Department of Computer Engineering
University of Peradeniya

Data Mining and Machine Learning
Lab 04

July 18, 2017

1 Introduction to Linear Regression

Linear regression, is the simplest and most classic linear method for regression. Linear regression finds the parameters such that minimize the mean squared error between predictions and the true regression targets, on provided dataset. Linear regression has no parameters, which is a benefit, but it also has no way to control model complexity.

Linear regression in scalar form,

$$Y_i = \beta_0 + \beta_1 * X_i + \epsilon_i \tag{1}$$

where $\epsilon_i \sim \mathcal{N}(0,\, \sigma^2)$, $i \in \{1, \ldots, n\}$ and n is the number of observations. In matrix form, for n observations above expression (1) can be rewritten as bellow.

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & X_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \epsilon_n \end{pmatrix} \tag{2}$$

Thus, by using above result (2) this can be simplified into this form

$$Y = X\beta + \epsilon \tag{3}$$

where X is the design matrix, β is the vector of parameters, ϵ is the error vector and Y is the response vector. In order to determine the optimal solution for β, model's prediction errors/residuals (ϵ) is needed to be minimized. The way is calculating residuals are shown in Fig. 1. Ordinary least squares is the one way to achieve this. Here by minimizing sum of squared residuals (ϵ_i^2) which is equal to the inner product of the residuals vector with itself($\sum \epsilon_i^2 = \epsilon^T \epsilon$), optimal β vector can be found. To minimize $\epsilon^T \epsilon$ which is equivalent to $(Y - X\beta)^T(Y - X\beta)$, bellow condition should be satisfied.

$$\frac{\partial}{\partial \beta}\left[\epsilon^T \epsilon\right] = 0 \tag{4}$$

Following equation (5) can be derived after solving this ($\epsilon^T \epsilon$),

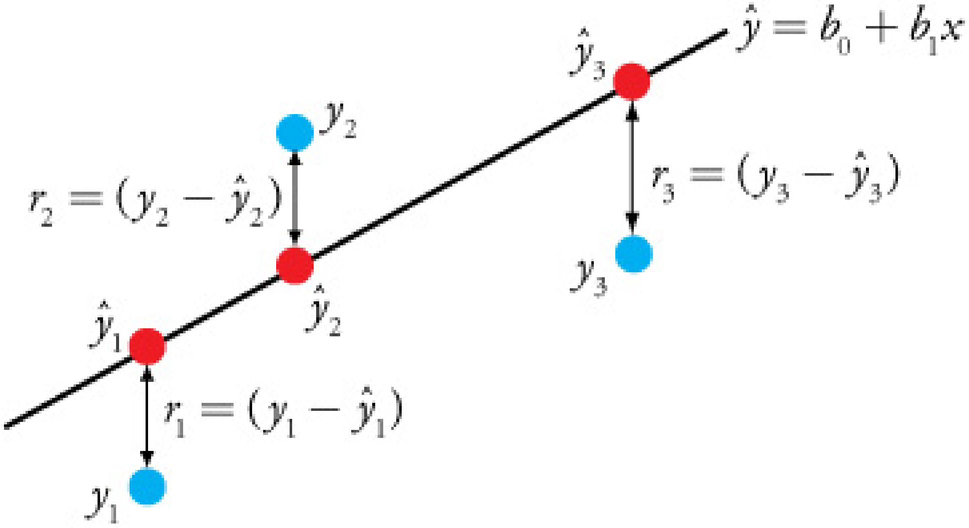

$\hat{y} = b_0 + b_1 x$
$\hat{y}_3$
y_2
$r_2 = (y_2 - \hat{y}_2)$
$r_3 = (y_3 - \hat{y}_3)$
y_3
$\hat{y}_1$
$\hat{y}_2$
$r_1 = (y_1 - \hat{y}_1)$
y_1

Figure 1: One way of calculating residuals

$$y^T y - 2\beta^T X^T y + \beta^T X^T X\beta. \tag{5}$$

Due to this identity $\frac{\partial \mathbf{a}^T \mathbf{b}}{\partial \mathbf{a}} = \mathrm{b}$, for vectors a and b, solution for β can be found as bellow.

$$\beta = (X^T X)^{-1} X^T y \tag{6}$$

2 Introduction to Logistic Regression

Logistic Regression is commonly used for classification, as it can output a value that corresponds to the probability of belonging to a given class or get a more accurate estimate based on some threshold values. Model computes a weighted sum of the input features. Example formula for the logistic function would be something similar to this

$$p = \frac{1}{1 + 10^{-A(d-B)}} \tag{7}$$

As you can see, the formula has two parameters, A and B. Changing these parameters affects the exact shape of the logistic function which means result of regression depends on the selecting proper values for A and B. Above expression can be rewritten as follows,

$$-\log(\frac{1-p}{p}) = A(d - B) \tag{8}$$

This can be rearrange into a formula which is equivalent to y = a x + b, where a = A, b = -AB, x=d and $y = -\log(\frac{1-p}{p})$.

To find parameters (A, B), you can use Stochastic Gradient Descent algorithm.

3 Understanding the scikit-learn estimator API

There are three methods of estimators:fit, $fit\_transform$ and $predict$. fit method is used to learn the parameters from the training data. $fit\_transform()$ method uses those parameters to transform the data into desired format. Predictions about new data samples via the predict(). Basic view of scikit-learn estimator API is shown in Fig. 2

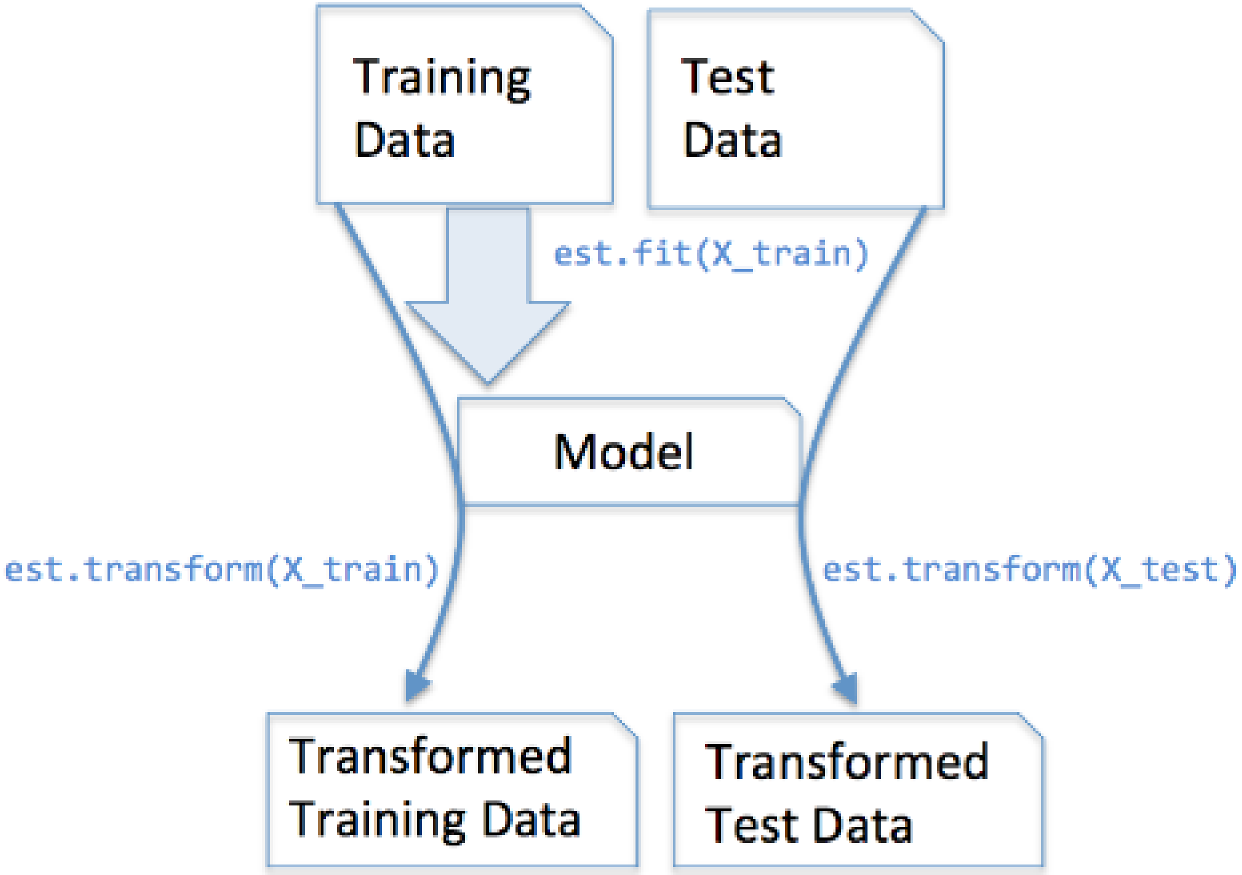


Figure 2: Scikit-learn estimator API basic view

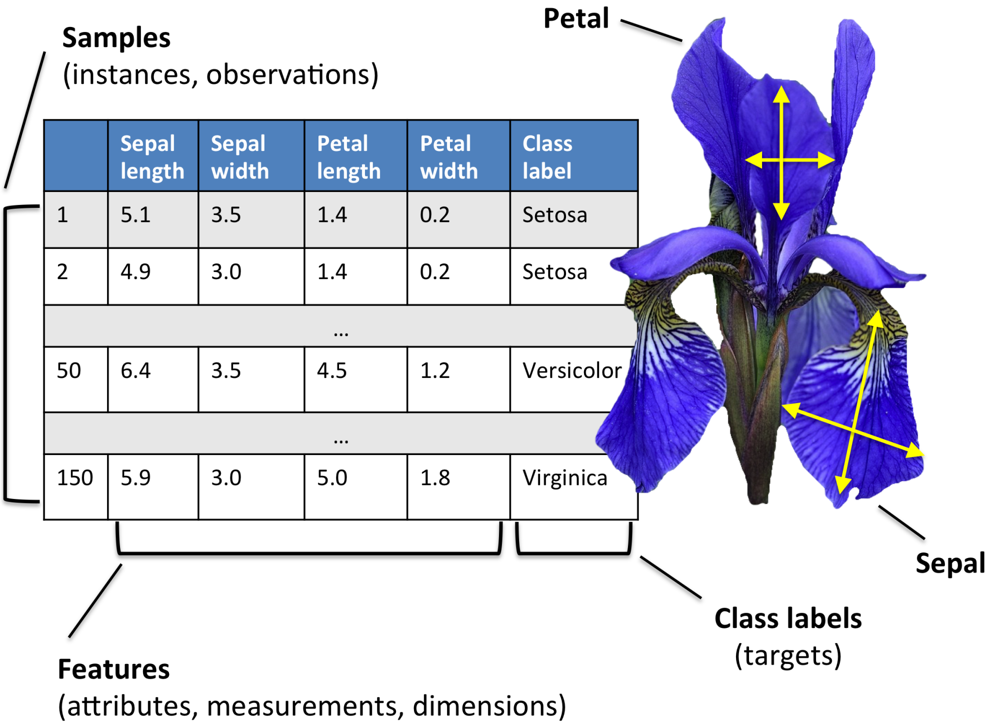
4 Try Out

4.1 To familiar with scikit-learn estimate API

Iris dataset is a famous dataset that contains 150 iris flowers of three different species: Iris-Setosa, Iris-Versicolor and Iris-Virginica and each entry consists of four features: sepal length, sepal width, petal width and petal length which are illustrated in Fig. 3. Let's try to build a model to detect the Iris-Virginica type based only on the petal width feature.

1. To load the iris dataset.

```
from sklearn import datasets
iris=datasets.load_iris()
```



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			…		
50	6.4	3.5	4.5	1.2	Versicolor
			…		
150	5.9	3.0	5.0	1.8	Virginica

Figure 3: A sample view of Iris dataset

2. Print *iris* and see what it contains

3. Load the petal width feature into X

4. Load the class values into Y which should meet following condition.

```
if class value (target) is equals to 2 then
    $value \leftarrow 1$
else
    $value \leftarrow 0$
end if
```

5. To train a logistic regression model

```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
X= ?
y= ?
X_train, X_val, y_train, y_val = train_test_split(X, y, test_size=0.2)
log_reg=LogisticRegression()
log_reg.fit(X_train,y_train)
y_proba=log_reg.predict(X_val)
```

6. To train a linear regression model

```
from sklearn.linear_model import LinearRegression
X= ?
y= ?
X_train, X_val, y_train, y_val = train_test_split(X, y, test_size=0.2)
lin_reg=LinearRegression()
lin_reg.fit(X_train,y_train)
lin_reg.predict(X_val)
```

5 Lab Exercise

5.1 Try to create a model for predicting angles

1. Load lab04ExerciseAngles and lab04ExerciseChannels dataset while specifying column names for lab04ExerciseChannels (channel1, channel2, channel3, channel4, channel5) and for lab04ExerciseAngles (angle1, angle2, angle3) using *pandas.read_csv* function.

2. Merge them into a one data frame.

3. Select either linear regression or logistic regression for the model building and comment on reasons for selecting it.

4. Select one or more channels as features and one angle data for predicting. Comment on selection criteria.

5. Using scikit-learn estimator API build the model and comment on the accuracy of your model

6 Submission

Submit a single .py file as [12|13]xxxlab04.py where xxx is your registration number. Add answers for questions 3,4 and 5 as comments in the same file.

7 Important

Make sure that you have the basic understanding of visualizations. This lab is really important for successive labs. If you do not understand any concepts, make sure you get some help from instructors.

8 Deadline

August 01, 23:59:59 GMT+5:30.