Homework 1

CSCI 5525: Machine Learning

Due on September 26th 11am (before class)

Please type in your info:

- **Name**: Chih-Tien Kuo
- **Student ID**: 5488927
- **Email**: kuo00013@umn.edu
- **Collaborators, and on which problems:**

Homework Policy. (1) You are encouraged to collaborate with your classmates on homework problems, but each person must write up the final solutions individually. You need to fill in above to specify which problems were a collaborative effort and with whom. (2) Regarding online resources, you should **not**:

- Google around for solutions to homework problems,
- Ask for help on online.
- Look up things/post on sites like Quora, StackExchange, etc.

Submission. Submit a PDF using this LaTeX template for written assignment part and submit Python jupyter or Colab python notebooks (.ipynb) for all programming part. You should upload all the files on Canvas.

Written Assignment

Instruction. For each problem, you are required to write down a full mathematical proof to establish the claim.

Problem 1. Two helpful matrices.

Let us first recall the notations in linear regression. The design matrix and the response vector are are defined as:

$$A = \begin{bmatrix} \leftarrow x_1^\intercal \rightarrow \\ \vdots \\ \leftarrow x_n^\intercal \rightarrow \end{bmatrix} \qquad \mathbf{b} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

For this problem, we will assume the covariance matrix $A^\intercal A$ is invertible, and so $(A^\intercal A)^{-1}$ is well-defined (**Clearly mention the properties of matrix operations used while solving**).

Problem 1.1. The residual matrix. For any weight vector $\mathbf{w}$, let us define the vector of least squares residuals as

$$e = \mathbf{b} - A\mathbf{w}$$

Now if $\mathbf{w}$ is the least square solution given by $\mathbf{w} = (A^\intercal A)^{-1}A^\intercal\mathbf{b}$, we can rewrite e as

$$e = \mathbf{b} - A(A^\intercal A)^{-1}A^\intercal\mathbf{b} = \left(I - A(A^\intercal A)^{-1}A^\intercal\right)\mathbf{b}$$

Now let $M = \left(I - A(A^\intercal A)^{-1}A^\intercal\right)$. Show that

- M is symmetric (i.e. $M = M^\intercal$). (**2 points**)
- M is idempotent (i.e. $M^2 = M$). (**2 points**)
- $MA = 0$. (**1 point**)

Answer

(a) From the property of $(A + B)^\intercal = A^\intercal + B^\intercal$ and $(AB)^\intercal = B^\intercal A^\intercal$:

$$\begin{aligned} M^\intercal &= \left(I - A(A^\intercal A)^{-1}A^\intercal\right)^\intercal \\ &= I^\intercal - \left(A(A^\intercal A)^{-1}A^\intercal\right)^\intercal \\ &= I - A\left(A(A^\intercal A)^{-1}\right)^\intercal \\ &= I - A\left(A^\intercal A\right)^{-\intercal}A^\intercal \\ &= I - A\left(A^\intercal A\right)^{-1}A^\intercal \\ &= M \end{aligned}$$

(b)

$$\begin{aligned} M^2 &= \left(I - A(A^\intercal A)^{-1}A^\intercal\right)^2 \\ &= I - A(A^\intercal A)^{-1}A^\intercal - A(A^\intercal A)^{-1}A^\intercal + \left(A(A^\intercal A)^{-1}A^\intercal\right)^2 \\ &= I - A(A^\intercal A)^{-1}A^\intercal - A(A^\intercal A)^{-1}A^\intercal + A(A^\intercal A)^{-1}A^\intercal \\ &= I - A(A^\intercal A)^{-1}A^\intercal \\ &= M \end{aligned}$$

(c)

$$MA = \left(I - A(A^\intercal A)^{-1}A^\intercal\right)A = A - A(A^\intercal A)^{-1}A^\intercal A = A - A = 0$$

Problem 1.2. The hat matrix. Using the residual maker, we can derive another matrix, the hat matrix or projection matrix $P = I - M = A(A^\intercal A)^{-1}A^\intercal$. Note that the predicted value by the least squares solution is given by $P\mathbf{b}$. Show that

- P is symmetric. (**1 point**)
- P is idempotent. (**1 point**)

Answer

(a)

$$P^\intercal = (A(A^\intercal A)^{-1}A^\intercal)^\intercal = A\,(A(A^\intercal A)^{-1})^\intercal = A\,(A^\intercal A)^{-\intercal}\,A^\intercal = A\,(A^\intercal A)^{-1}\,A^\intercal = P$$

(b)

$$P^2 = (A(A^\intercal A)^{-1}A^\intercal)^2 = A(A^\intercal A)^{-1}(A^\intercal A)(A^\intercal A)^{-1}A^\intercal = A(A^\intercal A)^{-1}A^\intercal = P$$

Problem 2. Gradient of conditional log-likelihood.

For any $a \in \mathbb{R}$, let $\sigma(a) = \frac{1}{1+\exp(-a)}$. For each example $(x_i, y_i) \in \mathbb{R}^d \times \{0, 1\}$, the conditional log-likelihood of logistic regression is

$$\ell(y_i \mid x_i, \mathbf{w}) = y_i \ln(\sigma(\mathbf{w}^\intercal x_i)) + (1 - y_i) \ln(\sigma(-\mathbf{w}^\intercal x_i))$$

Derive the gradient of $\ell(y_i \mid x_i, \mathbf{w})$ with respect to w_j (i.e. the j-th coordinate of $\mathbf{w}$) by following the following steps (**Clearly mention the properties of derivatives used while solving**).

- Derive $\frac{\partial}{\partial a}\sigma(a)$. (**2 points**)
- Derive $\frac{\partial}{\partial w_j}\sigma(\mathbf{w}^\intercal x_i)$. (**1 point**)
- Derive $\frac{\partial}{\partial w_j}\ln\sigma(\mathbf{w}^\intercal x_i)$. (**2 points**)
- Derive $\frac{\partial}{\partial w_j}\ln\sigma(-\mathbf{w}^\intercal x_i)$. (**1 point**)
- Derive $\frac{\partial}{\partial w_j}\ell(y_i \mid x_i, \mathbf{w})$. (**2 points**)

Answer For all of the five problems below, chain rule is used.

(a)

$$\begin{aligned}\frac{\partial}{\partial a}\sigma(a) &= \frac{\partial}{\partial a}(1+\exp(-a))^{-1} \\ &= -(1+\exp(-a))^{-2}(-\exp(-a)) \\ &= \frac{\exp(-a)}{(1+\exp(-a))^2}\end{aligned}$$

(b)

$$\begin{aligned}\frac{\partial}{\partial w_j}\sigma(\mathbf{w}^\intercal x_i) &= \frac{\partial(\sigma(\mathbf{w}^\intercal x_i))}{\partial(\mathbf{w}^\intercal x_i)}\frac{\partial(\mathbf{w}^\intercal x_i)}{\partial w_j} \\ &= \underbrace{\frac{\exp(-\mathbf{w}^\intercal x_i)}{(1+\exp(-\mathbf{w}^\intercal x_i))^2}}_{\text{From part (a)}} x_{i,j}\end{aligned}$$

(c)

$$\begin{aligned}
\frac{\partial}{\partial w_j} \ln \sigma(\mathbf{w}^\intercal x_i) &= \frac{\partial(\ln \sigma(\mathbf{w}^\intercal x_i))}{\partial(\sigma(\mathbf{w}^\intercal x_i))} \frac{\partial \sigma(\mathbf{w}^\intercal x_i)}{\partial w_j} \\
&= \frac{1}{\sigma(\mathbf{w}^\intercal x_i)} \underbrace{\frac{\exp(-\mathbf{w}^\intercal x_i)}{(1+\exp(-\mathbf{w}^\intercal x_i))^2} x_{i,j}}_{\text{From part (b)}} \\
&= (1+\exp(-\mathbf{w}^\intercal x_i)) \frac{\exp(-\mathbf{w}^\intercal x_i)}{(1+\exp(-\mathbf{w}^\intercal x_i))^2} x_{i,j} \\
&= \frac{\exp(-\mathbf{w}^\intercal x_i)}{1+\exp(-\mathbf{w}^\intercal x_i)} x_{i,j} \\
&= \sigma(-\mathbf{w}^\intercal x_i) x_{i,j}
\end{aligned}$$

(d)

$$\begin{aligned}
\frac{\partial}{\partial w_j} \ln \sigma(-\mathbf{w}^\intercal x_i) &= \frac{\partial(\ln \sigma(-\mathbf{w}^\intercal x_i))}{\partial(\sigma(-\mathbf{w}^\intercal x_i))} \frac{\partial \sigma(-\mathbf{w}^\intercal x_i)}{\partial w_j} \\
&= \frac{1}{\sigma(-\mathbf{w}^\intercal x_i)} \underbrace{\frac{\exp(\mathbf{w}^\intercal x_i)}{(1+\exp(\mathbf{w}^\intercal x_i))^2} (-x_{i,j})}_{\text{From part (b)}} \\
&= (1+\exp(\mathbf{w}^\intercal x_i)) \frac{\exp(\mathbf{w}^\intercal x_i)}{(1+\exp(\mathbf{w}^\intercal x_i))^2} (-x_{i,j}) \\
&= -\frac{\exp(\mathbf{w}^\intercal x_i)}{1+\exp(\mathbf{w}^\intercal x_i)} x_{i,j} \\
&= -\sigma(\mathbf{w}^\intercal x_i) x_{i,j}
\end{aligned}$$

(e)

$$\begin{aligned}
\frac{\partial}{\partial w_j} \ell(y_i \mid x_i, \mathbf{w}) &= \frac{\partial[y_i \ln(\sigma(\mathbf{w}^\intercal x_i)) + (1-y_i)\ln(\sigma(-\mathbf{w}^\intercal x_i))]}{\partial w_j} \\
&= y_i \frac{\partial(\ln(\sigma(\mathbf{w}^\intercal x_i))}{\partial w_j} + (1-y_i) \frac{\partial(\ln(\sigma(-\mathbf{w}^\intercal x_i)))}{\partial w_j} \\
&= y_i \sigma(-\mathbf{w}^\intercal x_i) x_{i,j} - (1-y_i)\sigma(\mathbf{w}^\intercal x_i) x_{i,j}
\end{aligned}$$

So the gradient of $\ell(y_i \mid x_i, \mathbf{w})$ is $y_i \sigma(-\mathbf{w}^\intercal x_i) x_i - (1-y_i)\sigma(\mathbf{w}^\intercal x_i) x_i$

□

Problem 3. Derivation of Ridge Regression Solution.

Recall that in class we claim that the solution to ridge regression ERM:

$$\min_{\mathbf{w}} \ (\|A\mathbf{w} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{w}\|_2^2)$$

is $\mathbf{w}^* = (A^\intercal A + \lambda I)^{-1} A^\intercal \mathbf{b}$. Now provide a proof. (**5 points**)
(Hint: recall that $\nabla F(\mathbf{w}) = \mathbf{0}$ is a sufficient condition for $\mathbf{w}$ to be a minimizer of any convex function F.) (**Clearly mention the properties of matrix calculus used while solving**)

Answer

Proof. $\min_{\mathbf{w}}\ (\|A\mathbf{w}-\mathbf{b}\|_2^2+\lambda\|\mathbf{w}\|_2^2)$ is equivalent to solving $\nabla(\|A\mathbf{w}-\mathbf{b}\|_2^2+\lambda\|\mathbf{w}\|_2^2)=0$.

$$\begin{aligned}\nabla(\|A\mathbf{w}-\mathbf{b}\|_2^2+\lambda\|\mathbf{w}\|_2^2) &= \nabla(A\mathbf{w}-\mathbf{b})^\intercal(A\mathbf{w}-\mathbf{b})+\lambda\nabla\mathbf{w}^\intercal\mathbf{w}\\ &= \nabla(\mathbf{w}^\intercal\underbrace{A^\intercal A}_{X}\mathbf{w}-2\mathbf{w}^\intercal A^\intercal\mathbf{b}+\mathbf{b}^\intercal\mathbf{b})+\lambda\nabla\mathbf{w}^\intercal\mathbf{w}\\ &= (2A^\intercal A\mathbf{w}-2A^\intercal\mathbf{b})+\lambda(2\mathbf{w})\\ &\text{(The gradient of inner products } \nabla\mathbf{w}X^\intercal\mathbf{w}=(X+X^\intercal)\mathbf{w}\\ &\text{and if X is symmetric, the gradient becomes } 2X\mathbf{w})\\ &= 0\end{aligned}$$

We then can solve $\mathbf{w}^*$ from $2A^\intercal A\mathbf{w}-2A^\intercal\mathbf{b}+2\lambda\mathbf{w}=0$

$$\begin{aligned}2A^\intercal A\mathbf{w}-2A^\intercal\mathbf{b}+2\lambda\mathbf{w}=0 &\Rightarrow (A^\intercal A+\lambda I)\mathbf{w}=A^\intercal\mathbf{b}\\ &\Rightarrow \mathbf{w}^*=(A^\intercal A+\lambda I)^{-1}A^\intercal\mathbf{b}\end{aligned}$$

□

Programming Assignment

Instruction. For each problem, you are required to submit Python notebook (.ipynb). The python notebook should be self-sufficient and explanatory in terms of the code, comments & equations you will implement in a cell and the required plots. The submitted python notebook must show all the steps and **DO NOT INCLUDE MORE THAN ONE FUNCTIONALITY IN ONE CELL**.

- **Python** version: Python 3.
- Please follow PEP 8 style of writing your Python code for better readability in case you are wondering how to name functions & variables, put comments and indent your code
- **Environment**: For this homework, you can use this environment `https://colab.research.google.com/notebooks/welcome.ipynb` If you want to work on your local machine, we recommend you to use Anaconda 3 and notebooks.
- **Packages allowed**: numpy, pandas, matplotlib
- **Submission**: For programming parts, **ONLY THE PYTHON 3 NOTEBOOKS WILL BE ACCEPTED**

Problem 1. Ridge Regression.

For this problem, you will use Housing dataset (Housing.csv). The dataset has 505 points, 13 features, and 1 target variable (label). First row has headers naming the features. First 13 columns are features and the last column named Price is the target. Submit a python notebook (name HW1-Ridge.ipynb).

a) (**12 Points**) Your goal is to implement ridge regression. You can design or structure your code to fulfil the requirements but it should have at least these methods, X being features and y target.:

Cross validation: Make sure you randomly shuffle the dataset and partition it into almost equal (k=5) folds. Save each of the 5 folds into dictionary X_shuffled and y_shuffled.

X_train, y_train, X_valid, y_valid = get_next_train_valid(X_shuffled, y_shuffled, itr) where itr is iteration number.

model_weights, model_intercept = train(X_train, y_train, lambda)

y_predict = predict(X_valid, model_weights, model_intercept)

We should be able to call any of these methods. **DO NOT USE GRADIENT DESCENT** to solve it in this assignment.

b) (**3 Points**) Briefy describe the approach in one paragraph (no more than 5 lines) along with any equations used in your implementation.

c) (**5 Points**) Report mean RMSE on train and validation sets using 5-fold cross-validation for different lambda varying from [0 to 100]. Include a plot showing lambda (x-axis) vs mean RMSE on y-axis.

Problem 2. Logistic regression.

For this problem, you will use the IRIS dataset. Features are in the file IRISFeat.csv and labels in the file IRISlabel.csv. The dataset has 150 samples. A python notebook (name HW1-Logistic.ipynb) with all the steps need to be submitted.

a) (**12 Points**) Your goal is to implement logistic regression. You can design or structure your code to fulfil the requirements but it should have at least these methods, X being features and y target.

Cross validation: Similar to problem 1.
X_train, y_train, X_valid, y_valid = get_next_train_valid(X_shuffled, y_shuffled, itr) where itr is iteration number.

θ = train(X_train, y_train)

y_predict_class = predict(X_valid, θ)

USE GRADIENT DESCENT to solve it. You should initialize the weights randomly to begin with.

b) (**3 Points**) At the beginning, briefly describe the approach in one paragraph (no more than 5 lines) along with any equations and methods used in your implementation.

c) (**5 Points**) Report the training and validation set error rates (number of misclassified samples/total number of samples) from 5-fold cross validation. Explain your selection of learning rate and How does it affect the performance/training of your model?