CSE 6740 Homework 3
Yuzhou Wang

Problem 1 Linear Regression

(a)

Solution:

Let $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$, $Y = (Y^1, \ldots, Y^n)^\top$, then we have $Y = X\theta + \varepsilon$, where X is the matrix with i^{th} row being X^i. Notice that $\mathbb{E}[\varepsilon]$ is the zero vector, We can compute the expectation:

$$\begin{aligned}\mathbb{E}[\hat{\theta}] &= \mathbb{E}\big[(X^\top X)^{-1}X^\top Y\big] = \mathbb{E}\big[(X^\top X)^{-1}X^\top(X\theta+\varepsilon)\big]\\ &= (X^\top X)^{-1}X^\top X\theta + (X^\top X)^{-1}X^\top\mathbb{E}[\varepsilon]\\ &= \theta.\end{aligned}$$

(b)

Solution:

As given, we have $\mathbb{V}\mathrm{ar}[\varepsilon] = \mathbb{E}[\varepsilon\varepsilon^\top] = \sigma^2 I$. It is also clear,from the previous part, that

$$\mathbb{V}\mathrm{ar}[\hat{\theta}] = \mathbb{E}[\hat{\theta}\hat{\theta}^\top] - \mathbb{E}[\hat{\theta}]^2 = \mathbb{E}[\hat{\theta}\hat{\theta}^\top] - \theta^2.$$

So we can focus the second moment. Also from the previous part, we can see that

$$\hat{\theta} = (X^\top X)^{-1}X^\top(X\theta+\varepsilon) = \theta + (X^\top X)^{-1}X^\top\varepsilon \tag{1}$$

Using this gives

$$\begin{aligned}\mathbb{E}[\hat{\theta}\hat{\theta}^\top] &= \theta^2 - 2\mathbb{E}[\theta((X^\top X)^{-1}X^\top\varepsilon)^\top] + \mathbb{E}[(X^\top X)^{-1}X^\top\varepsilon((X^\top X)^{-1}X^\top\varepsilon)^\top]\\ &= \theta^2 + (X^\top X)^{-1}X^\top\mathbb{E}[\varepsilon\varepsilon^\top]X(X^\top X)^{-1}\\ &= \theta^2 + (X^\top X)^{-1}X^\top\sigma^2 IX(X^\top X)^{-1}\\ &= \theta^2 + \sigma^2(X^\top X)^{-1}.\end{aligned}$$

In conclusion we have

$$\mathbb{V}\mathrm{ar}[\hat{\theta}] = \theta^2 + \sigma^2(X^\top X)^{-1} - \theta^2 = \sigma^2(X^\top X)^{-1}.$$

(c)

Solution:

Again we observed before in equality (1), $\hat{\theta}$ is a linear transformation of ε. Under the assumption that ε is follows a multivariate Gaussian distribution $\mathcal{N}(0, \sigma^2 I)$, $\hat{\theta}$ also follows a multivariate Gaussian distribution, more specifically, we have $\hat{\theta} \sim \mathcal{N}(\theta, \sigma^2(X^\top X)^{-1})$.

If each data point has a different noise variance, in other words, $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ for all $i \in [n]$, then we have $\varepsilon \sim \mathcal{N}(0, \Sigma)$, where

$$\Sigma = \begin{pmatrix}\sigma_1^2 & 0 & \cdots & 0\\ 0 & \sigma_2^2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \sigma_n^2.\end{pmatrix}$$

Since we still have $\mathbb{E}[\varepsilon]$ is the zero vector, the expectation of $\hat{\theta}$ is unchanged, that is, $\mathbb{E}[\hat{\theta}] = \theta$.

And we repeat the process of computing the second moment, the only place will be different is that $\mathbb{E}[\varepsilon\varepsilon^\top] = \Sigma$ in this case,

$$\begin{aligned}\mathbb{E}[\hat{\theta}\hat{\theta}^\top] &= \theta^2 - 2\mathbb{E}[\theta((X^\top X)^{-1}X^\top\varepsilon)^\top] + \mathbb{E}[(X^\top X)^{-1}X^\top\varepsilon((X^\top X)^{-1}X^\top\varepsilon)^\top] \\ &= \theta^2 + (X^\top X)^{-1}X^\top\mathbb{E}[\varepsilon\varepsilon^\top]X(X^\top X)^{-1} \\ &= \theta^2 + (X^\top X)^{-1}X^\top\Sigma X(X^\top X)^{-1}.\end{aligned}$$

Hence $\mathbb{V}\text{ar}[\hat{\theta}] = (X^\top X)^{-1}X^\top\Sigma X(X^\top X)^{-1}$.

When assuming that each every point is independent to each other and that every data point has mean zero and same variance, the Gauss–Markov theorem states that the ordinary least squares (OLS) estimator has the lowest sampling variance within the class of linear unbiased estimators. If each data point has a different noise variance, the OLS is still valid ($\mathbb{E}[\hat{\theta}] = \theta$) but it won't have the lowest sampling variance.

(d)

1.False. For linear regression, the cost function (such as Mean Squared Error) is convex in terms of the parameters θ. This means that the cost function has a single global minimum and no local minima. Therefore, gradient descent, when applied to linear regression, will always converge to the global minimum as long as the learning rate is appropriately chosen.

2.True. If is the initial θ is exactly at the global minimum, the gradient of the cost function at that point is zero, hence one iteration will not change θ.

Problem 2 Ridge Regression

(a)

Proof. The posterior distribution of θ is $p(\theta|X,y)$, using Bayes Rule gives

$$p(\theta|X,y) = \frac{p(y|\theta,X)p(\theta|X)}{p(y|X)}.$$

Notice that the denominator is a constant independent of θ, so it is irrelevant for find the mean of posterior distributions. We are given that $y \sim \mathcal{N}(X\theta, \sigma^2 I)$ and $\theta \sim \mathcal{N}(0, \tau^2 I)$, which means

$$\begin{aligned} p(y|X,\theta) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\|y - X\theta\|^2\right) \\ p(\theta|X) = p(\theta) &= \frac{1}{(2\pi\tau^2)^{m/2}} \exp\left(-\frac{1}{2\tau^2}\|\theta\|^2\right). \end{aligned}$$

Again, ignoring the normalizing constants independent of θ, we have

$$\begin{aligned} p(\theta|X,y) &\propto \exp\left(-\frac{1}{2\sigma^2}\|y - X\theta\|^2 - \frac{1}{2\tau^2}\|\theta\|^2\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}y^\top y + \frac{1}{\sigma^2}y^\top X\theta - \frac{1}{2\sigma^2}\theta^\top X^\top X\theta - \frac{1}{2\tau^2}\theta^\top\theta\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}y^\top y + \frac{1}{\sigma^2}y^\top X\theta - \theta^\top\left(\frac{1}{2\sigma^2}X^\top X + \frac{1}{2\tau^2}I\right)\theta\right) \end{aligned}$$

Let

$$\begin{aligned} \Sigma &= \left(\frac{1}{\sigma^2}X^\top X + \frac{1}{\tau^2}I\right)^{-1} \\ \mu &= \left(\frac{1}{\sigma^2}X^\top X + \frac{1}{\tau^2}I\right)^{-1} \frac{X^\top y}{\sigma^2} = \left(X^\top X + \frac{\sigma^2}{\tau^2}I\right)^{-1} X^\top y, \end{aligned}$$

Then we have, again manipulating the constants independent of θ,

$$p(\theta|X,y) \propto \exp\left(-\frac{1}{2}(\theta-\mu)^\top\Sigma^{-1}(\theta-\mu)\right),$$

which is Guassian distribution.

Thus the mean of $p(\theta|X,y)$ is μ, which the ridge regression estimate with $\lambda = \frac{\sigma^2}{\tau^2}$. ■

(b)

Solution:

The K-Fold Cross-Validation algorithm :

Initialize:

- Create an empty list `mse_scores` to store the average MSE for each λ.

Loop over each $\lambda \in \Lambda$:

- For each λ, perform K-Fold Cross-Validation:
  - Split the dataset X and y into K folds.
  - Initialize an empty list `fold_mses` to store the MSE for each fold.
  - Perform the following K-Fold Cross-Validation step.

K-Fold Cross-Validation:

- For each fold $k = 1, \ldots, K$:
  - **Split the data:**
    - Let X_{train} and y_{train} be the training data (all folds except fold k).
    - Let X_{val} and y_{val} be the validation data (fold k).
  - **Train the Ridge Regression model:**
    - Fit the Ridge Regression model on X_{train} and y_{train} using the current λ.
  - **Evaluate the model:**
    - Predict the target values for the validation set: $\hat{y}_{\text{val}} = X_{\text{val}}\theta_{\text{ridge}}$.
    - Compute the MSE for the validation set:

      $$\text{MSE}_k = \frac{1}{m}\sum_{i=1}^{m}(y_{\text{val},i} - \hat{y}_{\text{val},i})^2,$$

      where m is the number of samples in the validation set.
    - Append MSE_k to `fold_mses`.

Compute the average MSE:

- Compute the average MSE across the K folds:

$$\text{avg\_mse} = \frac{1}{K}\sum_{k=1}^{K}\text{MSE}_k.$$

- Append avg_mse to `mse_scores`.

Select the best λ:

$$\lambda_{\text{best}} = \arg\min_{\lambda\in\Lambda} \text{mse\_scores}.$$

Return λ_{best}.

Problem 3 Bayes Classifier

3.1 Joint Bayes vs Naive Bayes Classifier

(a)

Solution: According to the joint Bayes classifier: $\mathbb{P}(K = 1|a = 1 \wedge b = 1 \wedge c = 0) = \frac{1}{2}$, since there are 2 rows (5th and 9th) with $a = 1 \wedge b = 1 \wedge c = 0$ only in the 5th row $K = 1$.

Similarly, $\mathbb{P}(K = 0|a = 1 \wedge b = 1) = \frac{1}{3}$, since there are 3 rows (4th, 5th and 9th) with $a = 1 \wedge b = 1$ and $K = 1$ in the 4th and 5th row.

(b)

Solution: According to the naive Bayes classifier

$$\begin{aligned}
\mathbb{P}(K = 1|a = 1 \wedge b = 1 \wedge c = 0) &= \frac{\mathbb{P}(a = 1 \wedge b = 1 \wedge c = 0|K = 1)\mathbb{P}(K = 1)}{\mathbb{P}(a = 1 \wedge b = 1 \wedge c = 0)} \\
&= \frac{\mathbb{P}(a = 1|K = 1)\mathbb{P}(b = 1|K = 1)\mathbb{P}(c = 0|K = 1)\mathbb{P}(K = 1)}{\mathbb{P}(a = 1 \wedge b = 1 \wedge c = 0)} \\
&= \frac{4}{5} \times \frac{2}{5} \times \frac{2}{5} \times \frac{5}{9} \times \frac{9}{2} = \frac{8}{25}.
\end{aligned}$$

For the second probability, we follow the similar procedure as above,

$$\begin{aligned}
\mathbb{P}(K = 0|a = 1 \wedge b = 1) &= \frac{\mathbb{P}(a = 1 \wedge b = 1|K = 0)\mathbb{P}(K = 0)}{\mathbb{P}(a = 1 \wedge b = 1)} \\
&= \frac{\mathbb{P}(a = 1|K = 0)\mathbb{P}(b = 1|K = 0)\mathbb{P}(K = 0)}{\mathbb{P}(a = 1 \wedge b = 1)} \\
&= \frac{1}{4} \times \frac{3}{4} \times \frac{4}{9} \times \frac{9}{3} = \frac{1}{5}.
\end{aligned}$$

3.2 Bayes Classifier with Gaussian Class Conditional Distribution

(a)

Solution:

Assume that the class-conditional distributions are multivariate Gaussians of dimension d:

$$p(x|y = i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i)\right), \quad i \in \{1, -1\},$$

where μ_i and Σ_i are the mean and covariance matrix of class of $y = i$.

Taking the logarithm of the ratio of class-conditional densities:

$$\begin{aligned}
h(x) &= \log \frac{p(x|y = 1)p(y = 1)}{p(x|y = -1)p(y = 1)} \\
&= \log \frac{\frac{1}{(2\pi)^{d/2}|\Sigma_1|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_1)^T \Sigma_1^{-1}(x - \mu_1)\right)}{\frac{1}{(2\pi)^{d/2}|\Sigma_{-1}|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_{-1})^T \Sigma_{-1}^{-1}(x - \mu_{-1})\right)} + \log \frac{p(y = 1)}{p(y = -1)} \\
&= -\frac{1}{2}(x - \mu_1)^T \Sigma_1^{-1}(x - \mu_1) + \frac{1}{2}(x - \mu_{-1})^T \Sigma_{-1}^{-1}(x - \mu_{-1}) \underbrace{-\frac{1}{2}\log \frac{|\Sigma_1|}{|\Sigma_{-1}|} + \log \frac{p(y = 1)}{p(y = -1)}}_{=c}.
\end{aligned}$$

The decision boundary is defined by setting $h(x) = 0$, which gives

$$(x - \mu_1)^T \Sigma_1^{-1}(x - \mu_1) - (x - \mu_{-1})^T \Sigma_{-1}^{-1}(x - \mu_{-1}) + c = 0,$$

where c is a constant depending on priors and determinants of covariance matrices as indicated above.

Since this equation involves quadratic terms in x, the decision boundary is a **quadratic surface**. In two dimensions, this means the boundary could take the form of an ellipse, hyperbola, or parabola.

(b)

Solution: If the two Gaussians have identical covariance matrices, i.e., $\Sigma_1 = \Sigma_{-1} = \Sigma$, the discriminant function simplifies as follows:

$$\begin{aligned} h(x) &= -\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_{-1})^T\Sigma^{-1}(x-\mu_{-1}) + \log\frac{p(y=1)}{p(y=-1)} \\ &= -\frac{1}{2}\left(x^T\Sigma^{-1}x - 2\mu_1^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}\mu_1\right) + \frac{1}{2}\left(x^T\Sigma^{-1}x - 2\mu_{-1}^T\Sigma^{-1}x + \mu_{-1}^T\Sigma^{-1}\mu_{-1}\right) + \log\frac{p(y=1)}{p(y=-1)} \\ &= \left(\mu_1 - \mu_{-1}\right)^T\Sigma^{-1}x + \underbrace{\frac{1}{2}(\mu_{-1}^T\Sigma^{-1}\mu_{-1} - \mu_1^T\Sigma^{-1}\mu_1) + \log\frac{p(y=1)}{p(y=-1)}}_{=c} \end{aligned}$$

The decision boundary is defined by setting $h(x) = 0$, which gives

$$\left(\mu_1 - \mu_{-1}\right)^T\Sigma^{-1}x + c = 0,$$

where c is a constant that depends on μ_1, μ_{-1}, Σ, and the priors as indicated above. Since this equation is linear in x, the decision boundary is a **hyperplane**. In two dimensions, this corresponds to a straight line.

(c)

Solution:

If the covariance matrix of both are the identity matrix, i.e., $\Sigma = I$, then the discriminant function further simplifies to:

$$h(x) = (\mu_1 - \mu_{-1})^T x + \underbrace{\frac{1}{2}(\mu_{-1}^T\mu_{-1} - \mu_1^T\mu_1) + \log\frac{p(y=1)}{p(y=-1)}}_{=c}$$

The decision boundary is defined by setting $h(x) = 0$, which gives

$$\left(\mu_1 - \mu_{-1}\right)^T x + c = 0,$$

where c is a constant that depends on μ_1, μ_{-1}, and the priors as indicated above. Since this equation is linear in x, the decision boundary is a **hyperplane**. In two dimensions, this corresponds to a straight line.

Problem 4 Basics of optimization

(a)

Proof. First we compute the derivative of the sigmoid function $\sigma(z) = \frac{1}{1+e^{-z}}$:

$$\sigma'(z) = -(1+e^{-z})^{-2} \cdot (-e^{-z}) = \frac{e^{-z}}{(1+e^{-z})^2} = \sigma(z)(1-\sigma(z)).$$

To show $L(z)$ is convex, we just need to show its second derivative is ≥ 0.

$$\begin{aligned} L'(z) &= -y\frac{\sigma'(z)}{\sigma(z)} - (1-y)\frac{-\sigma'(z)}{1-\sigma(z)} \\ &= -y(1-\sigma(z)) + (1-y)\sigma(z) \\ &= \sigma(z) - y. \end{aligned}$$

and

$$L''(z) = \sigma'(z) = \frac{e^{-z}}{(1+e^{-z})^2} \geq 0.$$

It implies that the binary cross-entropy loss function is convex in z when $y \in \{0,1\}$. ■

(b)

Proof. We can use induction to prove the Jansen's Inequality:

Base Case: For $m = 2$, the concavity of f implies that for any $\alpha_1, \alpha_2 \geq 0$ with $\alpha_1 + \alpha_2 = 1$,

$$f(\alpha_1 x_1 + \alpha_2 x_2) \geq \alpha_1 f(x_1) + \alpha_2 f(x_2).$$

Inductive Step: Assume that the inequality holds for some m, i.e.,

$$f\left(\sum_{i=1}^{m} \alpha_i x_i\right) \geq \sum_{i=1}^{m} \alpha_i f(x_i).$$

We need to show that it holds for $m+1$. Consider weights $\alpha_1, \ldots, \alpha_{m+1}$ such that $\sum_{i=1}^{m+1} \alpha_i = 1$. Define

$$\beta = \sum_{i=1}^{m} \alpha_i, \quad \text{so that} \quad \alpha_{m+1} = 1 - \beta.$$

We can rewrite the weighted sum of the first m term as:

$$y = \sum_{i=1}^{m} \frac{\alpha_i}{\beta} x_i.$$

By the induction hypothesis, we apply the assumption to y:

$$f(y) \geq \sum_{i=1}^{m} \frac{\alpha_i}{\beta} f(x_i).$$

Using concavity for the two points y and x_{m+1}, we write:

$$f\left(\beta y + \alpha_{m+1} x_{m+1}\right) \geq \beta f(y) + \alpha_{m+1} f(x_{m+1}).$$

Substituting the induction hypothesis,

$$f\left(\sum_{i=1}^{m+1}\alpha_i x_i\right) \geq \beta\sum_{i=1}^{m}\frac{\alpha_i}{\beta}f(x_i) + \alpha_{m+1}f(x_{m+1}) = \sum_{i=1}^{m+1}\alpha_i f(x_i).$$

Thus, the inequality holds for $m+1$, completing the induction step. By induction, the inequality holds for all $m \geq 2$.

■

Problem 5 Logistic Regression

5.1 Binary Classification

(a)

Solution: The probability of $Y = 1$ given $X = x$ in logistic regression is given by:

$$P(Y = 1|X = x) = \frac{\exp(w_0 + w^T x)}{1 + \exp(w_0 + w^T x)}.$$

The decision boundary is defined where the probability of $Y = 1$ equals the probability of $Y = 0$, i.e.,

$$P(Y = 1|X = x) = P(Y = 0|X = x) = 0.5.$$

Simplify it gives

$$\begin{aligned} &\exp(w_0 + w^T x) = 1 \\ \implies &w_0 + w^T x = 0. \end{aligned}$$

The equation $w_0 + w^T x = 0$ represents a **hyperplane** in the feature space.

5.2 Multi-class Classification

(a)

Solution: The new regularized loss function is

$$\mathcal{L}(\theta) = -\sum_{i=1}^{m}\sum_{c=1}^{C} y_c^i \theta_c^T x^i + \sum_{i=1}^{m}\sum_{c=1}^{C} y_c^i \log\left(\sum_{c'=1}^{C}\exp(\theta_{c'}^T x^i)\right) + \frac{\lambda}{2}\sum_{c=1}^{C}\|\theta_c\|_2^2$$

And the gradient w.r.t θ_c is

$$\begin{aligned} \nabla_{\theta_c}\mathcal{L} &= -\sum_{i=1}^{m} y_c^i x^i + \sum_{i=1}^{m}\sum_{c=1}^{C} y_c^i \frac{\exp(\theta_c^\top x^i)}{\sum_{c'=1}^{C}\exp(\theta_{c'}^T x^i)} \cdot x^i + \lambda\theta_c \\ &= -\sum_{i=1}^{m} y_c^i x^i + \sum_{i=1}^{m} \frac{\exp(\theta_c^\top x^i)}{\sum_{c'=1}^{C}\exp(\theta_{c'}^T x^i)} \cdot x^i + \lambda\theta_c \\ &= \sum_{i=1}^{m}\left(\frac{\exp(\theta_c^\top x^i)}{\sum_{c'=1}^{C}\exp(\theta_{c'}^T x^i)} - y_c^i\right) x^i + \lambda\theta_c. \end{aligned}$$

Recommendation System

(a)

Solution: Let I_v be the set of items rated by user v and U_j be the set of users rated item j, we have

$$\begin{aligned}
\frac{\partial E(U,V)}{\partial U_{v,\ell}} &= \sum_{i \in I_v} 2\left(M_{v,i} - \sum_{k=1}^{r} U_{v,k} V_{i,k}\right) \cdot (-V_{i,\ell}) = -2 \sum_{i \in I_v} \left(M_{v,i} - \sum_{k=1}^{r} U_{v,k} V_{i,k}\right) V_{i,\ell} \\
\frac{\partial E(U,V)}{\partial V_{j,\ell}} &= \sum_{u \in U_j} 2\left(M_{u,j} - \sum_{k=1}^{r} U_{u,k} V_{j,k}\right) \cdot (-U_{u,\ell}) = -2 \sum_{u \in U_j} \left(M_{u,j} - \sum_{k=1}^{r} U_{u,k} V_{j,k}\right) U_{u,\ell}
\end{aligned}$$

(b)

Solution: After adding the regularization terms, we have

$$\begin{aligned}
\frac{\partial E(U,V)}{\partial U_{v,\ell}} &= -2 \sum_{i \in I_v} \left(M_{v,i} - \sum_{k=1}^{r} U_{v,k} V_{i,k}\right) V_{i,\ell} + 2\lambda U_{v,\ell} \\
\frac{\partial E(U,V)}{\partial V_{j,\ell}} &= -2 \sum_{u \in U_j} \left(M_{u,j} - \sum_{k=1}^{r} U_{u,k} V_{j,k}\right) U_{u,\ell} + 2\lambda V_{j,\ell}
\end{aligned}$$

(c)

Solution: In the enhanced model, for any fixed user v and any fixed item j, we have

$$\begin{aligned}
\frac{\partial E(U,V,b_u,b_i)}{\partial b_v} &= -2 \sum_{i \in I_v} \left(M_{v,i} - \mu - b_v - b_i - \sum_{k=1}^{r} U_{v,k} V_{i,k}\right) + 2\lambda b_v \\
\frac{\partial E(U,V,b_u,b_i)}{\partial b_j} &= -2 \sum_{u \in U_j} \left(M_{u,j} - \mu - b_u - b_j - \sum_{k=1}^{r} U_{u,k} V_{j,k}\right) + 2\lambda b_u
\end{aligned}$$

Report

Below is the table of train/test RMSE and running time for different ranks.

Configuration	train_rmse	test_rmse	time(s)
SVD-noReg-1	0.8632	0.9241	129.37
SVD-withReg-1	0.8709	0.9186	119.69
SVD-noReg-3	0.8081	0.9552	129.19
SVD-withReg-3	0.8272	0.9087	122.41
SVD-noReg-5	0.7522	0.9820	130.22
SVD-withReg-5	0.7967	0.9129	124.93

Table 1: Matrix Factorization Results with Different Configurations

Observation of varied low rank: Small rank(e.g., 1, 3) may lead to underfitting because the model is too simple and lacks the capacity to capture user-item interactions. So it is not performing as well in the training data as hight rank (i.e rank 5).

hyper-parameters: The global bias μ is just the average of all test data. The max iteration is 100, since larger iterations do not improve the performance significantly but definitely take more than 3 minutes to run. The learning rate $\lambda = 0.01$ which is the best after experimenting $\lambda \in \{0.01, 0.1, 0.2, 0.5\}$. The regularization parameter are set differently, namely, $u_{reg} = v_{reg} = 0.1$ and $b_{reg} = 0.01$, Since bias terms are scalars rather than vectors, they don't require as much regularization as U and V.