SortSearch

Release 0.1.3

Daniel Mesejo

Aug 15, 2022

NOTES

This a place to sort the notes by Daniel Mesejo to be able to easily search them later.

CHAPTER

ONE

A DEEP DIVE INTO SKLEARN.LINEARREGRESSION [WIP]

This post is inspired by a similar one for the R programming language, A Deep Dive Into How R Fits a Linear Model. With the difference that we would take a look at the relative simple mathematics behind fitting a line using linear regression

1.1 A Simple Example

```
import numpy as np
from sklearn.linear_model import LinearRegression
X = np.array([[1, 1], [1, 2], [2, 2], [2, 3]])
# y = 1 * x_0 + 2 * x_1 + 3
y = np.dot(X, np.array([1, 2])) + 3
reg = LinearRegression().fit(X, y)
```

1.2 The Big Picture

Before deep diving one can learn a lot by reading the documentation of sklearn.LinearRegression, from the documentation:

> From the implementation point of view, this is just plain Ordinary Least Squares (scipy.linalg.lstsq) or Non Negative Least Squares (scipy.optimize.nnls) wrapped as a predictor object.

Additionally if one reads the documentation of scipy.linalg.lstsq, one learns that

> Compute least-squares solution to equation Ax = b. Compute a vector x such that the 2-norm $||b - Ax||$ is minimized.

Going one layer below we notice that the least squares problem is solved using the gelsd LAPACK driver and in turn we get that the problem is solved as in three steps:

1. Reduce the coefficient matrix A A to bidiagonal form with Householder transformations, reducing the original problem into a "bidiagonal least squares problem" (BLS).
2. Solve the BLS using a divide and conquer approach.
3. Apply back all the Householder transformations to solve the original least squares problem.

The text above is a broad view of the steps involved when calling LinearRegression.fit

---

Note: Using ordinary least squares is not the only way to solve the fitting problem, but it is attractive for several reasons:

1. It's really mathematically attractive. $||x||^2$ is a smooth function of x, and the solution to the least squares problem is a linear function of b
2. There's a nice picture that goes with it – the least squares solution is the projection of b onto the span of A, and the residual at the least squares solution is orthogonal to the span of A.
3. It's a mathematically reasonable choice in statistical settings when the data vector b is contaminated by Gaussian noise (see Gauss-Markov theorem).

1.3 The Python Layer

This section includes the Python layer of the code