Homework Assignment 2

COGS 118A: Introduction to Machine Learning I

Due: 11:59pm, Monday, January 22nd 2018 (Pacific Time).

Instructions: Answer the questions below, attach your code, and insert figures to create a PDF file; submit your file via TritonEd (ted.ucsd.edu). You may look up the information on the Internet, but you must write the final homework solutions by yourself.

Late Policy: 5% of the total points will be deducted on the first day past due. Every 10% of the total points will be deducted for every extra day past due.

System Setup: You can install Anaconda to setup the Jupyter Notebook environment. Most packages have been already installed in Anaconda. If some package is not installed, you can use `pip` to install the missing package, that is, just type `pip install PACKAGE_NAME` in the terminal.

Grade: ____ out of 100 points

1 (10 points) Matrix Calculus

1.1 (5 points)

Suppose $x \in \mathbb{R}$, for $f(x) = \lambda(1 - x^2)$ where λ is a constant, determine the derivative of $f(x)$ with respect to x.

1.2 (5 points)

Several particular derivatives are useful for the course. For matrix $\mathbf{A}$ and vector $\mathbf{x}$ and vector $\mathbf{a}$, we have

- $\dfrac{\partial \mathbf{a}^T \mathbf{x}}{\partial \mathbf{x}} = \dfrac{\partial \mathbf{x}^T \mathbf{a}}{\partial \mathbf{x}} = \mathbf{a}$,
- $\dfrac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}$. If $\mathbf{A}$ is symmetric, $\dfrac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$.

Applying the above rules, for $f(\mathbf{x}) = \lambda(1 - \mathbf{x}^T \mathbf{A} \mathbf{x})$ where $\mathbf{A}$ is a symmetric matrix and λ is a constant, derive $\dfrac{\partial f(\mathbf{x})}{\partial \mathbf{x}}$.

2 (20 points) Decision Boundary

2.1 (10 points)

We are given a classifier that performs classification in $\mathbb{R}^2$ (the space of data points with 2 features (x_1, x_2)) with the following classification rule:

$$h(x_1, x_2) = \begin{cases} 1, & \text{if} \quad x_1 + 2x_2 - 4 \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

Draw the decision boundary of the classifier and shade the region where the classifier predicts 1. Make sure you have marked the x_1 and x_2 axes and the intercept points on those axes.

2.2 (10 points)

We are given a classifier that performs classification on $\mathbb{R}^2$ (the space of data points with 2 features (x_1, x_2)) with the following classification rule.

$$h(x_1, x_2) = \begin{cases} 1, & \text{if} \quad w_1 x_1 + w_2 x_2 + b \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

Here, the normal vector of the hyperplane (decision boundary) must be normalized, i.e.:

$$||\mathbf{w}||_2 = \sqrt{w_1^2 + w_2^2} = 1.$$

Compute the parameters w_1, w_2 and b for the decision boundary in Figure 1.
Hint: Utilize the intercepts in the figure to find the relation between w_1, w_2 and b. Then, substitute it into the normalization constraint to solve the value for parameters.

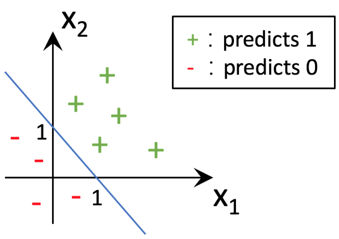


Figure 1: Decision boundary to solve the parameters.

3 (10 points) One-hot Encoding

A dataset S is denoted as $S = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$, where each sample refers to the specification of a car.

	Length (inch)	Height (inch)	Make	Color
$\mathbf{x}_1$	183	62	Toyota	Blue
$\mathbf{x}_2$	181	65	BMW	Silver
$\mathbf{x}_3$	182	59	BMW	Red
$\mathbf{x}_4$	179	68	Ford	Blue
$\mathbf{x}_5$	182	53	Toyota	Black

Represent this dataset S using a matrix and briefly explain your design decisions.
Hint: (1) For the categorical feature, you may use the one-hot encoding strategy. (2) You may choose either a row vector or a column vector to represent each data sample in your result. If you use a row vector to represent each data sample, the shape of the result matrix should be 5×9.

4 (20 points) Conditional Probability

Oftentimes, the performance of a binary medical diagnostic tests is measured as follows:

1. True positive rate (correctly identified) = $P(test + |sick+)$ = the probability that a sick people correctly diagnosed as sick.
2. False positive rate (incorrectly identified) = $P(test + |sick-)$ = the probability that a healthy people incorrectly identified as sick.
3. True negative rate (correctly rejected) = $P(test - |sick-)$ = the probability that a healthy people correctly identified as healthy.
4. False negative rate (incorrectly rejected) = $P(test - |sick+)$ = the probability that a sick people incorrectly identified as healthy.

A particular mammogram tests for breast cancer. The true positive rate is 98%. The true negative rate is 94%. The incident rate of breast cancer among a certain population is 0.06%. Suppose that a person is randomly drawn from the population.

4.1 (7 points)

Given that the person just tested positive, what is the probability of having breast cancer? In other words, what is $P(cancer + |test+)$?

4.2 (7 points)

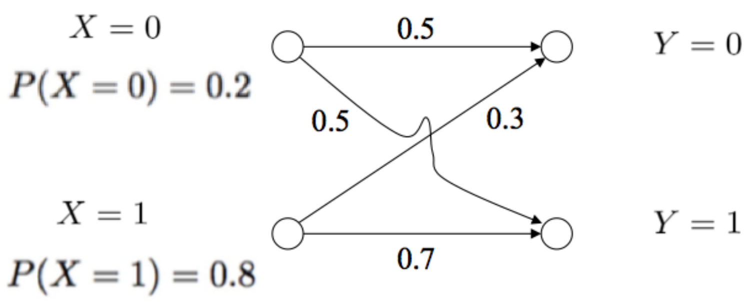
Given that the person just tested negative, what is the probability of **not** having breast cancer? In other words, what is $P(cancer - |test-)$?

4.3 (6 points)

Compute $precision$, $recall$, and $F - value = \frac{2 \times precision \times recall}{precision + recall}$.

5 (15 points) Binary Communication System

For the binary communication system shown below, compute the following probabilities.



(a) **(3 points)** $P(X = 2)$

(b) **(4 points)** $P(Y = 0|X = 1)$

(c) **(4 points)** $P(Y = 0)$

(d) **(4 points)** $P(X = 1|Y = 0)$

6 (25 points) Decision Stump

In this problem, we will perform a binary classification task on the Iris dataset. This dataset has 150 data points, where each data point $\mathbf{x} \in \mathbb{R}^4$ has 4 features and its corresponding label $y \in \{0, 1\}$.

To classify these 2 labels above, we decide to utilize a decision stump. The decision stump works as follows (for simplicity, we restrict our attention to uni-directional decision stumps):

- Given the j-th feature $\mathbf{x}(j)$ and a threshold Th, for each data point with index i, the classification function is defined by $y = f(\mathbf{x}, j, Th)$ as:

$$f(\mathbf{x}, j, Th) = \begin{cases} 1 & if\ \mathbf{x}(j) \geq Th \\ 0 & otherwise. \end{cases}$$

Based on the decision stump above, we wish to write an algorithm to find the **best feature** and **best threshold** on training set to create a "best" decision stump, in a sense that such decision stump achieves the **highest accuracy on training set**.

Follow the instructions in the skeleton code and report:

- All 4 histograms in last part of the code.
- The best feature, best threshold, training and test accuracy in last part of the code.