1. Short Answer:

1. The time complexity of the test-time evaluation is O(M\*d\*N).

2. The model with k=10 is expected to have a “smoother” decision boundry.

3. (w,b) pair could be ((1,1),0.5), in which (1,1)=(w1,w2).

4. The largest singular value of A is the square root of the largest eigenvalue of A⊤A.

5. The naïve bayes assumption sometimes failed for text analysis or classification in the field of NLP since sometimes the meaning of the words are related in the context and they may appear together for certain grammar structure.

2. Linear Regression:

1. Since X is full-rank, which means it’s rows are linearly independent of each other. Because its column belongs to R, which is all the real numbers, there exists at least one solution of the linear equations in X. Also, since d>n, which means there might exist infinite solutions for this system. Thus, at least one w exists such that , which is the empirical risk with squared loss is zero.
2. Since rank(X) = n and the rank of a matrix is equal to the number of non-zero singular values, the rank of Σ is thus n. This is because Σ contains the singular values of X, and only the first n of these are non-zero (corresponding to the rank of X).
3. Since is an n\*n matrix and rank(X) = n, rank() <= n since the rank of a product of matrices is at most the minimum rank of the individual matrices. However, as the rows of X are linearly independent, ’s rows are also linearly independent. Thus, it’s full rank, which means rank() = n. Hence, since the determinant of is non-zero and it’s a square matrix of full rank, it is thus invertible.

3. SVM:

1. Theoretically, the minimum number of support vectors in a linearly separable binary classification problem should be d+1.

2. The smallest possible number of support vectors is 3 because each non-zero alpha indicates the corresponding xi is a support vector. The largest possible number is the number of the dataset because it’s possible for a point to contribute to another optimal solution which means it’s greater than 0. Thus, this could be the whole dataset which means the largest possible number of the support vectors is the number of the points in D.

1. a. We can expand the kernel function here:

k(x,z)==

=

= ,

b. For x=(-1,-1): =[1,1,-];

for x=(1,1): =[1,1,];

for x=(-1,1): =[1,1,-];

for x=(1,-1): =[1,-1,].

We can see that the sign of variates depends on both x1 and x2, which we need to heavily weights on since we need to make sure has the correct outputs with the XNOR. And we also need a bias term adjusted to ensure = 1 for XNOR true cases and 0 for false cases. Thus, we can get a solution that w=[0,0,0,0,C,b], which C accounts for term and b accounts for the slightly adjustment of . Thus, both C and b should be greater than 0 for XNOR and one solution might be w=[0,0,0,0,1,0.5].

1. Gaussian Naïve Bayes:
2. Since , .

From Bayes Theorem, we have that = .

Since ,

= .

Now, define A= and B=

And insert A and B into the equation and it becomes: .

1. Assumptions and definitions:

-Assume all features in the vector x are conditionally independent given the class label y (Naive bayes assumption).

-Assume the feature follow a Gaussian distribution within each class

-The covariance matrix is the identity matrix I implies that all features are considered to have equal variance and are uncorrelated across both classes

-Assume and

Prove:

.

Thus, and .

Thus, , which can be simplified to , and which is thus the form of , in which w= is the weight vector and b= is the bias term.

A graph with blue dots and red line

Description automatically generated