Intro To ML – Kernels, GD and PCA

Amit Zulan – 207299033, Dvir Ben Zikri - 315409508

1. Code is added separately.
2. 2222222

We need to show that this function cannot be a kernel function, to do that we will show that this function violates the commutative property of the inner product function, since a kernel function is defined as: .

If we switch and in the above function, we get:

It can also be seen by the fact that we use different index values of the vectors x and x’ in the above function.

We need to show that this function cannot be a kernel function.

We shall look at the case when :

Since and in contradiction to definition of K as an inner product of two vector, which in this case are the same which results in norm.

We need to find :

Thus, we can see that should be:

1. Consider a linear combination of convex functions

For those we get g:

To prove that this function is not convex we need to show that for all two vectors

\* Since f is convex function.

So, we got:

Since equality is only for linear or constant functions which two are always convex, if is not either, we get:

If is indeed a constant or linear function, we can choose a different function which is not a constant or linear function, and if such a function does not exist than any linear combination of linear or constant functions is convex -> g will be convex.

We need to show that for any two vectors

\* This inequality comes from the fact that each is convex and that each which does not change the inequality.

And for

So, we got that for g is convex.

1. Let where the input domain is and the labels are from . Consider learning of a linear predictor using the following optimization for :

Where .

1. First, we’ll write this in matrix form.

Where

To find the w that minimizes the above formula we shall find the formula’s gradient with respect to w.

The gradient is 0 when:

Since is invertible:

1. To calculate the step of the GD algorithm we need to find the gradient of

Which we did above so:

1. To calculate the step of the SGD we need to separate to :

The step of SGD is the gradient of and for a uniformly randomly selected . The gradients are:

So, the SGD step is:

* 1. In the 1st experiment it turned out that in all times t,

And thus

We got that and are linearly dependent. All in all, we get that .

Since at least two of A’s eigenvalues are equal to 0.

Since the best distortion is equal to the 2 lowest eigenvalues thus .

* 1. In another experiment it turned out that in all times t,

Since now the are not necessarily dependent, if we would pick at least , then for at least 3 linearly independent samples we would get at least 3 eigenvalues different than 0.

Since A is positive semi definite all 3 will be positive thus the best distortion must be larger than 0 and thus larger than section a.

For example, Let:

The eigenvalues of A are:

In this case the best distortion is 0.8.