Supervised Learning and a Bit of Statistical Decision Theory

compiled from Elements of Statistical Learning by D. Gueorguiev, June 29, 2025

# A Bit of Statistical Decision Theory

Let denote a real valued random vector, and a real valued random output variable, with joint distribution . We seek a function for predicting given values of the input . This approach requires a *loss function* for penalizing errors in prediction. The most common loss function is the *squared error loss* . This leads us to a criterion for choosing ,

(1)

We condition on which allows us to write:

(2)

We rewrite (2) as :

(3)

where is the conditional probability density function of given and is the probability density function of .

Substituting (3) in (1) yields

(4)

One can show that in order to construct it suffices to minimize pointwise:

(5)

The solution is

(6)

the conditional expectation also known as the *regression function*. Thus the best prediction of at any point is the conditional mean, when best is measured by average squared error.

The nearest-neighbor methods attempt to directly implement this recipe using the training data – at each point we can ask for the average of those ’s with input . Thus we can write

where avg denotes average along the neighborhood of training data points closest to .

Note: two approximations are happening in this discourse:

* Expectation is approximated by averaging over sample data;
* Conditioning at a point is relaxed to conditioning on some region close to the target point

For large training sample size , the points in the neighborhood are likely to be close to , and as gets large the average will get more stable. In fact, under mild regularity conditions on the joint probability distribution , one can show that as such that , .

**Question**: when the nearest neighbor model is not appropriate?

(i) When we do not have large enough sample, we may get more stable estimate than the k-nearest neighbors.

(ii) Also when the number of dimensions p gets large the metric size of the k nearest neighbor increases, and this leads to decrease in the rate of convergence for the nearest neighborhood as a surrogate for conditioning.

So , let us instead of nearest neighbors consider linear model.

Now we assume that , where is a r.v. with mean and variance , and . The expected predicted error () under the squared error loss is:

(7)

We regard the expression (15) as a function of , a column vector of length . So we find the minimum of as

(8)

Thus we deduce

(9)

# 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