COMP 632: Assignment 2

Due on Wednesday, February 18 2015

Presented to Dr. Doina Precup

Geoffrey Stanley
Student ID: 260645907

Question 1

A)

For a function to be considered a kernel function the kernel matrix defined as $K_{ij} = K(x_i, x_j)$ must have two properties:

1. be symmetric
2. be positive semidefinite

As such, a Kernel matrix must abide by the following:

$$K_{ij} = K_{ji} \tag{1}$$

$$z^T K z \geq 0 \tag{2}$$

Where z is an arbitrary vector.

A kernel function is also one that can be expressed as a dot product of the feature vectors of the instances:

$$K(x, z) = \phi(x) \cdot \phi(x) \tag{3}$$

where ϕ is a feature maping of input features to a vector space. Further more, as described by Bishop(2006) p.296 a kernel function can be a construction of kernel functions such that

$$k(x, z) = k_1(x, z) + k_2(x, z) \tag{4}$$

In this particular case it would be usefull to decompose our K_l kernal function into:

$$k_l(x, z) = k_1(x, z) + k_2(x, z) + ... + k_l(x, z) \tag{5}$$

where the components of the decomposition evaluate the similarity of a particular character length such that $k_1('ar','ark') = 1$ and $k_2('ar','ark') = 2$. Now, in order to show that K_l is a kernal function, we are left with having to create a feature mapping onto some vector where dot products of those vector would result in the similarity required.

For simplicity let's define our alphabet as being a, b, c and our mapping to be onto a three dimensional vector representing our alphabet. Now we can map an input string such as "ab" onto a vector $x = [1, 1, 0]$. Given another string "bc" we can create $y = [0, 1, 1]$ where $x^T y = 1$. Expanding this mapping into longer character combinations we can now define our vector space as being all combination of character sequences of length 1 through l.

We now have multiple functions that can be expressed in terms of a dot product of vectors whose sum is equal to our original K_l function and have thus demonstrated that it is a kernal function.

B)

As l increases words will have a tendency of having higher scores when compared with itself then any other words. This will result in a diagonal matrix.

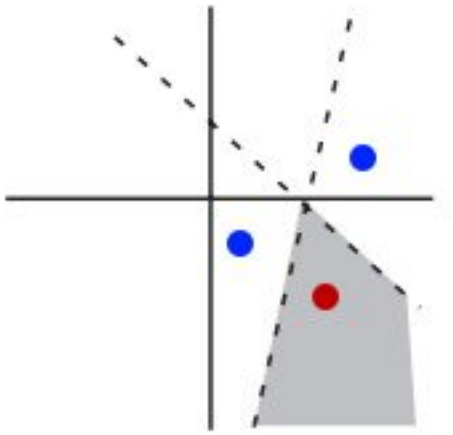
C)

Yes.

D)

Question 2

A)



B)

The VC-dimension of this hypthesis class is 4. This is because it can successfully shatter all configurations of 4 points. However, it would not be able to do so for all configuration of 5 points.

C)

The VC-dimension of any type of boolean combination of 2 linear classifiers is also 4.

Question 3

A)

Given the log-likelihood of a hypothesis h :

$$\log L(h) = \sum_{i=1}^{m} \log P(y_i|x_i, h) \tag{6}$$

And the probability of an example x belonging to class K as being :

$$P(K|x) = 1 - \sum_{i=1}^{K-1} h^i(x) \tag{7}$$

We can derive the log likelihood for a set of hypotheses and a given data set D as:

$$\log L(h) = \sum_{i=1}^{m} \sum_{j=1}^{K} \log \left(1 - \sum_{l=1}^{K-1} h^l(x_i)\right) \tag{8}$$

B)

C)

Question 4

A)

	Folds				
	1	2	3	4	5
log L Train	-2.103	-2.077	-2.031	-1.979	-2.033
log L Test	-2.064	-2.090	-2.016	-1.885	-2.129
Training Accuracy	0.81%	0.81%	0.83%	0.81%	0.80%
Testing Accuracy	0.69%	0.69%	0.72%	0.87%	0.71%

B)

C)

	Folds									
	1		2		3		4		5	
log L Train	-0.231	-1.428	-0.174	-1.435	-0.222	-1.560	-0.253	-1.434	-0.223	-1.371
log L Test	-0.410	-1.418	-0.497	-1.627	-0.547	-1.701	-0.549	-1.807	-0.456	-1.664
Training Accuracy	0.82%		0.84%		0.83%		0.81%		0.83%	
Testing Accuracy	0.72%		0.67%		0.69%		0.85%		0.74%	

D)

E)