Neural Network Learning: Theoretical Foundations

Chap.8, 9, 10, 11

Martin Anthony and Peter L. Bartlett

2017.08.14

Contents

8. Vapnik-Chervonenkis Dimension Bounds for Neural Networks

Part 2: Pattern Classification with Real-Output Networks
9. Classification with Real-Valued Functions

10. Covering Numbers and Uniform Convergence

11. The Pseudo-Dimension and Fat-Shattering Dimension

Contents

8. Vapnik-Chervonenkis Dimension Bounds for Neural Networks

Part 2: Pattern Classification with Real-Output Networks
9. Classification with Real-Valued Functions

10. Covering Numbers and Uniform Convergence

11. The Pseudo-Dimension and Fat-Shattering Dimension

Reviews

- Definition 7.5 Let G be a set of real-valued functions defined on $\mathbb{R}^d$. We say that G has solution set components bound B if for any $1 \leq k \leq d$ and any $\{f_1, \dots f_k\} \subseteq G$ that has regular zero-set intersetions, we have

$$\mathrm{CC}\Big(\bigcap_{i=1}^{k} \{a \in \mathbb{R}^d : f_i(a) = 0\}\Big) \leq B.$$

- Theorem 7.6 Suppose that F is a class of real-valued functions defined on $\mathbb{R}^d \times X$, and that H is a k-combination of $\mathrm{sgn}(F)$. If F is closed under addition of constants, has solution set components bound B, and functions in F are C^d in their parameters, then

$$\Pi_H(m) \leq B \sum_{i=0}^{d} \binom{mk}{i} \leq B \Big(\frac{emk}{d}\Big)^d,$$

for $m \geq d/k$.

8.2 Function Classes that are Polynomial in their Parameters

- Consider classes of functions that can be expressed as boolean combinations of thresholded real-valued functions, each of which is polynomial in its parameters.
- Lemma 8.1 Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is a polynomial of degree l. Then the number of connected components of $\{a \in \mathbb{R}^d : f(a) = 0\}$ is no more than $l^{d-1}(l+2)$.
- Corollary 8.2 For $l \in \mathbb{N}$, the set of degree l polynomials defined on $\mathbb{R}^d$ has solution set components bound $B = 2(2l)^d$.

- Theorem 8.3 Let F be a class of functions mapping from $\mathbb{R}^d \times X$ to $\mathbb{R}$ so that, for all $x \in X$ and $f \in F$, the function $a \mapsto f(a,x)$ is a polynomial on $\mathbb{R}^d$ of degree no more than l. Suppose that H is a k-combination of $\mathrm{sgn}(F)$. Then if $m \geq d/k$,

$$\Pi_H(m) \leq 2\left(\frac{2emkl}{d}\right)^d,$$

and hence $\mathrm{VCdim}(H) \leq 2d\log_2(12kl)$.

- Theorem 8.4 Suppose h is a function from $\mathbb{R}^d \times \mathbb{R}^n$ to $\{0,1\}$ and let

$$H = \{x \mapsto h(a,x) : a \in \mathbb{R}^d\}$$

  be the class determined by h. Suppose that h can be compoted by an algorithm that takes as input the pair $(a,x) \in \mathbb{R}^d \times \mathbb{R}^n$ and returns $h(a,x)$ after no more than t operations of the following types:
    - the arithmetic operations $+, -, \times$, and $/$ on real numbers,
    - jumps conditioned on $>, \geq, <, \leq, =$, and $\neq$ comparisions of real numbers, and
    - output 0 or 1.

  Then $\mathrm{VCdim}(H) \leq 4d(t+2)$.

- Theorem 8.5 For all $d, t \geq 1$, there is a class H of functions, parametrized by d real numbers, that can be computed in time $O(t)$ using the model of computation defined in Thoerem 8.4, and that has $\mathrm{VCdim}(H) \geq dt$.

8.3 Piecewise-Polynomial Networks

- Theorem 8.6 Suppose N is a feed-forward linear threshold network with a total of W weights, and let H be the class of functions computed by this network. Then $\text{VCdim}(H) = O(W^2)$.

- This theorem can easily be generalized to network with piecewise-polynomial activation functions. A piecewise-polynomial function $f : \mathbb{R} \to \mathbb{R}$ can be written as $f(\alpha) = \sum_{i=1}^{p} 1_{A(i)}(\alpha) f_i(\alpha)$, where $A(1), \ldots, A(p)$ are disjoint real intervals whose union is $\mathbb{R}$, and $f_1, \ldots, f_p$ are polynomials. Define the degree of f as the largest degree of the polynomials f_i.

- Theorem 8.7 Suppose N is a feed-forward network with a total of W weights and k computation units, in which the output unit is a linear threshold unit and every other computation unit has a piecewise-polynomial activation function with p pieces and degree no more than l. Then, if H is the class of functions computed by N, $\text{VCdim}(H) = O(W(W + kl \log_2 p))$.

- Theorem 8.8 Suppose N is a feed-forward network of the form described in Theorem 8.7, with W weights, k computation units, and all non-output units having piecewise-polynomial activation functions with p pieces and degree no more than l. Suppose in addition that the computation units in the network are arranged in L layers, so that each unit has connections only from units in earlier layers. Then if H is the class of functions computed by N,

$$\Pi_H(m) \leq 2^L (2emkp(l+1)^{L-1})^{WL},$$

and

$$\mathrm{VCdim}(H) \leq 2WL\log_2(4WLpk/\ln 2) + 2WL^2\log_2(l+1) + 2L.$$

For fixed p, l, $\mathrm{VCdim}(H) = O(WL\log_2 W + WL^2)$.

- Theorem 8.9 Suppose $s : \mathbb{R} \to \mathbb{R}$ has the following properties:
  1. $\lim_{\alpha \to \infty} s(\alpha) = 1$ and $\lim_{\alpha \to -\infty} s(\alpha) = 0$, and
  2. s is differentiable at some point $\alpha_0 \in \mathbb{R}$, with $s'(\alpha_0) \neq 0$.

  For any $L \geq 1$ and $W \geq 10L - 14$, there is a feed-forward network with L layers and a total of W parameters, where every computation unit but the output unit has activation function s, the output unit being a linear threshold unit, and for which the set H of functions computed by the network has

$$\mathrm{VCdim}(H) \geq \left\lfloor \frac{L}{2} \right\rfloor \left\lfloor \frac{W}{2} \right\rfloor$$

8.4 Standard Sigmoid Networks
Discrete inputs and bounded fan-in

- Consider networks with the standard sigmoid activation, $\sigma(\alpha) = 1/(1 + e^{-\alpha})$.
- We define the fan-in of a computation unit to be the number of input units or computation units that feed into it.
- Theorem 8.11 Consider a two-layer feed-forward network with input domain $X = \{-D, -D + 1, \ldots, D\}^n$ (for $D \in \mathbb{N}$) and k first-layer computation units, each with the standard sigmoid activation function. Let W be the total number of parameters in the network, and suppose that the fan-in of each first-layer unit is no more than N. Then the class H of functions computed by this network has $\mathrm{VCdim}(H) \leq 2W \log_2(60ND)$.

- Theorem 8.12 Consider a two-layer feed-forward linear threshold network that has W parameters and whose first-layer units have fan-in no more than N. If H is the set of functions computed by this network on binary inputs, then $\mathrm{VCdim}(H) \le 2W \log_2(60N)$. Furthermore, there is a constant c s.t. for all W there is a network with W parameters that has $\mathrm{VCdim}(H) \ge cW$.

General standard sigmoid networks

- Theorem 8.13 Let H be the set of functions computed by a feed-forward network with W parameters and k computation units, in which each computation unit other than the output unit has the standard sigmoid activation function (the output unit being a linear threshold unit). Then

  $$\Pi_H(m) \leq 2^{(Wk)^2/2}(18Wk^2)^{5Wk}\left(\frac{em}{W}\right)^W$$

  probided $m \geq W$, and

  $$\mathrm{VCdim}(H) \leq (Wk)^2 + 11Wk\log_2(18Wk^2).$$

- Theorem 8.14 Let h be a function from $\mathbb{R}^d \times \mathbb{R}^n$ to $\{0,1\}$, determining the class

$$H = \{x \mapsto h(a,x) : a \in \mathbb{R}^d\}.$$

Suppose that h can be computed by an algorithm that takes as input the pair $(a,x) \in \mathbb{R}^d \times \mathbb{R}^n$ and returns $h(a,x)$ after no more than t of the following oprations:

  - the exponential function $\alpha \mapsto e^{\alpha}$ on real numbers,
  - the arithmetic operations $+, -, \times$, and $/$ on real numbers,
  - jumps conditioned on $>, \geq, <, \leq, =$, and $\neq$ comparisions of real numbers, and
  - output 0 or 1.

Then $\mathrm{VCdim}(H) \leq t^2 d(d + 19\log 2(9d))$. Furthermore, if the t steps include no more than q in which the exponential function is evaluated, then

$$\Pi_H(m) \leq 2^{(d(q+1))^2/2}(9d(q+1)2^t)^{5d(q+1)}\left(\frac{em(2^t-2)}{d}\right)^d,$$

and hence $\mathrm{VCdim}(H) \leq (d(q+1))^2 + 11d(q+1)(t + \log_2(9d(q+1)))$.

Proof of VC-dimension bounds for sigmoid networks and algorithms

- Lemma 8.15 Let $f_1, \ldots, f_q$ be fixed affine functions of $a_1, \ldots, a_d$, and let G be the class of polynomials in $a_1, \ldots, a_d, e^{f_1(a)}, \ldots, e^{f_q(a)}$ of degree no more than l. Then G has solution set components bound

$$B = 2^{q(q-1)/2}(l+1)^{2d+q}(d+1)^{d+2q}.$$

- Lemma 8.16 Suppose G is the class of functions defined on $\mathbb{R}^d$ computed by a circuit satisfying the following conditions: the circuit contains q gates, the output gate computes a rational function of degree no more than $l \geq 1$, each non-output gate computes the exponential function of a rational function of degree no more than l, and the denominator of each rational function is never zero. Then G has solution set components bound $2^{(qd)^2/2}(9qdl)^{5qd}$.

Contents

8. Vapnik-Chervonenkis Dimension Bounds for Neural Networks

Part 2: Pattern Classification with Real-Output Networks
9. Classification with Real-Valued Functions

10. Covering Numbers and Uniform Convergence

11. The Pseudo-Dimension and Fat-Shattering Dimension

9.2 Large Margin Classifiers

- Suppose F is a class of functions defined on the set X and mapping to the interval $[0, 1]$.

- Definition 9.1 Let $Z = X \times \{0, 1\}$. If f is a real-valued function in F, the margin of f on $(x, y) \in Z$ is

$$\text{margin}(f(x), y) = \begin{cases} f(x) - 1/2 & \text{if } y = 1 \\ 1/2 - f(x) & \text{otherwise.} \end{cases}$$

  Suppose γ is a nonnegative real number and P is a probability distribution on Z. We define the error $er_P^\gamma(f)$ of f w.r.t. P and γ as the probability

$$er_P^\gamma(f) = P\{\text{margin}(f(x), y) < \gamma\},$$

  and the misclassification probability of f as

$$er_P(f) = P\{\text{sgn}(f(x) - 1/2) \neq y\}.$$

- Definition 9.2 A classification learning algorithm L for F takes as input a margin parameter $\gamma > 0$ and a sample $z \in \bigcup_{i=1}^{\infty} Z^i$, and returns a function in F s.t., for any $\epsilon, \delta \in (0,1)$ and any $\gamma > 0$, there is an integer $m_0(\epsilon, \delta, \gamma)$ s.t. if $m \geq m_0(\epsilon, \delta, \gamma)$ then, for any probability distribution P on $Z = X \times \{0,1\}$,

$$P^m\left\{er_P(L(\gamma, z)) < \inf_{g \in F} er_P^{\gamma}(g) + \epsilon\right\} \geq 1 - \delta.$$

- Sample error of f w.r.t. γ on the sample z :

$$\hat{er}_z^{\gamma}(f) = \frac{1}{m}|\{i : \text{margin}(f(x_i), y_i) < \gamma\}|$$

- Proposition 9.3 For any function $f : X \to \mathbb{R}$ and any sequence of labelled examples $((x_1, y_1), \ldots, (x_m, y_m))$ in $(X \times \{0,1\})^m$, if

$$\frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2 < \epsilon$$

then

$$\hat{er}_z^{\gamma}(f) < \epsilon/(1/2 - \gamma)^2$$

for all $0 \leq \gamma < 1/2$.

Contents

8. Vapnik-Chervonenkis Dimension Bounds for Neural Networks

Part 2: Pattern Classification with Real-Output Networks
9. Classification with Real-Valued Functions

10. Covering Numbers and Uniform Convergence

11. The Pseudo-Dimension and Fat-Shattering Dimension

10.2 Covering Numbers

- Recall that the growth function

$$\Pi_H(m) = max\{\left|H_{|S}\right| : S \subseteq X \ and \ |S| = m\}.$$

- Since H maps into $\{0,1\}$, $\left|H_{|S}\right|$ is finite for every finte S. Howevere, if F is a class of real-valued functions, $\left|F_{|S}\right|$ may be infinite.
- Use the notion of covers to measure the 'extent' of $F_{|S}$

10.2 Covering Numbers

- Covering numbers for subsets of Euclidean space

  - Definition Given $W \subseteq \mathbb{R}^k$ and a positive real number ϵ, we say that $C \subseteq \mathbb{R}^k$ is a d_∞ ϵ - cover for W if $C \subseteq W$ and for every $w \in W$ there is a $v \in C$ such that

$$max\{|w_i - v_i| : i = 1, \ldots, k\} < \epsilon$$

  - Definition We could also define an ϵ-cover for W$\subseteq \mathbb{R}^k$ as a subset C of W for which W is contained in the union of the set of open d_∞ ball of radius ϵ centred at the points in C.

  - Definition The d_∞ ϵ-covering number of W, $\mathcal{N}(\epsilon, W, d_\infty)$, to be the minimum cardinality of a d_∞ ϵ-cover for W.

10.2 Covering Numbers

- Uniform covering numbers for a function class
  - Definition Suppose that F is a class of functions from X to $\mathbb{R}$. Given a sequence $x = (x_1, x_2, \ldots, x_k) \in X^k$, we let $F_{|x}$ be the subset of $\mathbb{R}^k$ given by

$$F_{|x} = \{(f(x_1), f(x_2), \ldots, f(x_k)) : f \in F\}$$

  - Definition For a positive number ϵ, we define the uniform covering number $\mathcal{N}_\infty(\epsilon, F, k)$ to be the maximum, over all $x \in X^k$, of the covering number $\mathcal{N}(\epsilon, F_{|x}, d_\infty)$ that is,

$$\mathcal{N}_\infty(\epsilon, F, k) = max\{\mathcal{N}(\epsilon, F_{|x}, d_\infty) : x \in X^k\}$$

  - The uniform covering number is a generalization of the growth function. Suppose that functions in H map into $\{0, 1\}$. Then for all $x \in X^k, H_{|x}$ is finite and, for all $x \in X^k, H_{|x}$ is finite and, for all $\epsilon < 1$, $\mathcal{N}(\epsilon, F_{|x}, d_\infty) : x \in X^k = \left|H_{|x}\right|$, so $\mathcal{N}_\infty(\epsilon, F, k) = \Pi_H(m)$

10.3 A Uniform Convergence Results

- Theorem 10.1 Suppose that F is a set of real-valued functions defined on the domain X. Let P be any probability distribution on Z = X $\times\{0,1\}, \epsilon$ any real number between 0 and 1, γ any positive real number, and m any positive integer. Then,

$$P^m\{er_p(f) \geq \hat{er}_z^{\gamma}(f) + \epsilon \text{ for some } f \text{ in } F\} \leq 2\mathcal{N}_\infty(\gamma/2, F, 2m)exp(-\epsilon^2 m/8)$$

10.3 A Uniform Convergence Results

- Symmetrization : bound the desired probability in terms of the probability of an event based on two samples.

- Lemma 10.2 With the notation as above, let

$$Q = \{z \in Z^m : \textit{some } f \textit{ in } F \textit{ has } er_P(f) \geq \hat{er}_z^{\gamma}(f) + \epsilon\}$$

  and

$$R = \{(r,s) \in Z^m \times Z^m : \textit{some } f \textit{ in } F \textit{ has } \hat{er}_s(f) \geq \hat{er}_r^{\gamma}(f) + \epsilon/2\}$$

  Then for $m \geq 2/\epsilon^2$,

$$P^m(Q) \leq 2P^{2m}(R)$$

  .

10.3 A Uniform Convergence Results

- Permutations : involving a set of permutations on the labels of th double sample.
- Let Γ_m be the set of all permutations of $\{1, 2, \ldots, 2m\}$ taht swap i and m+i. For instance, $\sigma \in \Gamma_3$ might give

$$\sigma(z_1, z_2, \ldots, z_6) = (z_1, z_5, z_6, z_4, z_2, z_3).$$

- Using Lemma 4.5 we can get

$$P^{2m}(R) = \mathbb{E} Pr(\sigma z \in R) \leq \max_{z \in Z^{2m}} Pr(\sigma z \in R)$$

10.3 A Uniform Convergence Results

- Lemma 10.3 For the set $R \subseteq Z^{2m}$ defined in Lemma 10.2, and for a permutation σ chosen uniformly at random from γ_m

$$\max_{z \in Z^{2m}} Pr(\sigma z \in R) \leq \mathcal{N}_\infty(\gamma/2, F, 2m) exp(-\epsilon^2 m/8)$$

- (proof) Fix a minimal $\gamma/2$-cover T of $F_{|x}$. Then for all f in F there is an $\hat{f}$ in T with $\left| f(x_i) - \hat{f}_i \right| < \gamma/2 \text{ for } 1 \leq i \leq 2m$. Define $v(\hat{f}, i) = I(margine(\hat{f}_i, y_i) < \gamma/2)$ and use Hoeffding's inequality.

10.3 A Uniform Convergence Results

- When the set $\{f(x) : f \in F\} \subset \mathbb{R}$ is unbounded, then $\mathcal{N}_\infty(\gamma/2, F, 1) = \infty$ for all $\gamma > 0$
- Consider $\pi_\gamma : \mathbb{R} \to [1/2 - \gamma, 1/2 + \gamma]$ satisfies

$$\pi_\gamma(\alpha) = \begin{cases} 1/2 + \gamma \text{ if } \alpha \geq 1/2 + \gamma \\ 1/2 - \gamma \text{ if } \alpha \leq 1/2 + \gamma \\ \alpha \text{ if } \textit{otherwise} \end{cases}$$

- Theorem 10.4 Suppose that F is a set of real-valued functions defined on a domain X. Let P be any probability distribution on Z = X $\times \{0, 1\}$, ϵ nay real number between 0 and 1, γ any positive real number, and m any positive integer. Then,

$$P^m\{er_p(f) \geq \hat{er}_z^\gamma(f) + \epsilon \textit{ for some } f \textit{ in } F\} \leq 2\mathcal{N}_\infty(\gamma/2, \pi_\gamma(F), 2m) exp(-\epsilon^2 m/8)$$

10.4 Covering Numbers in General

- Recall that a metric space consists of a set A together with a metric, d, a mapping from A $\times$ A to the nonnegative reals with the following properties, for all $x, y, x \in A$: (i) d(x,y) $=$ 0 if and only if x$=$y (ii) d(x,y)$=$d(y,x), and (iii) d(x,z) $\leq$ d(x,y)$+$d(y,z)
- As same way, we can define the ϵ-covering number of W, $\mathbb{N}(\epsilon, W, d)$, to be the minimum cardinality of an ϵ-cover for W with respect to the metric d.
- Lemma 10.5 For any class F of real-valued functions defined on X, any $\epsilon > 0$, and any $k \in \mathbb{N}$,
$$\mathcal{N}_1(\epsilon, F, k) \leq \mathcal{N}_2(\epsilon, F, k) \leq \mathcal{N}_\infty(\epsilon, F, k)$$

10.5 Remark

- Pseudo-metric : A pseudo-metric d satisfies the second and third conditions in the definition of a metric, but the first condition does not necessarily hold. Instead, d(x,y) $\geq$ for all x,y and d(x,x)=0, but we can have x$\neq$y and d(x,y)=0.
- Improper coverings : if (A, d) is a metric space and W$\subseteq$A, then, for $\epsilon > 0$, we say that C$\subseteq$A is an ϵ-cover of W if C$\subseteq$W and for every $w \in W$ there is a $v \in C$ such that $d(w, v) < \epsilon$. If we drop the requirement that $C \subseteq W$ then we say that C is an improper cover.
- Lemma 10.6 Suppose that W is a totally bounded subset of a metric space (A,d). For $\epsilon > 0$, let $\mathcal{N}'(\epsilon, W, d)$ be the minimum cardinality of a finite improper ϵ-cover for W. Then,

  $$\mathcal{N}(2\epsilon, W, d) \leq \mathcal{N}'(\epsilon, W, d) \leq \mathcal{N}(\epsilon, W, d)$$

  for all $\epsilon > 0$

Contents

8. Vapnik-Chervonenkis Dimension Bounds for Neural Networks

Part 2: Pattern Classification with Real-Output Networks
9. Classification with Real-Valued Functions

10. Covering Numbers and Uniform Convergence

11. The Pseudo-Dimension and Fat-Shattering Dimension

11.2 The Pseudo-Dimension

- Recall that a subset S=$\{x_1, x_2, \ldots, x_m\}$ of X is shattered by H if $H_{|S}$ has cardinality 2^m. This means that for any binary vector $b = (b_1, b_2, \ldots, b_m) \in \{0, 1\}^m$, there is some corresponding function h_b in H such that
$$(h_b(x_1), h_b(x_2), \ldots, h_b(x_m)) = b$$
- Definition 11.1 Let F be a set of functions mapping from a domain X to $\mathbb{R}$ and suppose that $S = \{x_1, x_2, \ldots, x_m\} \subseteq X$. Then S is pseudo-shattered by F if there are real number $r_1, r_2, \ldots, r_m$ such that for each $b \in \{0, 1\}^m$ there is a function f_b in F with $sgn(f_b(x_i) - r_i) = b_i$ for $1 \leq i \leq m$. We say that $r = (r_1, r_2, \ldots, r_m)$ witnesses the shattering.

11.2 The Pseudo-Dimension

- Definition 11.2 Suppose that F is a set of functions from a domain X to $\mathbb{R}$. Then F has pseudo-dimension d if d is the maximum cardinality of a subset S of X that is pseudo-shattered by F. I f no such maximum exists, we say that F has infinite pseudo=dimension. The pseudo-dimension of F is denoted Pdim(F).

11.2 The Pseudo-Dimension

- Theorem 11.3 Suppose F is a class of real-valued functions and $\sigma : \mathbb{R} \to \mathbb{R}$ is a non-decreasing function. Let $\sigma(F)$ donote the class $\{\sigma \circ f : f \in F\}$. Then $Pdim(\sigma(F)) \leq Pdim(F)$.
- Theorem 11.4 If F is a vector space of real-valued functions then Pdim(F)=dim(F)
- (proof) Use theorem 3.5 : H=$\{sgn(f+g) : f \in F\}$ Then VCdim(H)=dim(F) and Pdim(F) $=$ VCdim(B_F) where $B_F = \{(x, y) \mapsto sgn(f(x) - y) : f \in F\}$
- Corollary 11.5 If F is a subset of a vector space F' of real-valued functions then $Pdim(F) \leq dim(F')$

11.2 The Pseudo-Dimension

- Suppose that F is the class of affine combinations of n real inputs of the form

$$f(x) = w_0 + \sum_{i=1}^{n} w_i x_i,$$

where $w_i \in \mathbb{R}$ and $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ is the input pattern. We can think of F as the class of functions computable by a linear computation unit, which has the identity function as its activation function.

- Theorem 11.6 Let F be the class of real functions computable by a linear computation unit on $\mathbb{R}^n$. Then Pdim(F)=n+1.

- (proof) F is a vector space. B=$\{f_1, f_2, \ldots, f_n, 1\}$ is a basis of F where$f_i(x) = x_i$ and 1 denotes the identically-1 function.

- Theorem 11.7 Let F be the class of real functions computable by a linear computation unit on $\{0,1\}^n$. Then Pdim(F)=n+1

11.2 The Pseudo-Dimension

- Consider the class of polynomial transformations. A polynomial transformation of $\mathbb{R}^n$ is a function of the form

$$f(x) = w_0 + w_1\phi_1(x) + w_2\phi_2(x) + \ldots + w_l\phi_l(x)$$

  where $\phi_i(x) = \prod_{j=1}^{n} x_i^{r_{ij}}$ for some nonnegative integers r_{ij}

- The degree of ϕ_i is $r_{i1} + r_{i2} + \ldots + r_{in}$.

- for instance, the polynomial transformations of degree at most two on $\mathbb{R}^3$ are the functions of the form

$$f(x) = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_1^2 + w_5x_2^2 + w_6x_3^2 + w_7x_1x_2 + w_8x_1x_3 + w_9x_2x_3.$$

- Theorem 11.8 Let F be the class of all polynomial transformations on $\mathbb{R}^n$ of degree at most k. Then

$$Pdim(F) = \binom{n+k}{k}$$

11.2 The Pseudo-Dimension

- (proof) F is a vector space. Let [n] denote $\{1, 2, \ldots, n\}$ and denote by $[n]^k$ the set of all selections of at most k objects from [n] where repetition is allowd. $\phi^T(x) = \prod_{i \in T} x_i$ We can state that

$$f(x) = \sum_{T \in [n]^k} w_T \phi^T(x)$$

Define B(n,k)=$\{\phi^T : T \in [n]^k\}$ and show that this set is linearly independent.

- Theorem 11.9 Let F be the class of all polynomial transformations on $\{0, 1\}^n$ of degree at most k. Then,

$$Pdim(F) = \sum_{i=0}^{k} \binom{n}{i}.$$

11.3 The Fat-Shattering Dimension

- Definition 11.10 Let F be a set of functions mapping from a domain X to $\mathbb{R}$ and suppose that $S = \{x_1, x_2, \ldots, x_m\} \subseteq X$. Suppose also that γ is a positive real number. Then S is γ-shattered by F if there are real numbers $r_1, r_2, \ldots, r_m$ such that for each $b \in \{0,1\}^m$ there is a function f_b in F with

$$f_b(x_i) \geq r_i + \gamma \ if \ b_i = 1, \ and \ f_b(x_i) \leq r_i - \gamma \ if \ b_i = 0, \ for 1 \leq i \leq m.$$

- Definition 11.11 Suppose that F is a set of functions from a domain X to $\mathbb{R}$ and that $\gamma > 0$. Then F has γ-dimension d if d is the maximum cardinality of a subset S of X that is γ-shattered by F. If no such maximum exists, we say that F has infinite γ-dimension. The γ-dimension of F is denoted $fat_F(\gamma)$.

11.3 The Fat-Shattering Dimension

- $f : [0,1] \to \mathbb{R}$ is of bounded variation if there is V such that for every integer n and every sequence $y_1, y_2, \ldots, y_n$ of numbers with $0 \leq y_1 < y_2 < \ldots < y_n \leq 1$, we have

$$\sum_{i=1}^{n-1} |f(y_{i+1}) - f(y_i)| \leq V$$

In this case, we say that f has total variation at most V.

- Theorem 11.12 Let F be the set of all functions mapping from the interval [0,1] to the interval [0,1] and having total variation at most V. Then,

$$fat_F(\gamma) = 1 + \left\lfloor \frac{V}{2\gamma} \right\rfloor$$

11.3 The Fat-Shattering Dimension

- Theorem 11.13 Suppose that F is a set of real-valued functions. Then,
  (i) For all $\gamma > 0$, $fat_F(\gamma) \le Pdim(F)$.
  (ii) If a finite set S is pseudo-shattered then there is γ_0 such that for all $\gamma < \gamma_0$, S is γ-shattered.
  (iii) The function fat_F is non-increasing with γ
  (iv) $Pdim(F) = \lim_{\gamma \downarrow 0} fat_F(\gamma)$ (where both sides may be infinite).

- Theorem 11.14 Suppose that a set F of real-valued functions is closed under scalar multiplication. Then, for all positive γ,

  $$fat_F(\gamma) = Pdim(F).$$

  In particular, F has finite fat-shattering dimension if and only if it has finite pseudo-dimension.