

COSC 594/650: Optimization for Machine Learning

Homework 4

Instructor: Dongsheng Ding

October 18, 2026

This homework is due on **10/30, 11:59pm**. Format your solution in PDF and submit it on Canvas.

1. (**Logistic regression with quadratic regularization**) Consider the logistic regression optimization problem

$$\underset{w \in \mathbb{R}^d}{\text{minimize}} \quad \frac{1}{n} \sum_{i=1}^n (\log(1 + \exp(w^T x_i)) - y_i w^T x_i) + \lambda \|w\|_2^2, \quad (1)$$

where $\lambda > 0$ is a regularization parameter. In this problem, you will use gradient descent to solve (1) and cross validation to tune λ .

- (a) Download the labeled training portion of the public [Madelon dataset](#) from the UCI Machine Learning Repository. Use the files `madelon_train.data` and `madelon_train.labels`. This dataset contains 2000 labeled samples and $d = 500$ features. The i -th row of the data matrix is the feature vector x_i . The original labels belong to $\{-1, +1\}$; convert them to $\{0, 1\}$ using

$$y_i \leftarrow \frac{y_i + 1}{2}.$$

After this conversion, y_i is the label used in (1).

Solution:

- (b) Using random seed 594, perform a stratified random split of the 2000 samples into a training pool of 1500 samples and a held-out test set of $n_{\text{test}} = 500$ samples. The class proportions should be the same in both sets. From the training pool, construct a stratified training subset $\mathcal{T}_{200}$ containing $n_{\text{tr}} = 200$ samples. Report the number of samples from each class in $\mathcal{T}_{200}$ and in the test set.

Solution:

- (c) Implement gradient descent to solve (1).

Solution:

- (d) Choose a grid Λ of positive candidate values for λ . Using random seed 594, perform stratified 5-fold cross validation entirely within $\mathcal{T}_{200}$. For each $\lambda \in \Lambda$, train on four folds and compute the misclassification rate on the remaining fold; repeat until every fold has served once as the validation fold. Select the value of λ with the smallest mean validation misclassification rate. Report Λ , the mean validation rate for each candidate, the selected value of λ , and your rule for breaking any ties. Do not use the held-out test set for any part of this selection procedure.

Solution:

- (e) Using the selected value of λ , retrain the classifier on all of $\mathcal{T}_{200}$ and report its misclassification rate on the held-out test set.

Solution:

- (f) Construct nested, stratified training sets

$$\mathcal{T}_{200} \subset \mathcal{T}_{700} \subset \mathcal{T}_{1500},$$

where $\mathcal{T}_{700}$ contains 700 samples and $\mathcal{T}_{1500}$ is the complete training pool. Repeat the optimization, cross-validation, and test-evaluation steps independently for $\mathcal{T}_{700}$ and $\mathcal{T}_{1500}$. Use the same candidate grid Λ , random seed, and held-out test set for all three training-set sizes. Tune a separate value of λ for each training set, then retrain on that complete training set before evaluating it on the test set.

Solution:

- (g) Based on your observations and our class discussions, describe in a few sentences the relationship between the selected value of λ and the number of training samples n_{tr} . Also describe the relationship between the test error rate and n_{tr} .

Solution:

2. (**Stochastic gradient method**) Consider the optimization problem in (1) using the training set $\mathcal{T}_{1500}$ and the corresponding value of λ selected in the previous problem. At each iteration k , draw an index j_k uniformly at random from $\{1, \dots, n_{\text{tr}}\}$ and use the stochastic gradient

$$g_k = \left(\frac{\exp(w_k^T x_{j_k})}{1 + \exp(w_k^T x_{j_k})} - y_{j_k} \right) x_{j_k} + 2\lambda w_k.$$

Let f denote the full objective function in (1), evaluated using all $n_{\text{tr}} = 1500$ training samples. Define

$$e_k \triangleq \log \left(\|\nabla f(w_k)\|_2^2 \right). \quad (2)$$

- (a) Implement stochastic gradient descent with a constant step size. Plot e_k versus the iteration number k , and report the constant step size used.

Solution:

- (b) Implement stochastic gradient descent with a diminishing step size. Plot e_k versus the iteration number k , and state the diminishing step-size rule used.

Solution: