Exercise

04

TUM Department of Informatics

Supervised by Prof. Dr. Stephan Günnemann

Informatics 3 - Professorship of Data Mining and Analytics

Submitted by Marcel Bruckner (03674122)

Julian Hohenadel (03673879)

Kevin Bein (03707775)

Submission date Munich, November 6, 2019

Least squares regression

Problem 1:

Using $T = diag(t_i)$, the derivative can be calculated similar to the lecture:

$$\begin{aligned}
\nabla_w E_{\text{weighted}}(w) &= \nabla_w \frac{1}{2} \sum_{i=1}^{N} t_i \left[w^T \phi(x_i) - y_i \right]^2 \\
&= \nabla_w \frac{1}{2} (\Phi w - y)^T T (\Phi w - y) \\
&= \nabla_w \frac{1}{2} (w^T \Phi^T - y^T)(T \Phi w - T y) \\
&= \nabla_w \frac{1}{2} (w^T \Phi^T T \Phi w - 2 w^T \Phi^T T y + y^T T y) \\
&= \nabla_w (\frac{1}{2} w^T \Phi^T T \Phi w - w^T \Phi^T T y + \frac{1}{2} y^T T y) \\
&= \Phi^T T \Phi w - \Phi^T T y \\
&\overset{!}{=} 0 \\
&\Rightarrow \Phi^T T y = \Phi^T T \Phi w \\
&\Rightarrow w = \Phi^T T y (\Phi^T T \Phi)^{-1}
\end{aligned}$$

Missing explanation for 1) and 2)

Ridge regression

Problem 2:

$$
\begin{aligned}
E_{\text{LS}} &= \frac{1}{2}\sum_{i=1}^{N}\left[w^T\phi(x_i) - y_i\right]^2 \\
&= \frac{1}{2}(\Phi w - y)^T(\Phi w - y) \\
E_{\text{ridge}} &= \frac{1}{2}\sum_{i=1}^{N}\left[w^T\phi(x_i) - y_i\right]^2 + \frac{\lambda}{2}||w||_2^2 \\
&= \frac{1}{2}(\Phi w - y)^T(\Phi w - y) + \frac{\lambda}{2}||w||_2^2
\end{aligned}
$$

Following the instructions, we can augment the design matrix Φ and y:

$$
\Phi = \begin{pmatrix} \phi_1(x_1) & \dots & \phi_M(x_1) \\ \vdots & \ddots & \vdots \\ \phi_1(x_N) & \dots & \phi_M(x_N) \end{pmatrix} \in \mathbb{R}^{N\times M}
$$

$$
\Rightarrow \Phi_A = \begin{pmatrix} \phi_1(x_1) & \dots & \phi_M(x_1) \\ \vdots & \ddots & \vdots \\ \phi_1(x_N) & \dots & \phi_M(x_N) \\ \sqrt{\lambda}I & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda}I \end{pmatrix} = \begin{pmatrix} \Phi \\ \sqrt{\lambda}I \end{pmatrix} \in \mathbb{R}^{N\times M+M}
$$

$$
y = \begin{pmatrix} y_1 \\ \vdots \\ y_M \end{pmatrix} \in \mathbb{R}^{M} \Rightarrow y_A = \begin{pmatrix} y_1 \\ \vdots \\ y_M \\ 0_1 \\ \vdots \\ 0_M \end{pmatrix} = \begin{pmatrix} y \\ 0 \end{pmatrix} \in \mathbb{R}^{M+M}
$$

Inserting Φ_A and y_A into $E_{\text{LS}}(w)$ directly gives ridge regression:

$$
E_{\text{LS}}(w) = \frac{1}{2}(\Phi_A w - y_A)^T(\Phi_A w - y_A) = \frac{1}{2}(\Phi w - y)^T(\Phi w - y) + \frac{\lambda}{2}||w||_2^2 = E_{\text{ridge}}(w)
$$

Problem 3:

$$
\begin{aligned}
\nabla_w E_{\text{ridge}}(w) &= \nabla_w \left[\frac{1}{2}(\Phi w - y)^T(\Phi w - y) + \frac{\lambda}{2}||w||_2^2\right] \\
&= \nabla_w \frac{1}{2}(\Phi w - y)^T(\Phi w - y) + \nabla_w \frac{\lambda}{2}||w||_2^2 \\
&= (\Phi^T\Phi w - \Phi^T y) + \lambda w \\
&\overset{!}{=} 0 \\
&\Rightarrow \Phi^T\Phi w - \Phi^T y + \lambda w = 0 \\
&\Rightarrow \Phi^T\Phi w + \lambda w = \Phi^T y \\
&\Rightarrow (\Phi^T\Phi + \lambda I)w = \Phi^T y \\
&\Rightarrow w = (\Phi^T\Phi + \lambda I)^{-1}\Phi^T y
\end{aligned}
$$

When $N < M$ then the covariance Matrix $\Phi^T\Phi$ becomes singular $(det(\Phi^T\Phi) = 0)$ and thus cannot be inverted anymore. Adding the L2-Regularization term in ridge regression $\Phi^T\Phi + \lambda I$ fixes this problem and the inverse can be computed again. L2 also increases the numerical stability by reducing the variance of the matrix and thus roundoff errors cannot accumulate as fast.

Multi-output linear regression

Problem 4:

The solution is very similar to the single target case which is given in the lecture:

$$w_{\text{ML}} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y} = \Phi^\dagger \mathbf{y}$$

For multiple outputs we can combine the weights $\mathbf{w}_i$ and the outputs $\mathbf{y}_i$ into matrices with n rows containing the respective m^{th} vector.

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \dots \\ y_m \end{pmatrix} \in \mathbb{R}^{n \times m}$$

$$\mathbf{W} = \begin{pmatrix} w_1 \\ \dots \\ w_m \end{pmatrix} \in \mathbb{R}^{n \times m}$$

The log-likelihood then looks similar to the single output case:

$$\begin{aligned} -\ln p(\mathbf{Y}|\mathbf{X}, \mathbf{W}, \beta) &= -\ln \left[\prod_{i=1}^{n} \mathcal{N}(\mathbf{y}_i | \mathbf{W}^T \phi(\mathbf{x}_i), \beta^{-1} I) \right] \\ &= \frac{\beta}{2} \sum_{i=1}^{n} (\mathbf{W}^T \phi(\mathbf{x}_i) - \mathbf{Y})^2 - \frac{n}{2} \ln \beta + \frac{n}{2} \ln 2\pi \\ &= \frac{\beta}{2} \sum_{i=1}^{n} \left(\sum_{j=1}^{m} (\mathbf{w}_j^T \phi(\mathbf{x}_i) - \mathbf{y}_j)^2 \right) - \frac{n}{2} \ln \beta + \frac{n}{2} \ln 2\pi \end{aligned}$$

When calculating the minimzer, the constant parts fall out and we are left with the sum of the previous result of the different $\mathbf{y}_i$ which ultimately yields to the resullt for the multivariate case:

$$\begin{aligned} \mathbf{W}_{\text{ML}} &= \Phi^\dagger \mathbf{y}_1 + \ldots + \Phi^\dagger \mathbf{y}_m \\ &= \sum_{i=1}^{m} \Phi^\dagger \mathbf{y}_i \\ &= \Phi^\dagger \sum_{i=1}^{m} \mathbf{y}_i \\ &= \Phi^\dagger \mathbf{Y} \end{aligned}$$

Comparison of linear regression models

Problem 5:

a)

$$\mathbf{w}^{*T} x_i = \mathbf{w}_{new}^T a\mathbf{x}_i \Rightarrow \mathbf{w}_{new} = \frac{\mathbf{w}^*}{a}$$

b) Closed form solution of w^*_{new}:

$$\begin{aligned}
\mathbf{w}^*_{new} &= \arg\min_{\mathbf{w}} \frac{1}{2}\sum_{i=1}^{N}(\mathbf{w}^T a\mathbf{x}_i - \mathbf{y}_i)^2 + \frac{\lambda_{new}}{2}\mathbf{w}^T\mathbf{w} \\
&= \nabla_w \left[\frac{1}{2}\sum_{i=1}^{N}(\mathbf{w}^T a\mathbf{x}_i - \mathbf{y}_i)^2 + \frac{\lambda_{new}}{2}\mathbf{w}^T\mathbf{w}\right] \\
&= \nabla_w \left[\frac{1}{2}\sum_{i=1}^{N}(\mathbf{w}^T a\mathbf{x}_i - \mathbf{y}_i)^2\right] + \nabla_w \left[\frac{\lambda_{new}}{2}\mathbf{w}^T\mathbf{w}\right] \\
&= \nabla_w \left[\frac{1}{2}(a\mathbf{X}\mathbf{w} - \mathbf{y})^T(a\mathbf{X}\mathbf{w} - \mathbf{y})\right] + \nabla_w \left[\frac{\lambda_{new}}{2}\mathbf{w}^T\mathbf{w}\right] \\
&= \nabla_w \left[\frac{1}{2}(a\mathbf{w}^T\mathbf{X}^T a\mathbf{X}\mathbf{w} - 2\mathbf{w}^{\mathbf{T}} a\mathbf{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{y})\right] + \nabla_w \left[\frac{\lambda_{new}}{2}\mathbf{w}^T\mathbf{w}\right] \\
&= a\mathbf{X}^T a\mathbf{X}\mathbf{w} - a\mathbf{X}^T\mathbf{y} + \lambda_{new}\mathbf{y} \\
&\overset{!}{=} 0 \\
&\Rightarrow a\mathbf{X}^T a\mathbf{X}\mathbf{w} + \lambda_{new}\mathbf{y} = a\mathbf{X}^T\mathbf{y} \\
&\Rightarrow w^*_{new} = a(a^2\mathbf{X}^T\mathbf{X} + \lambda_{new}I)^{-1}\mathbf{X}^T\mathbf{y}
\end{aligned}$$

Condition from a):

$$\begin{aligned}
\mathbf{w}^*_{new} &\overset{!}{=} \frac{\mathbf{w}^*}{a} \\
\Leftrightarrow a(a^2\mathbf{X}^T\mathbf{X} + \lambda_{new}I)^{-1}\mathbf{X}^T\mathbf{y} &= \frac{1}{a}(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y} \\
\Leftrightarrow a(a^2\mathbf{X}^T\mathbf{X} + a^2\frac{\lambda_{new}}{a^2}I)^{-1}\mathbf{X}^T\mathbf{y} &= \frac{1}{a}(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y} \\
\Leftrightarrow \frac{1}{a^2}a(\mathbf{X}^T\mathbf{X} + \frac{\lambda_{new}}{a^2}I)^{-1}\mathbf{X}^T\mathbf{y} &= \frac{1}{a}(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y} \\
\Leftrightarrow \frac{1}{a}(\mathbf{X}^T\mathbf{X} + \frac{\lambda_{new}}{a^2}I)^{-1}\mathbf{X}^T\mathbf{y} &= \frac{1}{a}(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y} \\
\Leftrightarrow \frac{\lambda_{new}}{a^2} &= \lambda \\
\Leftrightarrow \lambda_{new} &= a^2\lambda
\end{aligned}$$

Programming Task

Problem 6:

exercise_04_notebook

November 2, 2019

1 Programming assignment 2: Linear regression

```
In [1]: import numpy as np

        from sklearn.datasets import load_boston
        from sklearn.model_selection import train_test_split
```

1.1 Your task

In this notebook code skeleton for performing linear regression is given. Your task is to complete the functions where required. You are only allowed to use built-in Python functions, as well as any `numpy` functions. No other libraries / imports are allowed.

1.2 Exporting the results to PDF

Once you complete the assignments, export the entire notebook as PDF and attach it to your homework solutions. The best way of doing that is 1. Run all the cells of the notebook. 2. Export/download the notebook as PDF (File -> Download as -> PDF via LaTeX (.pdf)). 3. Concatenate your solutions for other tasks with the output of Step 2. On a Linux machine you can simply use `pdfunite`, there are similar tools for other platforms too. You can only upload a single PDF file to Moodle.

Make sure you are using `nbconvert` Version 5.5 or later by running `jupyter nbconvert --version`. Older versions clip lines that exceed page width, which makes your code harder to grade.

1.3 Load and preprocess the data

I this assignment we will work with the Boston Housing Dataset. The data consists of 506 samples. Each sample represents a district in the city of Boston and has 13 features, such as crime rate or taxation level. The regression target is the median house price in the given district (in $1000's).

More details can be found here: http://lib.stat.cmu.edu/datasets/boston

```
In [2]: X , y = load_boston(return_X_y=True)

        # Add a vector of ones to the data matrix to absorb the bias term
        # (Recall slide #7 from the lecture)
        X = np.hstack([np.ones([X.shape[0], 1]), X])
        # From now on, D refers to the number of features in the AUGMENTED dataset
```

```
    #  (i.e. including the dummy '1' feature for the absorbed bias term)

    # Split into train and test
    test_size = 0.2
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=test_size)
```

1.4 Task 1: Fit standard linear regression

```
In [3]: def fit_least_squares(X, y):
            """Fit ordinary least squares model to the data.

            Parameters
            ----------
            X : array, shape [N, D]
                (Augmented) feature matrix.
            y : array, shape [N]
                Regression targets.

            Returns
            -------
            w : array, shape [D]
                Optimal regression coefficients (w[0] is the bias term).

            """

            w = np.linalg.pinv(X) @ y
            return w
```

1.5 Task 2: Fit ridge regression

```
In [4]: def fit_ridge(X, y, reg_strength):
            """Fit ridge regression model to the data.

            Parameters
            ----------
            X : array, shape [N, D]
                (Augmented) feature matrix.
            y : array, shape [N]
                Regression targets.
            reg_strength : float
                L2 regularization strength (denoted by lambda in the lecture)

            Returns
            -------
            w : array, shape [D]
                Optimal regression coefficients (w[0] is the bias term).

            """
```

```
    # https://stats.stackexchange.com/questions/69205/how-to-derive-the-ridge-regressi
    N = X.shape[0]
    #regulator = reg_strength * np.identity(N)
    #XTX = X @ X.T
    #inv = np.linalg.inv((X @ X.T) + regulator)
    #w = (inv.T @ X).T @ y
    w = (np.linalg.inv((X @ X.T) + (reg_strength * np.identity(N))).T @ X).T @ y

    return w
```

1.6 Task 3: Generate predictions for new data

```
In [5]: def predict_linear_model(X, w):
            """Generate predictions for the given samples.

            Parameters
            ----------
            X : array, shape [N, D]
                (Augmented) feature matrix.
            w : array, shape [D]
                Regression coefficients.

            Returns
            -------
            y_pred : array, shape [N]
                Predicted regression targets for the input data.

            """

            y_pred = w @ X.T
            return y_pred
```

1.7 Task 4: Mean squared error

```
In [6]: def mean_squared_error(y_true, y_pred):
            """Compute mean squared error between true and predicted regression targets.

            Reference: `https://en.wikipedia.org/wiki/Mean_squared_error`

            Parameters
            ----------
            y_true : array
                True regression targets.
            y_pred : array
                Predicted regression targets.

            Returns
            -------
```

```
    mse : float
        Mean squared error.

    """

    mse = (1/len(y_true)) * (np.sum((y_true - y_pred)**2))
    #from sklearn.metrics import mean_squared_error
    #return mean_squared_error(y_true, y_pred)
    return mse
```

1.8 Compare the two models

The reference implementation produces * MSE for Least squares $\approx$ **23.98** * MSE for Ridge regression $\approx$ **21.05**

You results might be slightly (i.e. ±1%) different from the reference soultion due to numerical reasons.

```
In [7]: # Load the data
        np.random.seed(1234)
        X , y = load_boston(return_X_y=True)
        X = np.hstack([np.ones([X.shape[0], 1]), X])
        test_size = 0.2
        X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=test_size)

        # Ordinary least squares regression
        w_ls = fit_least_squares(X_train, y_train)
        y_pred_ls = predict_linear_model(X_test, w_ls)
        mse_ls = mean_squared_error(y_test, y_pred_ls)
        print('MSE for Least squares = {0}'.format(mse_ls))

        # Ridge regression
        reg_strength = 1
        w_ridge = fit_ridge(X_train, y_train, reg_strength)
        y_pred_ridge = predict_linear_model(X_test, w_ridge)
        mse_ridge = mean_squared_error(y_test, y_pred_ridge)
        print('MSE for Ridge regression = {0}'.format(mse_ridge))
```

```
MSE for Least squares = 23.964571384956837
MSE for Ridge regression = 21.034931232092767
```

```
In [ ]:
```

Appendix

We confirm that the submitted solution is original work and was written by us without further assistance. Appropriate credit has been given where reference has been made to the work of others.

Munich, November 6, 2019, Signature Marcel Bruckner (03674122)

Munich, November 6, 2019, Signature Julian Hohenadel (03673879)

Munich, November 6, 2019, Signature Kevin Bein (03707775)