CS 542: Machine Learning **Summer**

Problem Set 1

Lecturer: Prof. Peter Chin *Due: June 4*

- ◇ Please submit your solutions for the written questions (either type up your answers or scan your handwritten solution) in a single PDF on Gradescope under the assignment named `"PS1"` by 23:59PM on the due date. In your PDF submission, make sure you have a single PDF with EIGHT pages in total as specified in the problem prompts below.
- ◇ For the programming part, please follow the instructions in the shared colab notebook. Save a copy of the notebook to your own Google Drive and include your code in the copy together with necessary graphs, explanations and analysis. Keep the output and make sure the same results can be reproduced by running the colab notebook you submitted in sequence. Please include all necessary files that are required to run the notebook. If it requires additional packages, please specified clearly the procedure to install them. When submitting, submit a link (after enabling sharing) to your finished colab notebook.
- ◇ Late policy: there will be a penalty of 10% per day, up to TWO days late. After that no credit will be given. The written part and programming part are considered as a whole when calculating late penalties.

1. **Probability (20 pts, page 1)** Please include all your solutions for this question in page 1 of your PDF.

   (a) Show that if two variables x and y are independent, then their covariance is zero.

   (b) Table 1 represents a hypothetical study of a test for HIV in a population of intravenous drug users, some carrying HIV and others not. "+" denotes a positive outcome from the test and "–" denotes a negative outcome.

Outcome	with HIV	without HIV	Total
+	72	12	84
–	3	71	74
Total	75	83	158

Table 1.1. HIV test summary for intravenous drug users

   i. Calculate the probability of test being positive given the patient has HIV.

   ii. Calculate the probability of test being negative given the patient has HIV.

   iii. If the prevalence of HIV is 12% in intravenous drug users (12 cases per 100 patients), what is the probability of the patient having HIV after a positive test? A negative test?

2. **Bayes Theorem (20 pts, page 2)** Please include all your solutions for this question in page 2 of your PDF.

   (a) Monty Hall problem: On the game show, Let's make a Deal, you are shown four doors: A, B, C, and D, and behind exactly one of them is a big prize. Michael, the contestant, selects one of them, say door C, because he know from having watched countless number of past shows that door C has twice the probability of being the right door than door A or door B or door D. To make things more interesting, Monty Hall, game show host, opens one of the other doors, say door B, revealing that the big prize is not behind door B. He then offers Michael the opportunity to change the selection to one of the remaining doors (door A or door D). Should Michael change his selection? Justify your answer by calculating the probability that the prize is behind door A, the probability that the prize is behind door C, and the probability that the prize is behind door D, given that Monty Hall opened door B (to show that prize is not there), using Bayes Theorem.

   (b) Prof. Chin knows that historically 2 out 75 students in his CS 542 class cheats (!) on his exams. Last semester, he suspected one of the students engaged in cheating during the final, but when he gently confronted the student, the student vehemently denied that he was cheating. If Prof. Chin has 90% accuracy in identifying cheaters (i.e. if a student is cheating, 90% of the time, Prof. Chin will indeed identify that the student as a cheater), but also has 20% false alarm rate (i.e. even though a student is not cheating, Prof. Chin will erroneously identify that student as a cheater). What is the probability that the student Prof. Chin confronted in the last semester's final was indeed cheating?

3. **Linear Algebra (20 pts, page 3)** Please include all your solutions for this question in page 3 of your PDF.

   (a) Show that the inverse of a symmetric matrix is itself symmetric.

   (b) Find the eigenvalues and eigenvectors for

$$A = \begin{bmatrix} 3 & 4 & -1 \\ -1 & -2 & 1 \\ 3 & 9 & 0 \end{bmatrix}$$

   Is A positive definite?

4. **Probability Distributions (20 pts, pages 4-7)**

   (a) (Page 4) The form of the Bernoulli distribution given by (2.2) in the textbook is not symmetric between the two values of x. In some situations, it will be more convenient to use an equivalent formulation for which $x \in \{-1, 1\}$, in which case the distribution can be written

$$p(x|\mu) = \left(\frac{1-\mu}{2}\right)^{(1-x)/2} \left(\frac{1+\mu}{2}\right)^{(1+x)/2}$$

where $\mu \in [-1, 1]$. Show that the distribution defined above is normalized, and evaluate its mean, variance, and entropy.

(b) (Page 5) Using the property $\Gamma(x+1) = x\Gamma(x)$ of the gamma function, derive the following results for the mean, variance, and covariance of the Dirichlet distribution given by (2.38) in the textbook

$$\mathbb{E}[\mu_j] = \frac{\alpha_j}{\alpha_0}$$
$$\text{var}[\mu_j] = \frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)}$$
$$\text{cov}[\mu_j \mu_l] = -\frac{\alpha_j \alpha_l}{\alpha_0^2(\alpha_0 + 1)}, \quad j \neq l$$

where α_0 is defined by (2.39).

(c) (Page 6) The uniform distribution for a continuous variable x is defined by

$$\text{U}(x|a, b) = \frac{1}{b-a}, \quad a \leqslant x \leqslant b$$

Verify that this distribution is normalized, and find expressions for its mean and variance.

(d) (Page 7) Show that the entropy of the multivariate Gaussian $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is given by

$$\text{H}[\mathbf{x}] = \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{D}{2}(1 + \ln(2\pi))$$

where D is the dimensionality of $\mathbf{x}$

5. **Programming assignment (20 pts, page 8)**

Please follow the instructions in the Google Colab notebook below. Save a copy to your Google Drive and edit it as instructed. When submitting your homework on Gradescope, please share a link to your finished Colab file. The last edited time on colab is considered your submission time.

https://colab.research.google.com/drive/1MOo2PkgMurOWs80E-ZbnnOwz-G6oud7N?usp=sharing