Jared Yu

Data Mining

Module 7 Assignment

1. Question 1
   1. Show that the distance from the hyperplane to the point is by minimizing subject to the constraint .

Ans: References: [1.1], [1.2]

To solve this, I will use Lagrange multipliers. We are asked to minimize subject to the constraint . Then the function for the Lagrange multipliers is in the form of,

where

Then we must minimize this function w.r.t. each of the variables,

Therefore, the distance after taking the square root can be seen to as follows,

* 1. Show that the projection of onto the hyperplane is given by

Ans:

To prove this, we will first indicate what is. The textbook states that if and are both on the decision surface, then

or

This indicates that the constant vector is actually normal or perpendicular to the hyperplane.

Then, using the result from part a), we have that the distance between some arbitrary vector and the hyperplane can be found with . What we want to do then is to multiply this minimum distance by , which is the unit vector form of . Furthermore, let represent the projection of onto the hyperplane. This leads us to the following formula,

1. Let be -dimensional samples and be any nonsingular positive definite matrix. Show that the vector that minimizes

Is the sample mean, .

Ans: References: [2.1], [2.2]

Let the function be defined as follows

To try and find the vector that minimizes, we must first take the gradient w.r.t. . To begin, we can try to simplify .

Next, we can find the derivative of this function by utilizing the derivative of an inverse matrix w.r.t. a vector.

Then to show that it is indeed the minimum, the second derivative must also be examined.

Then, since is nonsingular positive definite, then is positive definite. Therefore, can be said to be the point that minimizes .

1. Consider a linear classifier with discriminant functions . Show that the decision regions are convex by showing that if and then if .

Ans: References: [3.1], [3.2], [3.3], [3.4], [3.5]

Let us define , where , as the convex combination of vectors and . Furthermore, the set of vectors within is convex if it contains all possible convex combinations of vectors. If this can be shown to be the case, then that implies that all decision regions , for are also convex.

Based on the linearity of the classifier, , we can also write

Now, since and , and the weights and are positive, then the following also holds,

From this it follows that,

Therefore, it can be concluded that,

This shows then that the decision regions , are convex.

1. In the gradient descent algorithm, is obtained from by

where is a positive scale factor that sets the step size. Consider the criterion function

where is the set of samples for which . Suppose that is the only sample in . Show that and that the matrix of second partial derivatives is given by . Use this to show that when the optimal is used in the gradient descent algorithm,

Ans: Reference: [4.1, 2.2]

The first step is to show that . In the case where only contains , then . Finding the derivative of this w.r.t. , we find that,

To find the matrix of second partial derivatives, we can take the partial derivative again to see that,

To find , we can use the formula for from the textbook.

Then, going back to the update formula we have the following,

1. Show that the partial derivatives of the functions used in multiple class logistic discrimination are given by

where

Ans: References: [5.1], [5.2]

To solve , we must look at two cases. We must look for when and when . This will yield a piecewise equation shown below.

Solving requires the use of the quotient rule, where when . In this case, can be thought of as and can be thought of as . With , the derivative w.r.t. for some arbitrary is always . However, looking at , the derivative w.r.t. for some arbitrary is only when .

To prove equation (5.1), we can first look at the case of . Solving for we get

Then in the case of we have the following.

Therefore, equation (5.1) leads to the following,

Next, we must define the Kronecker delta to be the following,

After combining the Kronecker delta into the equation (5.2) we can get the following,

**Reference:**

[1.1] <https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf>

[1.2] <https://math.stackexchange.com/questions/1210545/distance-from-a-point-to-a-hyperplane>

[1.3] <https://piazza.com/class/kc0jkwru805u1?cid=143>

[2.1] <https://piazza.com/class/kc0jkwru805u1?cid=147>

[2.2] Duda, R. O. (2000). R. O. Duda’s P. E. Hart’s D. G. Stork’s Pattern Classification (Pattern Classification (2nd Edition) [Hardcover])(2000) (2 edition). Wiley-Interscience.

[3.1] <https://www.cs.toronto.edu/~urtasun/courses/CSC411_Fall16/07_multiclass.pdf>

[3.2 ] <https://math.stackexchange.com/questions/404143/what-is-convex-combination-of-two-points>

[3.3] <https://en.wikipedia.org/wiki/Convex_combination>

[3.4] <https://mathworld.wolfram.com/ConvexCombination.html>

[3.5] <https://piazza.com/class/kc0jkwru805u1?cid=144>

[4.1] <https://piazza.com/class/kc0jkwru805u1?cid=148>

[5.1] <https://www.ics.uci.edu/~pjsadows/notes.pdf>

[5.2] <https://eli.thegreenplace.net/2016/the-softmax-function-and-its-derivative/>