

ICS273A: HW1

Due: Jan 18

Problem 1: Probability review. Let X be a Bernoulli random variable, which can be represented with a biased coin flip. If $\pi = .5$, then the coin is fair. Specifically,

$$P(X = x) = \pi^x(1 - \pi)^{(1-x)}$$

where $x \in \{0, 1\}$ and $\pi \in [0, 1]$ is the bias of the coin. Recall the following definitions of mean and variance

$$\begin{aligned}\mu &= E[X] = \sum_x xP(X = x) \\ \text{var}(X) &= E[(X - \mu)^2]\end{aligned}$$

Compute the mean and variance of a Bernoulli random variable, as a function of the bias π .

Problem 2: Calculus review. Let

$$f(z) = \frac{1}{1 + e^{-z}}$$

This is called the sigmoid, or logistic function, and comes in handy for classification.

1. Roughly sketch this function by hand. What is the min and max of $f(z)$?
2. Compute its derivative analytically $\frac{\partial f}{\partial z}$.
3. The derivative can be written as a quadratic function of f , or $\frac{\partial f}{\partial z} = a + bf(z) + cf(z)^2$. What are a, b, c ?
4. Let $\pi = f(z)$. Sketch the derivative from (3) as a function of π . Let us interpret π as the probability that a biased coin comes up heads. Does the derivative from (3) have any probabilistic interpretation? (Hint; look at the results from Problem 1).

Problem 3: Linear algebra overview. Define the following notation for 2×2 matrices

$$A_{2 \times 2} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

$$\text{tr} A = \sum_i a_{ii}$$

$$\nabla_A f(A) = \begin{bmatrix} \frac{\partial f(A)}{\partial a_{11}} & \frac{\partial f(A)}{\partial a_{12}} \\ \frac{\partial f(A)}{\partial a_{21}} & \frac{\partial f(A)}{\partial a_{22}} \end{bmatrix}$$

Show the following holds true in the case of A, B, C being 2×2 matrices and x a 2 dimensional column vector.

1. $\text{tr}(AB) = \text{tr}(BA)$
2. $\text{tr}(ABC) = \text{tr}(CAB)$ (Hint; use the above result to show this)
3. $\nabla_A \text{tr}(AB) = B^T$
4. $\nabla_x \text{tr}(x^T Ax) = Ax + A^T x$

Problem 4: Normal equations. We want to linearly predict $y^{(i)}$ from the n dimensional vector $x^{(i)}$. If we are given m points for training data, we can write

$$X = [x^{(1)} \quad x^{(2)} \quad \dots \quad x^{(m)}]^T$$

$$\bar{y} = [y^{(1)} \quad y^{(2)} \quad \dots \quad y^{(m)}]^T$$

$$\Theta = [\theta_1 \quad \theta_2 \quad \dots \quad \theta_n]^T$$

1. Write the squared error cost function $J(\Theta) = \frac{1}{2} \sum_{i=1}^m (\Theta^T x^{(i)} - y^{(i)})^2$ using matrix notation.
2. Using the trace rules from the above problem, derive the normal equations $X^T(X\Theta - \bar{y}) = 0$.

Problem 5: [MATLAB] Linear regression. You will implement online gradient descent, also called the LMS (least mean squared) algorithm. I will provide a datafile called hw1.dat. The first two columns of the data are components of the input vector x and the last column in the output value y . Note there is no constant term for this problem (no θ_0).

1. Load the data into MATLAB - I recommend using the `textread` function (type "help textread" at the prompt for help on any command). Solve for the optimal linear model θ_{opt} using the normal equations. What are the 2 model coefficients?
2. Compute the 2×2 matrix $C = X^t X$ where X is the design matrix. Compute the eigenvectors and eigenvalues of C using the `eig` command. (a) What are

they? (b) What are $v_1^T v_1$, $v_2^T v_2$, and $v_1^T v_2$? (c) Show that C can be decomposed into $C = V\Lambda V^T$, where

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

$$V^T V = V V^T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

3. Plot the error cost function $J(\theta)$ using MATLAB's `contour` command. To do this, you will need to compute the value of J over a grid of values of θ . The command `ndgrid` might be helpful to generate the parameter values.
4. Initialize the LMS algorithm from $\theta = [0 \ 0]^T$. Set the stepsize $\alpha = \beta \cdot \frac{1}{\lambda_{max}}$, where λ_{max} is the maximum eigenvalue from (2) and β varies from 2.5, 1.5, and .5. Generate 3 plots that show the path of the LMS algorithm on the contour plot from (3). Try other scaling factors β for yourself. Describe the qualitative convergence behaviour as a function of β .
5. Pre-conditioning. We will show that by applying a linear transformation to the data $x^{(i)}$, we can make the error surface more spherical. This will make it easier to select a stepsize α that will provide good convergence. Define $T = V\Lambda^{-\frac{1}{2}}$, where V is as defined from (2) and

$$\Lambda^{-\frac{1}{2}} = \begin{bmatrix} \frac{1}{\sqrt{\lambda_1}} & 0 \\ 0 & \frac{1}{\sqrt{\lambda_2}} \end{bmatrix}$$

Let $W = XT$. (a) Compute $W^T W$ using matlab. What are its eigenvectors and eigenvalues? (b) Prove that $W^T W$ is the identity matrix using the eigendecomposition $X^T X = V\Lambda V^T$. Show the contour plot for the error surface J using the transformed points W instead of X .

Please show all work including code. Work can be handed in class, under the door in my office (DBH 4072), or using the online drop box.