element of $\mathbb{R}^{\mathrm{D}}$ (as shown in Problem 1.1). In other words, you cannot *fall* out of $\mathbb{R}^{\mathrm{D}}$ by adding or multiplying. Subspaces are not necessarily closed under *all* mathematical operations.

(a) Show that $\mathbb{R}^{\mathrm{D}}$ is *not* closed under element-wise square roots.

(b) Give an example of a subspace that is closed under element-wise square roots (besides being closed under linear combinations, as *all* subspaces must be).

Problem 1.3. Let $\mathbf{u}_1, \ldots, \mathbf{u}_{\mathrm{R}} \in \mathbb{R}^{\mathrm{D}}$. Show that $\mathbb{U} = \mathrm{span}[\mathbf{u}_1, \ldots, \mathbf{u}_{\mathrm{R}}]$ is a subspace.

Problem 1.4 (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 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 diabetes?

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

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

Problem 1.5 (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 the traffic of 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 $\mathbb{P}(\mathrm{x}|\theta)$ denote the probability density of x. Give an explicit form for $\mathbb{P}(\mathrm{x}|\theta)$ and explain the rationale of your model.

Problem 1.6 (Simulating random variables)**.** In this problem you will simulate random variables 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 with $\mathrm{p} = 1/4, 1/2, 3/4$. 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 $\mathrm{n} = 10$ and $\mathrm{p} = 1/4, 1/2, 3/4$. Do these histograms match the distribution of your answer from (d)?

Problem 1.7 (Logistic Gradient)**.** The following expression describes the log-likelihood of the logistic regression model:

$$\ell(\boldsymbol{\theta}) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \left[\mathrm{y_i} \log \left(\frac{1}{1 + e^{-\boldsymbol{\theta}^\mathsf{T} \mathbf{x}_\mathrm{i}}} \right) + (1 - \mathrm{y_i}) \log \left(\frac{1}{1 + e^{\boldsymbol{\theta}^\mathsf{T} \mathbf{x}_\mathrm{i}}} \right) \right].$$

The goal in logistic regression is to maximize this quantity.

(a) Derive an expression for its gradient (w.r.t. $\boldsymbol{\theta}$).

(b) Derive an expression for its Hessian.

(c) Is $\ell(\boldsymbol{\theta})$ a scalar, a vector, or a matrix? What about its gradient? What about its Hessian?