Notes on Least Squares and Nearest Neighbors

compiled from Elements of Statistical Learning by D. Gueorguiev, Oct 25, 2024

# Linear Models and Least Squares

The linear model is defined as:

Given a vector of inputs the output is predicted as:

(1)

where is the *intercept* or *bias*. Often, we include a unit constant ‘variable’ additionally in X and include the bias in the vector of the coefficients . Then we can rewrite (1) as :

(2)

where denotes a row vector or matrix transpose ( is a column vector).

Here we are modeling a single output so is a scalar; in general can be a -vector, in which case would be a matrix of coefficients. In the – dimensional input-output space, represents a hyperplane. If the constant is included in , then the hyperplane includes the origin and is a subspace; If not, it is an affine set cutting the -axis at the point . From now we assume that the intercept is included in .

Obviously, is a linear function over the -dimensional space and the gradient is a vector in input space that points in the steepest uphill direction.

Fitting the linear model to a set of training data

Popular method is least squares- we pick the coefficients to minimize the *residual sum of squares (RSS)* as:

(3)

is a quadratic function of the parameters, and hence its minimum always exists, but may not be unique. The solution can be written in matrix notation as:

(4)

where is an matrix with each row an input vector, and is an -vector of the outputs in the training set. Differentiating w.r.t. we get the *normal equations*:

(5)

If is non-singular, then the unique solution is given by:

(6)

and the fitted value at the -th input is .

## Insight in the Optimality of the Linear Regression with an Example

Let us look into a linear model in a classification context. The Figure 1 below shows a scatterplot of the training data.

The data is simulated. The output class variable has the values Blue or Orange and is represented as such on the scatterplot. There are 100 points in each of these two classes. The linear regression model was fit to these data, with the response coded as 0 for Blue and 1 for Orange. The fitted values are converted to a fitted class variable

A graph of a linear regression

Description automatically generated with medium confidence

Figure: A classification example in two dimensions.

# References

[1] [The Elements of Statistical Learning; Data Mining, Inference, Prediction, Trevor Hastie, Robert Tibshirani, Jerome Friedman, Second Edition, 2017](https://github.com/dimitarpg13/statistical_learning_and_kernel_methods/blob/main/literature/books/EelementsOfStatisticalLearning_print12.pdf)

# Appendix

## Affine Space

### Informal discussion

Affine space is what is left from vector space after one has forgotten which point is the origin.

Imagine that there are two observers – observer *A* and observer *B*. Observer *A* knows the real origin, but Observer *B* believes that another point is the origin. Two vectors , and , are to be added. Observer *B* believes he has computed the distance but observer A knows better – B has actually computed .

Similarly, observers *A* and *B* may evaluate any linear combination of and or of any finite set of vectors and will generally get different answers. However, if the sum of the coefficients is unit then the observers will arrive at the same answer. Moreover, for any finite number of observers if the sum of the coefficients is unity each of them will come to the same answer.

Origin according to observer B

Origin according to observer A

*according to origin A*

*according to origin B*

according to origin A

*according to origin A*

Figure A.1: origins and vector computations from the perspectives of observer A (red) and observer B (blue)

In the example on Figure A.1 if observer *A* travels to and observer *B* travels using the same route then observer *A* will observe from his frame of reference the following path of observer *B* :

So under the condition all observers who have different frames of reference will describe the same point with the same linear combination despite the different origins. While only observer *A* on the earlier example knows the *linear structure* both observers *A* and *B* know the *affine structure* that is, the values of affine combinations defined as linear combinations in which the sum of the coefficients is unity.

**Definition** *Affine Space*

A set together with a vector space , and a transitive and free action of the additive group of on the set . The elements of the affine space are called points. The vector space is said to be associated with the affine space and its elements are called *translations,* or *free vectors*.

The action in affine space will be denoted as addition and it is a mapping having the following properties:

1) *Right identity*:

, where is the zero vector in

2) *Associativity*:

, (here the last is the addition in )

3) *Free and transitive action*:

, is a bijection

The first two properties are simply defining properties of a (right) group action. The third property characterizes free and transitive actions, and onto character coming from transitivity, and then the injective character follows from the action being free. Forth property follows from 1) and 2)

4) *Existence of one-to-one translations*

the mapping is a bijection

3) is often used in the following equivalent form :

5) *Subtraction*

denoted with such that