CS480/680, Spring 2026

Assignment 1

Designer: Wenyu Bo; Instructor: Hongyang Zhang

Released: May 13; Due: June 8, noon

Instructions

- We do not accept hand-written submissions.
- This assignment is due by noon on June 8, 2026.
- For questions labelled as "**coding**", please follow the instructions provided and implement the required features. Unless otherwise specified, all implementations should be in *Python* using a *Jupyter Notebook*. Before submission, please make sure that your code can run without any errors. Also, be sure to save the output of each cell, as any missing output may not be graded.
- Please submit the following TWO files to LEARN:
  - A write-up in PDF format: the written answers to ALL questions, including the reported results and plots of coding questions, in a single PDF file.
  - An IPYNB file: your implementations for ALL coding questions. Please save the output of each cell, or your coding questions may NOT be graded.

Question 1 (SVM - 10 points)

1. *(4 points) Assume that the data set is linearly separable.*
   (a) *(3 points) Prove that removing any non support vector ($\alpha = 0$) from data set does not change SVM solution using dual conditions.*
   (b) *(1 point) Let f_{-i} be the SVM trained on $D_{-i} = D \setminus \{(x_i, y_i)\}$, and L be the error rate, i.e. $L = \frac{|\{i: f_{-i}(x_i) \neq y_i\}|}{n}$, prove that $L \leq \frac{|support\ vector\ of\ f|}{n}$ where f is the SVM trained on D.*
2. *(6 points) Assume data points are 1 dimensional, and two classes. Consider the polynomial kernel $k(x, x') = (1 + xx')^2$.*
   (a) *(1 point) Find the map $\phi : \mathbb{R} \to \mathbb{R}^d$ such that $k(x, x') = \phi(x)^T \phi(x')$.*
   (b) *(1 point)Let $D = \{(-1, +1), (0, -1), (1, +1)\}$, Using the map ϕ found in the previous part, write down the primal SVM problem for this data set. You need to give the explicit numerical values.*
   (c) *(1 point)Write the expression of the dual SVM problem for this data set. You need to give the explicit numerical values.*
   (d) *(2 points) Solve the dual SVM problem.*
   (e) *(1 point) Write down the classifier $f(x) = sign(w^T \phi(x) + b)$, and report the dicision boundary.*

Question 2 (Regression on MNIST - 15 points) *In this problem, we will implement regularized classifiers for the MNIST data set. The task is to classify handwritten images of numbers between 0 to 9. Each example has features $x_i \in \mathbb{R}^d$ (with $d = 28 * 28 = 784$) and label $z_j \in \{0, \ldots, 9\}$. We wish to learn a predictor $\hat{f}$ that takes as input a vector in $\mathbb{R}^d$ and outputs an index in $\{0, \ldots, 9\}$. We define our training and testing classification error on a predictor f as*

$$\widehat{\epsilon}_{train}(f) = \frac{1}{N_{train}} \sum_{(x,z) \in Training\ Set} \mathbf{1}\{f(x) \neq z\},$$

$$\widehat{\epsilon}_{test}(f) = \frac{1}{N_{test}} \sum_{(x,z) \in Test\ Set} \mathbf{1}\{f(x) \neq z\}.$$

We will use one-hot encoding of the labels: for each observation (x, z), the original label $z \in \{0, \ldots, 9\}$ is mapped to the standard basis vector e_{z+1} where e_i is a vector of size k containing all zeros except for a 1 in the i^{th} position (positions in these vectors are indexed starting at one, hence the $z + 1$ offset for the digit labels). We adopt the notation where we have n data points in our training objective with features $x_i \in \mathbb{R}^d$ and label one-hot encoded as $y_i \in \{0, 1\}^k$. Here, $k = 10$ since there are 10 digits.

1. *(2 points)*

   In this problem we will choose a linear classifier to minimize the regularized least squares objective:

   $$\widehat{W} = argmin_{W \in \mathbb{R}^{d \times k}} \|XW - Y\|_F^2 + \lambda \|W\|_F^2.$$

   Note that $\|W\|_F$ corresponds to the Frobenius norm of W, i.e. $\|W\|_F^2 = \sum_{i=1}^{d} \sum_{j=1}^{k} W_{i,j}^2$. Show that

   $$\widehat{W} = (X^T X + \lambda I)^{-1} X^T Y.$$

2. *(5 points, coding)*

- *Implement a function* `train` *that takes as input* $X \in \mathbb{R}^{n\times d}$, $Y \in \{0,1\}^{n\times k}$, $\lambda > 0$ *and returns* $\widehat{W} \in \mathbb{R}^{d\times k}$.
- *Implement a function* `one_hot` *that takes as input* $Y \in \{0, ..., k-1\}^n$, *and returns* $Y \in \{0,1\}^{n\times k}$.
- *Implement a function* `predict` *that takes as input* $W \in \mathbb{R}^{d\times k}$, $X' \in \mathbb{R}^{m\times d}$ *and returns an* m*-length vector with the* i*th entry equal to* $\arg\max_{j=0,...,9} x_i' W e_{j+1}^T$.
- *Using the functions you coded above, train a model to estimate* $\widehat{W}$ *on the MNIST training data with* $\lambda = 10^{-4}$, *and make label predictions on the test data.*

3. *(*1 *point, coding) What are the training and testing errors of the classifier trained as above?*

4. *(*7 *points) Consider softmax classifier, the loss function is defined as*

$$l = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{\exp(x_i W_{y_i})}{\sum_{l=1}^{k} \exp(x_i W_l)} + \frac{\lambda}{2}\|W\|_F^2.$$

where W_{y_i} *is the* $y_i + 1$*-th column of* W *(label* 0 *means the first column),* W_l *is the* l*-th column of* W.

- *(*2 *points) Derive the gradient of the loss function with respect to* W.
- *(*3 *points, coding) Implement a function* `grad_descent` *that takes as imput* $X \in \mathbb{R}^{n\times d}$, $Y \in \{0,1\}^{n\times k}$, $\lambda > 0$, *learning rate* $\eta > 0$, *and number of iteration* T, *and returns* $\widehat{W} \in \mathbb{R}^{d\times k}$ *after performing* T *steps of gradient descent.*

  Using the function you coded above, train a softmax classifier on the MNIST training data with $\lambda = 10^{-4}$, *learning rate* $\eta = 0.1$, *and* $T = 500$, *and make label predictions on the test data (The prediction function is the same as in part (2)).*
- *(*2 *points, coding) Report the training and testing errors of the resulting classifier, and plot the training loss as a function of the iteration number.*