A Simple Introduction of Vapnik-Chervonenkis Dimension *

Jingfang Liu (15320171151900)& Xingyue Yu (15320171151888) †
Department of Economics, School of Economics, Xiamen University

March 17, 2019

1 Keywords

VC-Dimension/Classification/Linear Classifier/
Nonlinear Classifier/General Position/Shatter/Training/Test Data

2 Contents

Two parts

*I thank Muse for valuable comments. All errors are, of course, my own.
†Email: 675183025@qq.com

3 Introduction

1. Classification Problem

(There is two basic concepts you should know: Training/test data and Training/ test error. You can learn the concepts from our teacher class or you can learn from web https://jiamingmao.github.io/data-analysis/assets/Lectures/Foundations_of_Statistical_Learning.pdf Suppose there is 2-dimension training data which can be divided into two classes: Class1 and Class-1. We can find it has functions that classify these data in two classes. The function can be linear classifier or nonlinear classifier.

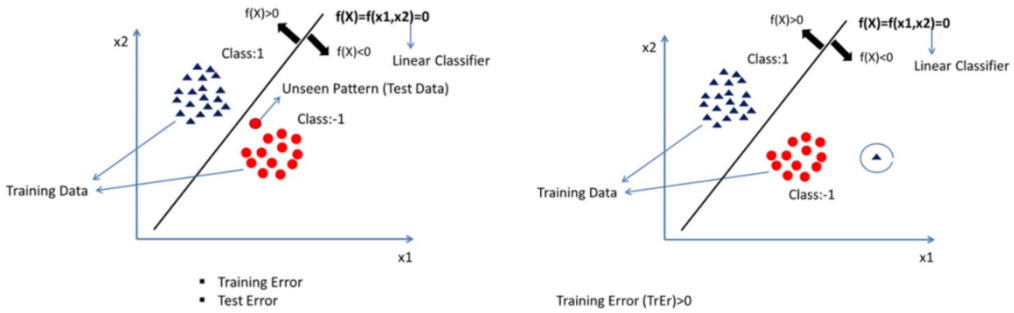


Figure 1: Classification

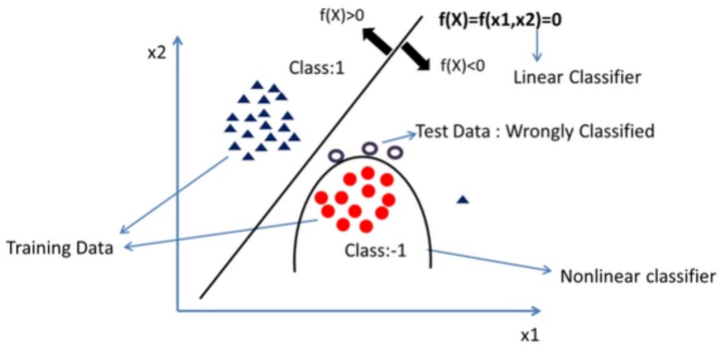


Figure 2: Classification

We can see from the Figure 1-2, though the nonlinear classifier can classify the data more precisely, the performance of classifier the test data will be bad, the test data may be wrongly classified. Comparing with the nonlinear classifier, the performance of linear classifier is better than nonlinear classifier.

2. The Relation Between Test and Training Error

1). Hoeffding's inequality In probability theory, Hoeffding's inequality provides an upper bound on the probability that the sum of bounded independent random variables deviates from its expected value by more than a certain amount

$$P(|\bar{X} - E(\bar{X})| \geqslant \sigma) \leqslant \exp(-2\delta^2 n^2) \tag{1}$$
$$\bar{X} = \frac{x1 + x2 + x3... + xn}{n}$$

2).Out of sample error:

$$E_{out}(h) = \underset{x\sim p}{\varepsilon} [h(x) \neq f(x)] \tag{2}$$

In-sample error:

$$Ein(h) = \frac{1}{N} \sum [h(x_n) \neq y_n] \tag{3}$$

According to Hoeffding's inequality:

$$P(Ein(h) - Eout(h) > \varepsilon) \leqslant \exp(-2\varepsilon^2 N) \tag{4}$$

For given h ,given large enough N,

$$Ein \approx Eout \tag{5}$$

3). VitalitemsbehindE_{in}andE_{out}error

(1) Two conditions to feasible learning

When the hypothesis space include M hypothesis, for the whole hypothesis space,we can infer

$P\ (|Ein(h1) - Eout(h1)| > \varepsilon \cup |Ein(h2) - Eout(h2)| > \varepsilon \cup .....|Ein(hm) - Eout(hm)| > \varepsilon)$
$\leqslant P(|Ein(h1) - Eout(h1)| > \varepsilon) + P(|Ein(h2) - Eout(h2)| > \varepsilon) + ...P(|Ein(hm) -$

$Eout(hm)| > \varepsilon) \leqslant 2M \exp(-2\varepsilon^2 N)$

For any given hypothesis h,the difference between $Ein(h)$ and$Eout(h)$ is bounded by $2M \exp(-2\varepsilon^2 N)$

There is two conditions for learning

(i). Suppose the space of hypothesis size M is finite, when the sample is large enough , that is the number of data N is large enough, thus for any arbitrary hypothesis in H, hypothesis space, out of sample error is approximately close to in sample error

(ii). If we find optimal function g(x) by loss function, $Ein(g)$ is close to 0, thus, $Eout(g)$ is approximately to 0.

The two core conditions for learning also correspond to the two processes of test and train, the process of train expect Ein(g) is small enough. The test process in population want to reach a result that the expected error is as small as possible.

(2) The vital role of M in two learning conditions:

(a) Trade-off on M:

i.. Can we make sure the $Eout(g)$ is close enough to $Ein(g)$?

ii. Can we make $Ein(g)$? small enough?

The size of M is vital, when M is small, N is large enough, Ein and Eout is close ,however, we might find it difficult to search for g(x) which makes Ein(g)is close to 0. When the M is small, thus the second condition cannot be satisfied. When M is large ,it is easy to find g(x) that makes $Ein(g)$?equals to 0 approximately, but the first term cannot satisfy.

(b) The size of M

M could be infinite for hypothesis space, we might find a finite factor mH to replace M in the inequality bound.

Establish a finite quantity that replaces M:

$$P\left[|Ein(h) - Eout(h)| > \varepsilon\right] \leq 2mH \exp(-2\varepsilon^2 N) \tag{6}$$

In the derivation above ,we used

$$P(h1 \cup h2 \cup h3... \cup hm) \leqslant P(h1) + P(h2) + ...P(hm) \tag{7}$$

In fact, each h hypothesis is not completely independent of each other, they have a lot of overlap, which means that among the M hypotheses, there may be some hypotheses that can be put into the same category. Thus,we can assume that the space size M is large, but on the sample set D, the number of valid hypothesis functions is limited.

(3) Effective number of Hypotheses

We might choose any function h from hypothesis space H, let this function classify the sample set D, for example, we might split two points use one straight line, which turns out (1,-1), (1,1), (-1,-1), (-1,1).We define each outcome as dichotomy. Effective number of hypotheses is defined as effective(N)=the number of dichotomies which the H split the sample set D. thus we might use effective (N) to replace M:

$$P\left[|Ein(h) - Eout(h)| > \varepsilon\right] \leq 2effictive(N) \exp(-2\varepsilon^2 N) \tag{8}$$

(4) Growth Function: The number of dichotomies which the H split D is related to hypothesis space and the sample point N,the number is labeled as is growth function which is related to sample point N.

growth function: remove dependence by taking max of all possible (x1,x2,. . . xn)

$$mH(N) = \max_{x1,x2,.....xn \in \chi} |H(x1, x2, ..., xn)| \tag{9}$$

Thus we lead to

$$P\left[|Ein(h) - Eout(h)| > \varepsilon\right] \leq 2mH(N)\exp(-2\varepsilon^2 N)| \tag{10}$$

however, this inequality is problematic because the possible value of$Ein(h)$is finite and the possible value of $Eout(h)$is infinite.

We use mathematical methods to derive ,finally we lead to VC bound:

$$P\left[\exists h \in Hs.t|Ein(h) - Eout(h)| > \varepsilon\right] \leqslant 4mH(2N)\exp(-\frac{1}{8}\varepsilon^2 N)| \tag{11}$$

That is to say for any function in H, $Ein(h)$is close to$Eout(h)$ if N is large enough.

3. Points to Remember

1). Always look for test error along with training error

2). Improving on training error not always improves test error

3). Increase in machine capacity may result in poor test performance

4). It is difficult to estimate true test error of a classifier

4 VC-Dimension

1. Basic Concepts Concept

1). Points in general position Statement: in a n-dimensional feature space a set of m points(m>n)is in general position if and only if no subset of (n+1) points lie on (n-1) dimensional hyperplane. Explanation:

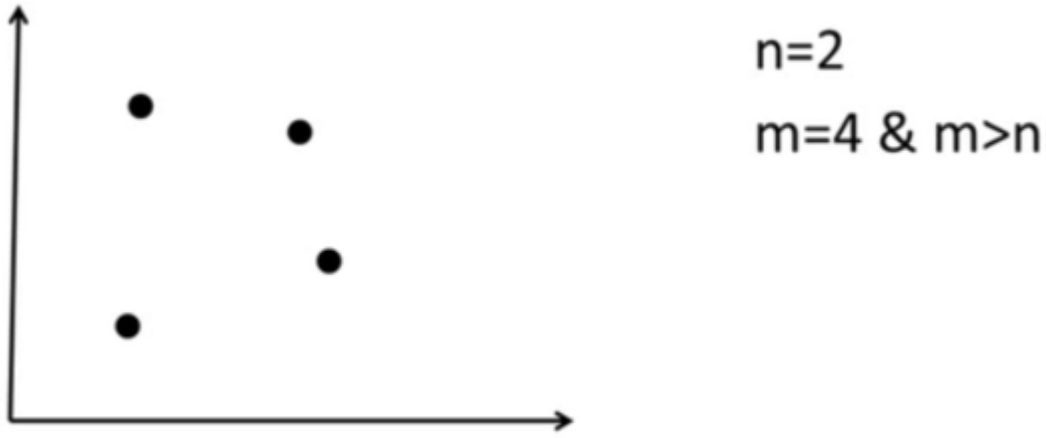


Figure 3: Points in general position

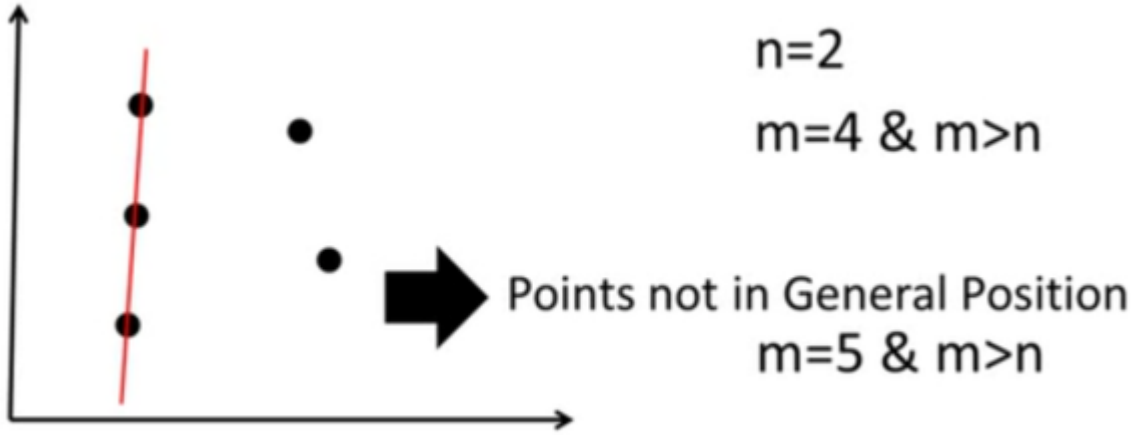
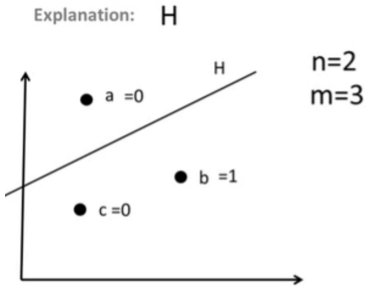


Figure 4: Points not in general position

From figure 3, we can see it satisfied the condition of statement, i.e. the n+1=3 is not on the straight line. So we can say that three point in a general position. While from Figure 4, it is not satisfied the statement because there are three points in a straight line. We say that points not in general position.

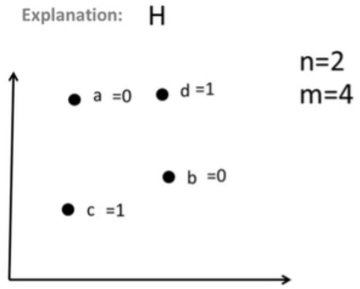
2). Shattering Statement: a hypothesis (H) shatters m points in n-dimensional space if all possible combinations of m points in n-dimensional spaces are correctly classified by H. Explanation From Figure 5, when m=3, there is 8 possible arrangements can be set. And we can see that every possible arrangements can be classified by a straight line(H is linear model). While from Figure 6 when there are four points in 2-dimension space, not all arrangements can be classified by a straight line. We say that four points in 2-dimension space are not shattered by straight line."



2^m possible arrangements

a	b	c
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

Figure 5: Shattering



2^m possible arrangements

a	b	c
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

Figure 6: Not Shattering

2. The definition of VC Dimension

Cardinality of the largest set of points that Hypothesis can shatter.

1). The VC dimension of H, denoted$d_{vc}(H)$, is the size of the largest data set that H can shatter.

2). $d_{vc}(H)$ is the largest value of N for which $m_H(N) = 2^N$

3). If arbitrarily large finite sets can be shattered by H, then$d_{vc}(H) = \infty$

4). $\exists$ some shattered set of size $d \rightarrow d_{vc}(H) \geq d$ No set of size d+1 is shatterd $\rightarrow$ $d_{vc}(H) \leq d$

2. Linear/Nonlinear

VC dimension of linear classifier: (n+1) {points should be in general position}

VC dimension of Nonlinear Classifier: very difficult to compute

Some empirical methods to compute VC dimension of nonlinear classifier are suggested by Vapnik et al.(http://yann.lecun.com/exdb/publis/pdf/vapnil-levin-lecun-94.pdf)

3. Points to Remember

1). VC dimension is directly related to Machine/Hypothesis Capacity

2). For a given training set size and training error VC dimension gives probabilistic upper bound of test error

3). VC dimension is a cardinality of the largest set of points that the Machine/Hypothesis can shatter.

4). For good generalization(less test error) VC dimension of a Machine/Hypothesis for asymptotical solutions. For small VC dimension, small training set may lead to good generalization.

5 Acknowledgement

Part of this notes is adapted from the following sources:

1. A lecture called A simple and gentle tutorial on Vapnik-Chervonenkis dimension presented by Himanshu Pant.

2. Http://www.svms.org/vc-dimension/

3. Kearns, M. J., & Vazirani, U. V. (1994). An introduction to computational learning theory. Cambridge, MA, USA: MIT Press.

4. Http://freemind.pluskid.org/slt/vc-theory-vapnik-chervonenkis-dimension/

5. Https://cs.nyu.edu/~yann/talks/lecun-ranzato-icml2013.pdf

6. Https://en.wikipedia.org/wiki/Vapnik–Chervonenkis_theory

7. Https://jiamingmao.github.io/data-analysis/assets/Lectures/