

Homework 0
CS534 Machine Learning, Spring 2019

This timed homework assignment tests prerequisite knowledge in linear algebra, statistics, and programming. Please observe the time limits in each problem (these limits include time to lookup facts you may have forgotten). Show your work, attach any code in subsequent pages, and include a filled and signed copy of the honor pledge.

Problem 1 (linear algebra - 2 hours)

1.a. Given the pairs of vectors $v_1 = [\frac{\sqrt{3}}{2}, \frac{1}{2}, 0]^T$, $v_2 = [0, \frac{1}{2}, \frac{\sqrt{3}}{2}]^T$ and $\hat{v}_1 = [\frac{1}{\sqrt{2}}, 0, \frac{-1}{\sqrt{2}}]^T$, $\hat{v}_2 = [\frac{-1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0]^T$, use software to find the matrix $A \in \mathbb{R}^{3 \times 3}$ such that

$$[\hat{v}_1, \hat{v}_2] = A[v_1, v_2], \quad [v_1, v_2] = A^{-1}[\hat{v}_1, \hat{v}_2]. \quad (1)$$

1.b. Show that the inverse of a nonsingular skew-symmetric matrix $A^T = -A$ is also skew-symmetric.

1.c. Show that the eigenvalues of a real symmetric matrix $A = A^T$ are always real (use the conjugate transpose).

Problem 2 (programming)

Gradient-based optimization is a fundamental technique in machine learning. This technique seeks to minimize a cost f that is a function of some parameters β by iteratively updating the parameters using information from the gradient of the cost.

Given Himmelblau's function

$$f(\beta_1, \beta_2) = (\beta_1^2 + \beta_2 - 11)^2 + (\beta_1 + \beta_2^2 - 7)^2 \quad (2)$$

2.a. Identify the *critical points* $\beta^* = [\beta_1^*, \beta_2^*]^T$ of f using a software solver. Classify each critical point as a local minimum, maximum, or saddle point by analyzing the Hessian matrix of second partial derivatives H at β^*

$$H f = \begin{bmatrix} \frac{\partial^2 f}{\partial \beta_1^2} & \frac{\partial^2 f}{\partial \beta_1 \partial \beta_2} \\ \frac{\partial^2 f}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 f}{\partial \beta_2^2} \end{bmatrix} \bigg|_{\beta=\beta^*} \quad (3)$$

Generate a contour plot of the cost f in the range $\hat{\beta}_1, \hat{\beta}_2 \in [-5, 5]$. Plot the critical points and indicate their classifications.

2.b. Implement gradient descent to find solutions minimizing f using the updating function

$$\hat{\beta}^{i+1} = \hat{\beta}^i - \gamma \nabla f(\hat{\beta}^i) \quad (4)$$

where $\hat{\beta}^i = [\hat{\beta}_1^i, \hat{\beta}_2^i]^T$ is the solution at iteration i , and the *learning rate* $\gamma = 1e-3$. Plot the trajectory of your solutions $\hat{\beta}^0, \hat{\beta}^1, \dots, \hat{\beta}^N$ where $N = 10^5$ iterations on the plot from **2.a.**. Illustrate trajectories for ten random initializations of $\hat{\beta}^0 \in [-5, 5]$.

2.c. Implement Newton's method to find solutions minimizing f using the updating function

$$\hat{\beta}^{i+1} = \hat{\beta}^i - \gamma [H f(\hat{\beta}^i)]^{-1} \nabla f(\hat{\beta}^i), \quad (5)$$

where H is the Hessian. Use *learning rate* $\gamma = 1e-2$. Illustrate trajectories for ten random initializations as in as in **2.b.**

Problem 3 (statistics & probability - 1 hour)

A discrete-valued random variable is defined by its probability mass function $P = \{p_1, p_2, \dots, p_n\}$, where $0 \leq p_i \leq 1$ is the probability of outcome X_i .

One interpretation of the *Shannon entropy* of a random variable \mathbf{X} is the degree of uncertainty in the outcome of \mathbf{X} . For discrete-valued random variables the Shannon entropy is calculated as

$$S(\mathbf{X}) = - \sum_i p_i \log(p_i). \quad (6)$$

3.a. Suppose there is a game with n possible outcomes and probability mass function $p = \{p_1, p_2, \dots, p_n\}$. What P maximizes the entropy of the game? Explain your answer in plain English.

3.b. Show that this choice of p maximizes S (Note that for a concave function f , $\sum_i a_i f(x_i) \leq f(\sum_i a_i x_i)$ for $a_i \geq 0$, $\sum_i a_i = 1$).