Statistical Methods Assignment

MAKE SURE TO READ THE ENTIRE PM BEFORE STARTING

1 Introduction

Linear regression is a fundamental machine learning approach and the oldest form of statistical inference. It is still highly relevant in many applications, in particular those where good statistics are more important than precision and accuracy. If we imagine an axis plotting how amenable a method is to statistical analysis, linear regression would be "most amenable" and neural networks would be on the other end of the scale. In this assignment, you will implement *multiple linear regression* with comprehensive statistics using only `numpy` for linear algebra and `scipy.stats` for additional statistics. You are obviously not allowed to use existing linear regressions from those or `pyplot`, `sklearn`, `seaborn` or other packages.

2 Method

Data

In this assignment we will use a dataset from the book we will be using in the Machine Learning course. However, here we will be treating the dataset quite differently from the later pure Machine Learning approach. This will be a *statistical* analysis, while machine learning is a purely *numerical* approach. The dataset concerns housing prices in California in 1997. Use your knowledge from previous courses to explore the dataset so you understand what you are supposed to be doing. You will have to make several decisions about your model based on this exploration, for example the number of features and what features to include. For the higher grade (VG) there is a feature in the dataset that must be included, but you have to make an informed choice about the others. One choice, that you might decide is appropriate, is to include everything.

Ordinary Least Squares

A *linear model* can be expressed mathematically as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... + \beta_d X_d$$

where d is the *dimension*, or *number of features* of the model. To find an approximation of the β coefficients, an Ordinary Least Squares method is usually employed. To calculate the coefficients, a vector denoted $\mathbf{b}$ after they have been approximated, the formula is:

$$\mathbf{b} = (X^T X)^{-1} X^T Y$$

Dependency checks

In order for a linear regression to be relevant, the variables X_i should not have linear dependencies between them. If they do, the right-hand side of the equation system is not necessarily significant. Even if we have some dependency, our statistics can still be good enough for some confidence interval. There are several commonly used statistics to check for linear dependence, but a very frequently used statistic is the Pearson correlation number (r), available as `scipy.stats.pearsonr(...)`:

$$r = \frac{\texttt{Cov}(X_a, X_b)}{\sqrt{\texttt{Var}X_a \texttt{Var}X_b}}$$

Quantitative statistics

In a linear regression, the *mean* or *expected value*, is the regression line itself. The other two moments we have discussed are *variance* (σ^2) and *standard deviation* (σ). The variance has a slightly different definition for *multiple* linear regression as opposed to *simple* linear regression. For the multiple case, an unbiased estimator of the variance is:

$$\hat{\sigma}^2 = \frac{\texttt{SSE}}{n - d - 1}$$

where

$$\texttt{SSE} = \sum_{i=1}^{n} (Y_i - X\mathbf{b})^2$$

The MSE, or Mean Squared Error is the mean of SSE and the RMSE is the square root of the MSE. Note that these values are *not* unbiased estimates, but they are frequently used in machine learning.

Confidence and significance

To analyse a linear regression there are two main analytical tools: *confidence intervals* and *significance tests*. To test whether the regression is significant, the test statistic below will follow an F distribution with d and $n - d - 1$ degrees of freedom.

$$\frac{\texttt{SSR}/d}{\hat{\sigma}^2}$$

where

$$S_{yy} = \sum_{i=1}^{n} (Y - \mu_y)^2 = \frac{n \sum_{i=1}^{n} (Y^2) - (\sum_{i=1}^{n} Y)^2}{n}$$
$$\texttt{SSR} = S_{yy} - \texttt{SSE}$$

The F-distribution is available through `scipy.stats.f(d, n-d-1)` and it's survival function is `scipy.stats.f.sf(...)`. The test rejects (survival function has low p-value) for large values of the statistic. This tests significance of *all* parameters at once. To test the significance of individual parameters the following statistic can be used:

$$\frac{\hat{\beta}_i}{\hat{\sigma}\sqrt{c_{ii}}}$$

where c_{ii} is the entry on row i and column i of the *variance/covariance* matrix given by $c = (X^T X)^{-1}\sigma^2$. This test follows a T-distribution with $n - d - 1$ degrees of freedom and is a two-sided test that rejects for too large or too small values of the statistic. The T-distribution is available through `scipy.stats.t(...)`. The cumulative distribution function and the survival function are available as `scipy.stats.t.cdf(...)` and `scipy.stats.t.sf(...)` respectively. A two-sided test is calculated as $p = 2 * \texttt{min}(\texttt{cdf}, \texttt{sf})$.

To determine the *coefficient of multiple determination*, R^2, use the following formula:

$$R^2 = \frac{SSR}{S_{yy}}$$

A confidence interval for the individual parameters (features) of the model is given by:

$$\hat{\beta}_i \pm t_{\alpha/2}\hat{\sigma}\sqrt{c_{ii}}$$

where $t_{\alpha/2}$ is the appropriate point based on the T_{n-d-1} distribution and a confidence level α.

3 Result

The assignment should be coded in a `.py` file and should not import anything else but modules `numpy` and `scipy.stats`. The mainstay of the code should be a `LinearRegression` class/type. A second file is needed for the hand-in, an `.ipynb` notebook. The latter file should call the code in the former and demonstrate the functionality.

For the passing grade(**G**) the following should be implemented:

- A function or method that performs a least squared approximation of the mean given a dataset.
- A property or field `d` that contains the number of features/parameters/dimensions of the model.
- A property or field `n` that contains the size of the sample.
- A function or method to calculate the sample variance.
- A function or method to calculate the standard deviation.
- A function or method to calculate the RMSE.

For the higher grade (**VG**) the code must adhere to the stated requirements, in particular with respect to modularity, and the following is also needed:

- A function or method that reports the significance of the regression.
- A function or method that reports the relevance of the regression (R^2).
- Significance tests on individual variables, in particular categorical variables.
- A function or method that calculates the Pearson number between all pairs of parameters.
- Confidence intervals on individual parameters.
- Being able to set a `confidence_level` for the statistics.
- Inclusion of categorical data in the model

4 Discussion

Hand in the assignment by giving a link to a git-repo, or include all code-files in the form on ITHS. The notebook file should present the results of running the code on the given data by using the components in a clear and simple manner. Don't use comments or print statements in your python cells.