UTSA

CS 6243 Machine Learning

EE 6363: Advanced Machine Learning

Assignment: HW4

## I. Assignment Instructions

Solve all given problems.

Submit your report in the dedicated folder on CANVAS.

Submit your report by the deadline. Delayed reports will not be graded.

Deadline: 11/29/2023, 11:59pm.

Reports must be typed and uploaded in PDF format. Handwritten reports will not be graded.

You can use any source (notes, books, online), but it must be cited in a References List at the end of your report.

Do not outsource this assignment (or parts of it) to another intelligent entity (whether human or AI). Except for what you explicitly cite, what you submit must be your own intellectual work.

Grade: This assignment corresponds to 6% of your final grade. You will be graded based on 1) correctness, 2) completeness, 3) clarity (equal weight).

Teams: Work individually (not in teams).

Terminology: Present = just show result. Derive = show all math steps. Discuss = discuss in words (no need for math).

**Problems**

1. **Present** GD updates on $w$ that minimize $J(w) = \|Aw - y\|_2^2 + \lambda\|w\|_2^2$ (arbitrary initialization and constant step-size). For adaptive step size, **derive** the Exact Line Search optimal step size for general update $n$. **Discuss** the type of regression that has objective $J(w)$ and when would it be preferred.

2. Consider data model with input $x$ and output $y(x) = f(x) + \epsilon$, where $f(\cdot)$ is unknown deterministic function and $\epsilon$ is unknown random noise, independent of $x$, with mean 0 and variance $\sigma_\epsilon^2$. Consider trained function $\hat{f}(\cdot)$, trained over random dataset, that estimates $f(x)$. For input $x$, define squared error $s(x) = f(x) - \hat{f}(x)$. **Derive** that the mean of $s(x)$ is equal to the sum of the squared bias and the variance of $\hat{f}(x)$.

3. **Discuss** the "curse of dimensionality" in non-parametric learning.

4. Consider parametric model such that for input $x$ the output is $y = b(x)^T w + \epsilon$, $w \in \mathbb{R}^D$, and training data $\{(y_i, x_i)\}_{i=1}^N$. Assume prior Gaussian distribution of $w$, with mean $m_0$ and covariance matrix $C_0$. **Derive** the maximum-a-posteriori-probability (MAP) parameters $w$. **Discuss** under what assumption MAP and MLE coincide.

5. Consider binary classification with logistic regression and training data $\{(t_n, x_n)\}_{n=1}^N$. **Present** the cross-entropy loss. **Derive** that its gradient is equal to:

$$g(w) = -\sum_{n=1}^{N} b(x_n) \cdot \left(t_n - \sigma\left(w^T b(x_n)\right)\right)$$

6. **Present** the ADAM SGD algorithm for the training of $w$. **Discuss** how ADAM is different than standard SGD and its potential merits.

7. Consider K-class Softmax Logistic Regression and training data $\{(t_n, x_n)\}_{n=1}^N$. **Present** the cross-entropy loss $L(W)$, $W \in \mathbb{R}^{D \times K}$. **Present** the gradient matrix of $L(W)$ as a function of $W$ and $\{(t_n, x_n)\}_{n=1}^N$.

8. Given data $\{(x_n \in \mathbb{R}^D)\}_{n=1}^N$, **present** the fixed-point iterations algorithm for L1-PCA:

$$\max_{G \in \mathbb{R}^{D \times K},\ G^T G = I_K} \sum_{n=1}^{N} \|G^T x_n\|_1$$