Machine Learning Exercise Sheet 04
Linear Regression

Anastasia Stamatouli : 03710902

10/11/19

• Least Squares Regression

Problem ①:

$$E_w(w) = \frac{1}{2}\sum_{i=1}^{N} t_i\,[w^T\varphi(x_i) - y_i]^2 =$$

$$\frac{1}{2}\sum_{i=1}^{N} t_i\,(w^T\varphi(x_i) - y_i)(w^T\varphi(x_i) - y_i) =$$

$$\frac{1}{2}(\phi w - y)^T T (\phi w - y) =$$

T: diagonal matrix of weights

$$= \frac{1}{2}(w^T\phi^T - y^T)\, T\, (\phi w - y) =$$

$$\frac{1}{2}\left[w^T\varphi^T T \phi w - w^T\phi^T T y - y^T T \phi w + y^T T y\right]$$

To find w^* that minimizes $E_w(w)$ we calculate $\nabla_w E_w(w)$.

$$\nabla_w E_w(w) = \frac{1}{2}\left[w^T(\phi^T T\phi + \phi^T T\phi) - (\phi^T T y)^T - y^T T\phi + 0\right]$$

$\Leftrightarrow$

$$\nabla_w E_w(w) = w^T\phi^T T\phi - y^T T\phi$$

In order to find the minimum we set $\nabla_w E_w(w) = 0$

Therefore: $\quad w^T\phi^T T\phi - y^T T\phi = 0 \Leftrightarrow$

$$w^T\phi^T T\phi = y^T T\phi \Leftrightarrow$$

$$(w^T\phi^T T\phi)^T = (y^T T\phi)^T \Leftrightarrow$$

$$\boxed{w = (\phi^T T\phi)^{-1}\phi^T T y}$$

1.1 It might happen that the data do not have the same variance (heteroskedasticity). In this case, ordinary least squares is no longer the optimal estimate as it gives constant noise variance to every observation (σ^2). So in the case of WLS we could set $w_i = \frac{1}{\sigma_i^2}$ and focus on the data with small variance.

1.2 When there are duplicates, regression is biased towards those points. In this case we could put more weight to the points which are not duplicated and less weight to duplicated ones.

- Ridge Regression

Problem 2 :

$\lambda > 0$ so it has a positive square root $\lambda = v^2$

- we define a new matrix augmented with rows corresponding v times the $m \times m$ identity matrix

$$\Phi_* = \begin{pmatrix} \Phi \\ vI \end{pmatrix}$$

- we define another augmented vector y whose additional terms are m zeros:

$$y_* = \begin{pmatrix} y \\ 0_{m \times 1} \end{pmatrix}$$

Now, we consider the objective function

$$E_{LS} = (\phi_* w - y_*)^T (\phi_* w - y_*)$$

Due to augmentation there will be m additional terms compared to what we would get with OLS.

$$(vw - 0)^2 = v^2 w^2 = \lambda w^T w$$

Therefore : $(\phi_* w - y_*)^T (\phi_* w - y_*) =$

$$(\phi w - y)^T (\phi w - y) + \lambda w^T w \quad (1)$$

In order to find minimizer w^*, we take the derivative of equation (1) and set it equal to zero

$$(1) \Rightarrow w^T \phi^T \phi - y^T \phi + \lambda w^T = 0 \Leftrightarrow$$
$$w = (\phi^T \phi + \lambda I)^{-1} \phi^T y$$

Problem 3 :

$$E_{ridge}(w) = \frac{1}{2} \sum_{i=1}^{N} (w^T \varphi(x_i) - y_i)^2 + \frac{\lambda}{2} w^T w$$

- We know that the derivative of $\left(\frac{1}{2} \sum_{i=1}^{N} (w^T \varphi(x_i) - y_i)^2\right)' =$
  $w^T \phi^T \phi - y^T \phi$

- the derivative of $\left(\frac{\lambda}{2} w^T w\right)' = \lambda w^T$

Therefore $\frac{d\,E_{ridge}(w)}{dw} = w^T\phi^T\phi - y^T\phi + \lambda w^T$

We set $\frac{d\,E_{rigde}(w)}{dw} = 0 \Leftrightarrow$

$$w^T\phi^T\phi - y^T\phi + \lambda w^T = 0 \Leftrightarrow$$

$$\boxed{w = (\phi^T\phi + \lambda I)^{-1}\phi^T y}$$

In case $N < M$ the covariance matrix $\phi^T\phi \in \mathbb{R}^{M \times M}$ will be singular, which means is not invertible.

when we use regularization, λI is added to the covariance matrix. Therefore, it fixes the degeneracy problem.

- Multi-output Linear Regression

Problem (4) :

Y is a vector which its rows are the targets $y_i \sim N(Wx_i, \Sigma)$, $W \in \mathbb{R}^{m \times n}$, $\Sigma \in \mathbb{R}^{m \times m}$.

We have n pairs (x_i, y_i) which are identically distributed. The likelihood is $\prod_i N(Wx_i, \Sigma)$. Therefore, negative likelihood is

$\frac{1}{2}\sum_i (y_i - Wx_i)^T \Sigma^{-1} (y_i - Wx_i) + \text{const}$ (slide 4?

In general, we solve m single least square problems for every $y_i \in Y$. Therefore $W_{MLE} = n\,(X^TX)^{-1}X^Ty$

Scanned with CamScanner

exercise_04_notebook

November 10, 2019

1 Programming assignment 2: Linear regression

[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

[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

[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).

    """
    # TODO

    w = np.linalg.solve(np.dot(np.transpose(X),X), np.dot(np.transpose(X),y))
    return w
```

1.5 Task 2: Fit ridge regression

[13]:
```
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).

    """
    # TODO

    w = np.linalg.solve((np.dot(np.transpose(X),X))+ reg_strength*np.identity(X.
↪shape[1]), np.dot(np.transpose(X),y))
    return w
```

1.6 Task 3: Generate predictions for new data

[14]:
```
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.

    """
    # TODO

    y = np.dot(X,w)
    return y
```

1.7 Task 4: Mean squared error

[26]:
```
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.

    """
    # TODO

    MSE = np.square(y_true-y_pred).mean()


    return MSE
```

1.8 Compare the two models

The reference implementation produces * MSE for Least squares ≈ **23.98** * MSE for Ridge regression ≈ **21.05**

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

[27]:
```
# 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 = 23.96457138495054
MSE for Ridge regression = 21.03493121591798
```

[ ]:
```
```