« Previous (/course/cs357-f15/flow-session/74262/2/)

下一页 » (/course/cs357-f15/flow-session/74262/4/)

跳转至最末 »

1 2 3 (/course/cs357-f15/flow-session/74262/0/) (/course/cs357-f15/flow-session/74262/1/)
4 5 (/course/cs357-f15/flow-session/74262/2/) (/course/cs357-f15/flow-
6 session/74262/4/) (/course/cs357-f15/flow-session/74262/5/)

Least Squares Theory

Some essential theory for this problem is given below. It is suggested that you also revisit the course notes on solving a Least Squares problem with QR factorization here (http://andreask.cs.illinois.edu/cs357-s15/public/notes/06-qr-applications.pdf)

Consider an example set of data D^{7x2} , which has 7 rows and 2 columns and can be used to predict a result (in our case this will be using the tumor characteristics of each patient to predict if the patient has a benign or malignant tumor). D looks like,

$$D = \begin{bmatrix} d_{0,0} & d_{0,1} \\ d_{1,0} & d_{1,1} \\ \vdots & \vdots \\ d_{6,0} & d_{6,1} \end{bmatrix}$$

Now consider you have a set of results b^{7x1} and you'd like to know how the data D can be used to best approximate the solution b. The least squares formulation serves this purposes by solving the following problem, $Aw = b$. Where A and b are known and A is some representation of the data D.

A **linear representation** of D would be,

$$A = \begin{bmatrix} d_{0,0} & d_{0,1} \\ d_{1,0} & d_{1,1} \\ \vdots & \vdots \\ d_{6,0} & d_{6,1} \end{bmatrix}$$

A **quadratic representation** of D would be,

$$\tilde{A} = \begin{bmatrix} d_{0,0} & d_{0,1} & d_{0,0}^2 & d_{0,1}^2 & d_{0,0} \times d_{0,1} \\ d_{1,0} & d_{1,1} & d_{1,0}^2 & d_{1,1}^2 & d_{1,0} \times d_{1,1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{6,0} & d_{6,1} & d_{6,0}^2 & d_{6,1}^2 & d_{6,0} \times d_{6,1} \end{bmatrix}$$

Once you have construct A or $\tilde{A}$, QR factorization can be used to solve the system for w.

Then, given a new set of data $E^{n,2}$ and the w just solved for, a predictive result can be solved for by putting E in the same form as either A or $\tilde{A}$ and simply performing a matrix-vector multiplication.

Lastly, the example given here had 2 columns, but least squares applies for problems with an arbitrary number of rows and columns. Although one needs more rows than columns. For the problem given in this homework, you will be asked to create a quadratic representation with 4 columns of data. The matrix $\tilde{A}$ will then have the following form:

$$\tilde{A} = [\alpha_0,\ \alpha_1,\ \alpha_2,\ \alpha_3,\ \alpha_0^2,\ \alpha_1^2,\ \alpha_2^2,\ \alpha_3^2,\ \alpha_0\alpha_1,\ \alpha_0\alpha_2,\ \alpha_0\alpha_3,\ \alpha_1\alpha_2,\ \alpha_1\alpha_3,\ \alpha_2\alpha_3]$$

Where each α_i represents a column.

1. Construct a linear least squares representation of the data in *breast-cancer-train.dat* and *breast-cancer-validate.dat*. (i.e. 2 matrices).
   - Call the linear least squares representation for *breast-cancer-train.dat* `A_linear`.
2. Construct a right-hand side vector b for both data sets. To create b, make a numpy 1D-array the same size as the *Malignant/Benign* column of your data set and set each entry to 1 if the patient's tumor was malignant, otherwise set it to -1.
   - Call the right-hand side vector for *breast-cancer-train.dat* `b`.