TRINITY COLLEGE DUBLIN

School of Computer Science and Statistics

Week 3 Assignment CS7CS4/CSU44061 Machine Learning

Rules of the game:

- **Its ok to discuss with others, but do not show any code you write to others. You must write answers in your own words and write code entirely yourself. All submissions will be checked for plagiarism.**
- **Reports must be typed (no handwritten answers please) and submitted as a separate pdf on Blackboard (not as part of a zip file please).**
- **Important: For each problem, your primary aim is to articulate that you understand what you're doing - not just running a program and quoting numbers it outputs. Long rambling answers and "brain dumps" are not the way to achieve this. If you write code to carry out a calculation you need to discuss/explain what that code does, and if you present numerical results you need to discuss their interpretation. Generally most of the credit is given for the explanation/analysis as opposed to the code/numerical answer. Saying "see code" is not good enough, even if code contains comments. Similarly, standalone numbers or plots without further comment is not good enough.**
- **When your answer includes a plot be sure to (i) label the axes, (ii) make sure all the text (including axes labels/ticks) is large enough to be clearly legible and (iii) explain in text what the plot shows.**
- **Include the source of code written for the assignment as an appendix in your submitted pdf report. Also include a separate zip file containing the executable code and any data files needed. Programs should be running code written in Python, and should load data etc when run so that we can unzip your submission and just directly run it to check that it works. Keep code brief and clean with meaningful variable names etc.**
- **Reports should typically be about 5 pages, with 10 pages the upper limit (excluding appendix with code). If you go over 10 pages then the extra pages will not be marked.**
- **When selecting a hyperparameter value show cross-validation analysis to justify your choice.**
- **When evaluating an ML algorithm on data, compare against a baseline model.**

Downloading Dataset

- Download the assignment dataset from `https://www.scss.tcd.ie/Doug.Leith/CSU44061/week3.php`. Important: You must fetch your own copy of the dataset, do not use the dataset downloaded by someone else.
- Please cut and paste the first line of the data file (which begins with a #) and include in your submission as it identifies your dataset.
- The data file consists of three columns of data (plus the first header line). The first two columns are input features and the third column is the real-valued target value.

.

Assignment

In this assignment you'll use sklearn to train and evaluate Lasso regression models on the data you downloaded. Recall that Lasso regression uses a linear model and mean square cost function with an L_1 penalty and that the L_1 penalty has a weight parameter C in the lecture notes (weight parameter $\alpha = 1/(2C)$ in sklearn).

(i) (a) Plot the data you downloaded as a 3D scatter plot i.e. with the first feature on the x-axis, the second feature in the y-axis and the target on the z-axis. You can use the matplotlib scatter function for this, e..g for training data with two features X and target y

```
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')
ax.scatter(X[:,0],X[:,1],y)
```

Does it look like the training data lies on a plane or a curve?

(b) In addition to the two features in the data file add extra polynomial features equal to all combinations of powers of the two features up to power 5 (you can use the sklearn PolynomialFeatures function to do this). Now train Lasso regression models with these polynomial features for a large range of values of C e.g. 1, 10, 1000 (you might need to adjust these values for your data, start by making C small enough that the trained model has all parameters zero, then increase from there). Report the parameters of the trained models (don't just give a list of numbers, say what feature each parameter value corresponds to), discuss how they change as C is varied.

(c) For each of the models from (b) generate predictions for the target variable. Generate these predictions on a grid of feature values. You can use a couple of nested for loops for this e.g.

```
import numpy as np
Xtest=[]
grid=np.linspace(-5,5)
for i in grid:
        for j in grid:
                Xtest.append([i,j])
Xtest = np.array(Xtest)
```

This grid should extend beyond the range of values in the dataset e.g. if the first feature in the dataset has values from 0 to 2 generate predictions for values from -5 to 5 or thereabouts. Plot these predictions on a 3D data plot and also show the training data. Adjust the grid range used for the predictions so that the training data can still be clearly seen in the plot. Its up to you to decide how best to plot this data but do try to make your plot easy to read (suggestion: it can be helpful to plot the predictions as a surface using the matplotlib plot_surface command and the training data as points using the matplotlib scatter command, be sure to add a legend to identify the different curves). With reference to this plot discuss how the predictions change as C is varied.

(d) What is under- and over-fitting? Using your parameter data from (b) and visualisation from (c) explain how penalty weight parameter C can be used to manage to trade-off between under- and over-fitting the data.

(e) Repeat (b)-(c) for a Ridge Regression model. This uses an L_2 penalty instead of an L_1 penalty in the cost function. Compare the impact on the model parameters of changing C with Lasso Regression and with Ridge Regression.

(ii) Using the Lasso model with polynomial features from (i) you'll now look at using cross-validation to select C.

(a) Use 5-fold cross-validation to plot the mean and standard deviation of the prediction error vs C. Use the matplotlib errorbar function for this. You will need to choose the range of values of C to plot, justify your choice.

(b) Based on the cross-validation data what value of C would you recommend be used here? Importantly, explain the reasons for your choice.

(c) Repeat (b)-(c) for a Ridge Regression model.