CSE 6740 Homework 3

Kai Wang, Fall 2024

Deadline: Nov 3rd, 11:59pm ET

- Submit your answers as an electronic copy on gradescope.
- Late homework incurs a penalty of 20% for each 24 hours that it is late. Thus, right after the deadline it will only be worth 80% credit and after five days it will not be worth any credit.
- We recommend the use of LaTeX for typing up your solutions. No credit will be given to unreadable handwriting.
- List explicitly with whom in the class you discussed which problem, if any. Cite all external ressources that you were using to complete the homeworks. For details, consult the collaboration policy in the class syllabus on canvas.
- Recommended reading: PRML[1] Section 3.1, 3.2

1 Linear Regression [30 pts]

In class, we derived a closed form solution (normal equation) for linear regression problem: $\hat{\theta} = (X^T X)^{-1} X^T Y$. A probabilistic interpretation of linear regression tells us that we are relying on an assumption that each data point is actually sampled from a linear hyperplane, with some noise. The noise follows a zero-mean Gaussian distribution with constant variance. Specifically,

$$Y^i = \theta^T X^i + \epsilon^i \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, $\theta \in \mathbb{R}^d$, and $\{X^i, Y^i\}$ is the i-th data point. In other words, we are assuming that each every point is independent to each other and that every data point has same variance.

(a) Using the normal equation, and the model (Eqn. 1), derive the expectation $\mathbb{E}[\hat{\theta}]$. Note that here X is fixed, and only Y is random, i.e. "fixed design" as in statistics. [6 pts]

(b) Similarly, derive the variance $\mathbf{Var}[\hat{\theta}]$. [6 pts]

(c) Under the white noise assumption above, someone claims that $\hat{\theta}$ follows Gaussian distribution with mean and variance in (a) and (b), respectively. Do you agree with this claim? Why or why not? [8 pts]

(d) Weighted linear regression

Suppose we keep the independence assumption but remove the same variance assumption. In other words, data points would be still sampled independently, but now they may have different variance σ_i. Thus, the

[1] Christopher M. Bishop, Pattern Recognition and Machine Learning, 2006, Springer.

covariance matrix of Y would be still diagonal, but with different values:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}. \tag{2}$$

Derive the estimator $\hat{\theta}$ (similar to the normal equations) for this problem using matrix-vector notations with Σ. [10 pts]

2 Regression contd. [10 pts]

2.1 Ridge Regression [5 pts]

For linear regression, it is often assumed that $y = \theta^\top \mathbf{x} + \epsilon$ where $\theta, \mathbf{x} \in \mathbb{R}^m$ by absorbing the constant term, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian random variable. Given n i.i.d samples $(\mathbf{x}^1, y^1), ..., (\mathbf{x}^n, y^n)$, we define $\mathbf{y} = (y^1, ..., y^n)^\top$ and $X = (\mathbf{x}^1, ..., \mathbf{x}^n)^\top$. Thus, we have $\mathbf{y} \sim \mathcal{N}(X\theta, \sigma^2 I)$. Show that the ridge regression estimate is the mean of the posterior distribution under a Gaussian prior $\theta \sim \mathcal{N}(0, \tau^2 I)$. Find the explicit relation between the regularization parameter λ in the ridge regression estimate of the parameter θ, and the variances σ^2, τ^2.

2.2 Tikhonov regularization [5 pts]

Consider the following objective :

$$\min_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{y} - \mathbf{X}^\top \theta\|_2^2 + \|\Gamma\theta\|_2^2$$

where $\mathbf{y} \in \mathbb{R}^n, \theta \in \mathbb{R}^d, \mathbf{X} \in \mathbb{R}^{d \times n}$ and $\Gamma \in \mathbb{R}^{m \times d}$.

(a) Show that $\Gamma^\top \Gamma$ is symmetric and positive semidefinite (psd). Also derive the optimal solution of this objective. [3 pts]

(b) For what values of Γ does this problem reduce to the ridge regression case? And discuss about the pros/cons of Tikhonov regularization over ridge regression. [2 pts]

3 Programming (Warmup): Regression [10 pts]

Take a look at the `regression.ipynb` notebook.

(a) Implement Linear Regression. Present the mean-squared error (MSE) loss value to notice the error between the predictions on the test set and the ground truth labels from the training set. [3 pts]

(b) Similarly, implement Ridge Regression. Use different values of the regularization parameter λ such as $\lambda = 0.1, 0.5, 1.0, 5.0, 10.0$ and turn in your scatter plot with λ on the x-axis (please scale the x-axis accordingly). Report the mean-squared error (MSE) loss value on the y-axis. [5 pts]

(c) Compare both these approaches. Please discuss which method performs better and why. [2 pts]

4 Programming: LR, SVM, NN, CNN [20 pts]

Please take a look at the `MNIST_all_methods.ipynb` notebook.

Implement the model, training loop and the evaluation code for the four approaches : (a) Logistic Regression (LR) [5 pts], (b) Support Vector Machine (SVM) [5 pts], (c) Neural Network (NN) [5 pts] and (d) Convolutional Neural Network (CNN) [5 pts].

Note: For the evaluation code, it suffices to use the `evaluate` function defined in the notebook.

Keep the hyperparameter values fixed, as given in the notebook.

Model: Use atmost two convolutional layers for your CNN implementation and atmost one intermediate linear layer for your NN implementation. The LR and SVM implementation may take only one linear layer with the input size and output classes as the parameters.

Finally, compare the test accuracies obtained upon evaluation across all the four methods and report your observations.

5 Programming: Recommendation System [30 pts]

Personalized recommendation systems are used in a wide variety of applications such as electronic commerce, social networks, web search, and more. Machine learning techniques play a key role to extract individual preference over items. In this assignment, we explore this popular business application of machine learning, by implementing a simple matrix-factorization-based recommender using gradient descent.

Suppose you are an employee in Netflix. You are given a set of ratings (from one star to five stars) from users on many movies they have seen. Using this information, your job is implementing a personalized rating predictor for a given user on unseen movies. That is, a rating predictor can be seen as a function $f : \mathcal{U} \times \mathcal{I} \to \mathbb{R}$, where $\mathcal{U}$ and $\mathcal{I}$ are the set of users and items, respectively. Typically the range of this function is restricted to between 1 and 5 (stars), which is the the allowed range of the input.

Now, let's think about the data representation. Suppose we have m users and n items, and a rating given by a user on a movie. We can represent this information as a form of matrix, namely rating matrix M. Suppose rows of M represent users, while columns do movies. Then, the size of matrix will be $m \times n$. Each cell of the matrix may contain a rating on a movie by a user. In $M_{15,47}$, for example, it may contain a rating on the item 47 by user 15. If he gave 4 stars, $M_{15,47} = 4$. However, as it is almost impossible for everyone to watch large portion of movies in the market, this rating matrix should be very sparse in nature. Typically, only 1% of the cells in the rating matrix are observed in average. All other 99% are missing values, which means the corresponding user did not see (or just did not provide the rating for) the corresponding movie. Our goal with the rating predictor is estimating those missing values, reflecting the user's preference learned from available ratings.

Our approach for this problem is matrix factorization. Specifically, we assume that the rating matrix M is a low-rank matrix. Intuitively, this reflects our assumption that there is only a small number of factors (e.g, genre, director, main actor/actress, released year, etc.) that determine like or dislike. Let's define r as the number of factors. Then, we learn a user profile $U \in \mathbb{R}^{m \times r}$ and an item profile $V \in \mathbb{R}^{n \times r}$. (Recall that m and n are the number of users and items, respectively.) We want to approximate a rating by an inner product of two length r vectors, one representing user profile and the other item profile. Mathematically, a rating by user u on movie i is approximated by

$$M_{u,i} \approx \sum_{k=1}^{r} U_{u,k} V_{i,k}. \tag{3}$$

We want to fit each element of U and V by minimizing squared reconstruction error over all training data points. That is, the objective function we minimize is given by

$$E(U,V) = \sum_{(u,i)\in M} (M_{u,i} - U_u^T V_i)^2 = \sum_{(u,i)\in M} (M_{u,i} - \sum_{k=1}^{r} U_{u,k} V_{i,k})^2 \tag{4}$$

where U_u is the uth row of U and V_i is the ith row of V. We observe that this looks very similar to the linear regression. Recall that we minimize in linear regression:

$$E(\theta) = \sum_{i=1}^{m} (Y^i - \theta^T x^i)^2 = \sum_{i=1}^{m} (Y^i - \sum_{k=1}^{r} \theta_k x_k^i)^2 \tag{5}$$

where m is the number of training data points. Let's compare (4) and (5). $M_{u,i}$ in (4) corresponds to Y^i in (5), in that both are the observed labels. $U_u^T V_i$ in (4) corresponds to $\theta^T x^i$ in (5), in that both are our estimation with our model. The only difference is that both U and V are the parameters to be learned in (4), while only θ is learned in (5). This is where we personalize our estimation: with linear regression, we apply the same θ to any input x^i, but with matrix factorization, a different profile U_u are applied depending on who is the user u.

As U and V are interrelated in (4), there is no closed form solution, unlike linear regression case. Thus, we need to use gradient descent:

$$U_{v,k} \leftarrow U_{v,k} - \mu \frac{\partial E(U,V)}{\partial U_{v,k}}, \qquad V_{j,k} \leftarrow V_{j,k} - \mu \frac{\partial E(U,V)}{\partial V_{j,k}}, \tag{6}$$

where μ is a hyper-parameter deciding the update rate. It would be straightforward to take partial derivatives of $E(U,V)$ in (4) with respect to each element $U_{v,k}$ and $V_{j,k}$. Then, we update each element of U and V using the gradient descent formula in (6).

(a) Derive the update formula in (6) **by solving the partial derivatives. [10 pts]**

(b) To avoid overfitting, we usually add regularization terms, which penalize for large values in U and V. Redo part (a) using the regularized objective function below. [5 pts]

$$E(U,V) = \sum_{(u,i)\in M} (M_{u,i} - \sum_{k=1}^{r} U_{u,k} V_{i,k})^2 + \lambda \sum_{u,k} U_{u,k}^2 + \lambda \sum_{i,k} V_{i,k}^2$$

(λ is a hyper-parameter controlling the degree of penalization.)

(c) Fill out the gradient descent part.

You are given a skeleton code. Using the training data `rateMatrix`, you will implement your own recommendation system of rank `lowRank`. The only function you need to edit is `my_recommender`. In the gradient descent part, repeat your update formula in (b), observing the average reconstruction error between your estimation and ground truth in training set. You need to set a stopping criteria, based on this reconstruction error as well as the maximum number of iterations. You should play with several different values for μ and λ to make sure that your final prediction is accurate.

Formatting information is here:

Input

- **rateMatrix**: training data set. Each row represents a user, while each column an item. Observed values are one of $\{1,2,3,4,5\}$, and missing values are 0.
- **lowRank**: the number of factors (dimension) of your model. With higher values, you would expect more accurate prediction.

Output

- **U**: the user profile matrix of dimension user count × low rank.
- **V**: the item profile matrix of dimension item count × low rank.

Evaluation [10 pts]

To test your code, try to run the cell below the function `my_recommender`. You may have noticed that the code prints both training and test error, in RMSE (Root Mean Squared Error), defined as follows:

$$\sum_{(u,i)\in M} (M_{u,i} - f(u,i))^2$$

where $f(u,i)$ is your estimation, and the summation is over the training set or testing set, respectively. For the grading, we will use another set-aside testing set, which is not released to you. If you observe your test error is less than 1.00 without cheating (that is, training on the test set), you may expect to see the similar performance on the unseen test set as well.

Grading criteria:

- Your code should output U and V as specified. The dimension should match to the specification.
- We will test your output on another test dataset, which was not provided to you. The test RMSE on this dataset should be at least 1.05 to get at least partial credit.
- We will measure elapsed time for learning. If your implementation takes longer than 3 minutes for rank 5, you should definitely try to make your code faster or adjust parameters. Any code running more than 5 minutes is not eligible for credit.
- Your code should not crash. Any crashing code will be not credited.

Report [5 pts]

In your report, show the performance (RMSE) both on your training set and test set, with varied `lowRank`. (The default is set to 1, 3, and 5, but you may want to vary it further.) Discuss what you observe with varied low rank. Also, briefly discuss how you decided your hyper-parameters (μ, λ).

Note

- Do not print anything in your code (e.g, iteration 1 : err=2.4382) in your final submission.
- Do not alter input and output format of the skeleton file. (E.g, adding a new parameter without specifying its defalut value) Please make sure that you returned all necessary outputs according to the given skeleton.
- Please do not use any additional file. This task is simple enough that you can fit in just one file.
- Submit your code with the best parameters you found. We will grade without modifying your code. (Applying cross-validation to find best parameters is fine, though you do not required to do.)
- Please be sure that your program finishes within a fixed number of iterations. Always think of a case where your stopping criteria is not satisfied forever. This can happen anytime depending on the data, not because your code is incorrect. For this, we recommend setting a maximum number of iteration in addition to other stopping criteria.

Grand Prize

Similar to the Netflix competition held in 2006 to 2009, the student who achives the lowest RMSE on the held out test set will earn the Grand Prize. We will award extra 10 bonus points to the winner, and share the student's code to everyone. (Note that the winner should satisfy all other grading criteria perfectly, including answer sanity check and timing requirement. Otherwise, the next student will be considered as the winner.)