Saarland University, Department of Computer Science
Neural Network Assignment 2

Deborah Dormah Kanubala (7025906) , Irem Begüm Gündüz (7026821), Anh Tuan Tran (7015463)

November 29, 2022

Exercise 3.1

3.1.a

A Matrix $\mathcal{M}$ is said to be a symmetric matrix if $\mathcal{M}^\top = \mathcal{M}$

$$\mathcal{M} = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}$$

$$\mathcal{M}^\top = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}$$

Here, $\mathcal{M}^\top = \mathcal{M}$ hence it is a symmetric matrix. Having a symmetric matrix would mean the eigenvalues are real and its eigenvectors are perpendicular. Therefore, this makes it **orthogonally diagonalizable** meaning there is an orthogonal matrix $\mathcal{U}$ and a diagonal matrix $\mathcal{D}$ such that $\mathcal{M} = \mathcal{U}\mathcal{D}\mathcal{U}^{-1}$

3.1.b

No. Since $\mathcal{M}^\top$ is a symmetrical matrix, its inverse would exist and hence it is not a singular matrix.

Exercise 3.2

3.2.a

✓ **Accuracy:** Accuracy measures the correct number of predictions (both true positives and true negatives) all over the total number of predictions made.

$$\frac{TruePositive + TrueNegatives}{TruePositive + TrueNegatives + FalsePositives + FalseNegatives}$$

- ✓ **Precision:** Precision measures the rate of true positive predictions to the number of all positive predictions (includes false positives)

$$\frac{TruePositive}{TruePositive + FalsePositives}$$

- ✓ **Recall:** Recall measures the rate of true positive predictions to the number of all predictions that should actually have been positive.

$$\frac{TruePositive}{TruePositive + FalseNegatives}$$

- ✓ **F1:** F1 is the harmonic mean of the recall and precision.

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

- ✓ **Example:** In the case of an imbalanced dataset, it would be better to use $F1score$ over the other. Also, if the interest involves maximizing the positive predictions then it would be better to use recall and precision often involves maximizing the negative predictions.

3.2.b

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 * X_{i1} + \beta_2 * lnX_{i2} + \beta_3 * X_{i3} \\ Y_i &= 10 + 10 * X_{i1} + 0.5 * lnX_{i2} - 5 * X_{i3} \end{aligned} \tag{1}$$

- a) let X_{i1} increases by one unit:

$$\begin{aligned} Y_i^{'} &= 10 + 10 * (X_{i1} + 1) + 0.5 * lnX_{i2} - 5 * X_{i3} \\ &= Y_i + 10 \end{aligned} \tag{2}$$

Considering

$$\begin{aligned} Y_i + 10 &= Y_i + \frac{1000}{100} * Y_i \\ &= Y_i + 10 * Y_i \end{aligned} \tag{3}$$

The statement is True when $Y_i \approx 1$

let X_{i1} decreases by one unit:

$$\begin{aligned} Y_i^{''} &= 10 + 10 * (X_{i1} - 1) + 0.5 * lnX_{i2} - 5 * X_{i3} \\ &= Y_i - 10 \end{aligned} \tag{4}$$

Considering

$$\begin{aligned} Y_i^{''} = Y_i - 10 &= Y_i - \frac{1000}{100} * Y_i \\ &= Y_i - 10 * Y_i \end{aligned} \tag{5}$$

$Y_i > 0$ since it represents company values, $Y_i^{''} = Y_i - 10 * Y_i 0 < 0$ and this cannot happen. The statement is always False in this case.

In conclusion, the statement is True only when $Y_i \approx 1$ and $X_i 1$ increase by 1.

- b) Let X_{i2} increase by one unit:

$$\begin{aligned} Y_i^{'} &= 10 + 10 * X_{i1} + 0.5 * ln(X_{i2} + 1) - 5 * X_{i3} \\ &= 10 + 10 * X_{i1} + 0.5 * ln(X_{i2} * (1 + \frac{1}{X_i 2})) - 5 * X_{i3} \\ &= Y_i + 0.5 * ln(1 + 1/X_{i2}) \end{aligned} \tag{6}$$

  Considering

$$\begin{aligned} &Y_i^{'} = Y_i + 0.5 * ln(1 + 1/X_{i2}) = Y_i + 0.5Y_i \\ &\Longrightarrow ln(1 + 1/X_{i2}) = Y_i \\ &\Longrightarrow X_{i2} = \frac{1}{e^{Y_i} - 1} \end{aligned} \tag{7}$$

  Thus the statement is True when $X_{i2} = \frac{1}{e^{Y_i - 1}}$ and X_{i2} increases by 1.

  Let X_{i2} decreases by one unit

$$\begin{aligned} Y_i^{''} &= 10 + 10 * X_{i1} + 0.5 * ln(X_{i2} - 1) - 5 * X_{i3} \\ &= 10 + 10 * X_{i1} + 0.5 * ln(X_{i2} * (1 - \frac{1}{X_i 2})) - 5 * X_{i3} \\ &= Y_i + 0.5 * ln(1 - 1/X_{i2}) \end{aligned} \tag{8}$$

  Considering

$$\begin{aligned} &Y_i^{'} = Y_i + 0.5 * ln(1 - 1/X_{i2}) = Y_i - 0.5Y_i \\ &\Longrightarrow ln(1 - 1/X_{i2}) = Y_i \\ &\Longrightarrow X_{i2} = \frac{1}{1 - e^{Y_i}} \end{aligned} \tag{9}$$

  Since $Y_i > 0$ (as it represents company values), $1 - e^{Y_i} < 0$ and thus $X_{i2} < 0$ which is not possible (as X_{i2} represents ”nitial stock value”). Thus the statement is False in this case.

  In conclusion, the statement is True only when $X_{i2} = \frac{1}{e^{Y_i - 1}}$ and X_{i2} increases by 1 othewise, it is False.

- c)

  Let X_{i2} increase 100% its values

$$\begin{aligned} Y_i^{'} &= 10 + 10 * X_{i1} + 0.5 * ln(X_{i2} * 2) - 5 * X_{i3} \\ &= Y_i + 0.5 * ln(2) \end{aligned} \tag{10}$$

  Considering

$$\begin{aligned} &Y_i^{'} = Y_i + Y_i + 0.5 * ln(2) = Y_i + 0.5Y_i \\ &\Longrightarrow ln(2) = Y_i \end{aligned} \tag{11}$$

  Therefore, the statement is true when we have X_i and Y_i such that $Y_i = ln(2)$.

  Let X_{i2} decrease 100% its values

$$Y_i^{'} = 10 + 10 * X_{i1} + 0.5 * ln(0) - 5 * X_{i3} \tag{12}$$

  This is not possible as $ln(0)$ does not exist.

  In conclusion, the statement is True only when we have X_i and Y_i such that $Y_i = ln(2)$.

- d) Bias term can be understand as the default value for the company value when all $X_{ij} \approx 0$.

3.2.c

We have the MSE:

$$MSE(a) = \frac{1}{m}\|Y - Xa\|_2^2 \tag{13}$$

Given $w, x, c \in \mathbb{R}^n$, $A, B\ in\, \mathbb{R}^{n\times n}$, we have the following identity from the assignment 2:

$$\begin{aligned} \nabla_x(x^T Ax) &= Ax + A^T x \\ \nabla_x(\|Bx - c\|_2^2) &= 2B^T(Bx - c) \end{aligned} \tag{14}$$

To find the least square solution, we find $\nabla_a MSE(a) = 0$

$$\begin{aligned} \nabla_a MSE(a) = &\qquad 2X^T(Xa - Y) = 0 \\ &\implies X^T Xa - X^T Y = 0 \\ &\implies a = (X^T X)^{-1} X^T Y \end{aligned} \tag{15}$$

Since $\mathbb{E}[X] = E[Y] = 0$, we have

$$\begin{aligned} a =&((X - \mathbb{E}[X])^T(X - \mathbb{E}[x]))^{-1}(X - E[X])^T(Y - E[Y]) \\ \frac{1}{n^2}a =&((X - \mathbb{E}[X])^T\frac{1}{n}I(X - \mathbb{E}[x]))^{-1}(X - E[X])^T\frac{1}{n}I(Y - E[Y]) \\ \frac{1}{n^2}a =&(\sigma_x)^{-1}\sigma_{xy} \\ a =&n^2(\sigma_x)^{-1}\sigma_{xy} \end{aligned} \tag{16}$$

3.2.d

We assume that the added noise and the data are independent. The proof is as follows:

$$\begin{aligned} \mathbb{E}[(y_i - \langle w, x_i + \epsilon_i\rangle)^2] =& \mathbb{E}[(y_i - \langle w, x_i\rangle - \langle w, \epsilon_i\rangle)^2] \\ =& \mathbb{E}[(y_i - \langle w, x_i\rangle)^2 - 2(y_i - \langle w, x_i\rangle)(\langle w, \epsilon_i\rangle) + \langle w, \epsilon_i\rangle^2] \\ =& \mathbb{E}[(y_i - \langle w, x_i\rangle)^2] - 2\,\mathbb{E}[(y_i - \langle w, x_i\rangle)(\langle w, \epsilon_i\rangle)] + \\ & \mathbb{E}[\langle w, \epsilon_i\rangle^2] \end{aligned} \tag{17}$$

Since w is a constant and ϵ_i is independent of y_i and x_i, we have

$$\begin{aligned} &2\,\mathbb{E}[(y_i - \langle w, x_i\rangle)(\langle w, \epsilon_i\rangle)] \\ &= 2\,\mathbb{E}[(y_i - \langle w, x_i\rangle)]\,\mathbb{E}[\langle w, \epsilon_i\rangle] \\ &= 2\,\mathbb{E}[(y_i - \langle w, x_i\rangle)](\langle w, \mathbb{E}[\epsilon_i]\rangle) \\ &= 2\,\mathbb{E}[(y_i - \langle w, x_i\rangle)]\langle w, 0\rangle \\ &= 0 \end{aligned} \tag{18}$$

Considering $\mathbb{E}[\langle w, \epsilon_i\rangle^2]$

$$
\begin{aligned}
&\mathbb{E}[\langle w, \epsilon_i \rangle^2] \\
&= \mathbb{E}[(w^T \epsilon_i)^2] \\
&= \mathbb{E}[(w^T \epsilon_i)(\epsilon_i^T w)] \\
&= \mathbb{E}[w^T (\epsilon_i \epsilon_i^T) w] \\
&= w^T \, \mathbb{E}[(\epsilon_i \epsilon_i^T)] w \\
&= w^T (\delta^2 I) w \\
&= \delta^2 * (w^T w) \\
&= \delta^2 \sum_{i=1}^{d} w_i^2
\end{aligned} \tag{19}
$$

Using results from Eq.18 and Eq.19 for Eq.20 we have:

$$\mathbb{E}[(y_i - \langle w, x_i + \epsilon_i \rangle)^2] = \mathbb{E}[(y_i - \langle w, x_i \rangle)^2] \tag{20}$$

Exercise 3.3

3.3.a

- **Bias:** Bias measures the difference between the true value and predictions. A model with a high bias (low model capacity) will fail to capture the true underlying distributions of the data.

- **Variance:** Variance measures the variations of the predictions across different datasets. A model with high variance (high model capacity) will fit perfectly to the training data but will turn to underperform when used on other unseen datasets.

- **Relation to Under-fitting and Over-fitting:** A model that exhibits a very high bias and low variance will turn to underfit. On the other hand, a model that exhibits a very high variance and low bias will turn to overfit.

3.3.b

Let X be the training data of $\mathbb{R}^{m \times d}$, Y be the labels of $\mathbb{R}^m$ and $w \in \mathbb{R}^d$ be the parameters. Ridge regression loss is as follows:

$$
\begin{aligned}
J(w) =& MSE_{train} + \lambda w^T w \\
=& \frac{1}{m} \|Y - Xw\|_2^2 + \lambda w^T w \\
=& \frac{1}{m} (Y - Xw)^T (Y - Xw) + \lambda w^T w
\end{aligned} \tag{21}
$$

Given $w, x, c \in \mathbb{R}^n$, $A, B \; in \, \mathbb{R}^{n \times n}$, we have the following identity from the assignment 2:

$$
\begin{aligned}
\nabla_x (x^T A x) &= Ax + A^T x \\
\nabla_x (\|Bx - c\|_2^2) &= 2B^T (Bx - c)
\end{aligned} \tag{22}
$$

Taking gradient w.r.t to w:

$$
\begin{aligned}
\nabla_w J(w) =& \nabla_w(\frac{1}{m}\|Y - Xw\|_2^2) + \nabla_w(\lambda w^T w) \\
=& \frac{1}{m}\nabla_w(\|Y - Xw\|_2^2) + \nabla_w(\lambda w^T I w) \\
=& \frac{1}{m}(2X^T(Xw - Y)) + Iw + I^T w \\
=& \frac{1}{m}(2X^T(Xw - Y)) + 2\lambda w
\end{aligned} \tag{23}
$$

Solve for $\nabla_w J(w) = 0$:

$$
\begin{aligned}
&\frac{1}{m}(2X^T(Xw - Y)) + 2\lambda w = 0 \\
\implies & m\lambda w = X^T Y - X^T X w \\
\implies & m\lambda w + X^T X w = X^T Y \\
\implies & (m\lambda I + X^T X)w = X^T Y \\
\implies & w = (m\lambda I + X^T X)^{-1} X^T Y
\end{aligned} \tag{24}
$$

Exercise 3.4

3.4.a

Cross-validation shows the model performance over multiple train-test splits while hold out only use a single train-test split. Cross-validation thus shows how well the model generalize better than hold out. Cross-validation is needed when we want to fine tune parameters when only a small amount of data is available. Cross-validation is needed when we want to know how well the model generalize.

3.4.b

Using k-fold cross-validation, we split the data into 5 folds and fixed these 5 folds for all the three models of polynomials 1, 5 and 9. For each model, we compute the 5-fold cross-validation MSE: we train a model from the scratch 5 times (each times with a different fold as the test set) and compute the MSE of test set each time; average of the 5 recorded MSE will be the MSE for the model. Finally, we compare the final MSE above among the three models to find the best setting.

3.4.c

The leave-one-out cross validation MSE after shuffle the data would be the same as before (35) because the shuffle does not change the 1-sample test sets.