Title: Uniform convergence in probability

URL: https://en.wikipedia.org/wiki/Uniform_convergence_in_probability

PageID: 22999791

Categories: Category:Combinatorics, Category:Machine learning, Category:Theorems in probability theory



-----

Uniform convergence in probability is a form of convergence in probability in statistical asymptotic theory and probability theory . It means that, under certain conditions, the empirical frequencies of all events in a certain event-family converge to their theoretical probabilities . Uniform convergence in probability has applications to statistics as well as machine learning as part of statistical learning theory .

The law of large numbers says that, for each single event A {\displaystyle A} , its empirical frequency in a sequence of independent trials converges (with high probability) to its theoretical probability. In many application however, the need arises to judge simultaneously the probabilities of events of an entire class S {\displaystyle S} from one and the same sample. Moreover it, is required that the relative frequency of the events converge to the probability uniformly over the entire class of events S {\displaystyle S} [ 1 ] The Uniform Convergence Theorem gives a sufficient condition for this convergence to hold. Roughly, if the event-family is sufficiently simple (its VC dimension is sufficiently small) then uniform convergence holds.

## Definitions

For a class of predicates H {\displaystyle H} defined on a set X {\displaystyle X} and a set of samples x = ( x 1 , x 2 , … , x m ) {\displaystyle x=(x_{1},x_{2},\dots ,x_{m})} , where x i ∈ X {\displaystyle x_{i}\in X} , the empirical frequency of h ∈ H {\displaystyle h\in H} on x {\displaystyle x} is

The theoretical probability of h ∈ H {\displaystyle h\in H} is defined as Q P ( h ) = P { y ∈ X : h ( y ) = 1 } . {\displaystyle Q_{P}(h)=P\{y\in X:h(y)=1\}.}

The Uniform Convergence Theorem states, roughly, that if H {\displaystyle H} is "simple" and we draw samples independently (with replacement) from X {\displaystyle X} according to any distribution P {\displaystyle P} , then with high probability , the empirical frequency will be close to its expected value , which is the theoretical probability. [ 2 ]

Here "simple" means that the Vapnik–Chervonenkis dimension of the class H {\displaystyle H} is small relative to the size of the sample. In other words, a sufficiently simple collection of functions behaves roughly the same on a small random sample as it does on the distribution as a whole.

The Uniform Convergence Theorem was first proved by Vapnik and Chervonenkis [ 1 ] using the concept of growth function .

## Uniform convergence theorem

The statement of the uniform convergence theorem is as follows: [ 3 ]

If H {\displaystyle H} is a set of { 0 , 1 } {\displaystyle \{0,1\}} -valued functions defined on a set X {\displaystyle X} and P {\displaystyle P} is a probability distribution on X {\displaystyle X} then for ε > 0 {\displaystyle \varepsilon >0} and m {\displaystyle m} a positive integer, we have:

And for any natural number m {\displaystyle m} , the shattering number Π H ( m ) {\displaystyle \Pi _{H}(m)} is defined as:

From the point of Learning Theory one can consider H {\displaystyle H} to be the Concept/Hypothesis class defined over the instance set X {\displaystyle X} . Before getting into the details of the proof of the theorem we will state Sauer's Lemma which we will need in our proof.

## Sauer–Shelah lemma

The Sauer–Shelah lemma [ 4 ] relates the shattering number Π h ( m ) {\displaystyle \Pi _{h}(m)} to the VC Dimension.

Lemma: Π H ( m ) ≤ ( e m d ) d {\displaystyle \Pi _{H}(m)\leq \left({\frac {em}{d}}\right)^{d}} , where d {\displaystyle d} is the VC Dimension of the concept class H {\displaystyle H} .

Corollary: Π H ( m ) ≤ m d {\displaystyle \Pi _{H}(m)\leq m^{d}} .

## Proof of uniform convergence theorem

[ 1 ] and [ 3 ] are the sources of the proof below. Before we get into the details of the proof of the Uniform Convergence Theorem we will present a high level overview of the proof.

Symmetrization: We transform the problem of analyzing | Q P ( h ) − Q ^ x ( h ) | ≥ ε {\displaystyle |Q_{P}(h)-{\widehat {Q}}_{x}(h)|\geq \varepsilon } into the problem of analyzing | Q ^ r ( h ) − Q ^ s ( h ) | ≥ ε / 2 {\displaystyle |{\widehat {Q}}_{r}(h)-{\widehat {Q}}_{s}(h)|\geq \varepsilon /2} , where r {\displaystyle r} and s {\displaystyle s} are i.i.d samples of size m {\displaystyle m} drawn according to the distribution P {\displaystyle P} . One can view r {\displaystyle r} as the original randomly drawn sample of length m {\displaystyle m} , while s {\displaystyle s} may be thought as the testing sample which is used to estimate Q P ( h ) {\displaystyle Q_{P}(h)} .

Permutation: Since r {\displaystyle r} and s {\displaystyle s} are picked identically and independently, so swapping elements between them will not change the probability distribution on r {\displaystyle r} and s {\displaystyle s} . So, we will try to bound the probability of | Q ^ r ( h ) − Q ^ s ( h ) | ≥ ε / 2 {\displaystyle |{\widehat {Q}}_{r}(h)-{\widehat {Q}}_{s}(h)|\geq \varepsilon /2} for some h ∈ H {\displaystyle h\in H} by considering the effect of a specific collection of permutations of the joint sample x = r | | s {\displaystyle x=r||s} . Specifically, we consider permutations σ ( x ) {\displaystyle \sigma (x)} which swap x i {\displaystyle x_{i}} and x m + i {\displaystyle x_{m+i}} in some subset of 1 , 2 , . . . , m {\displaystyle {1,2,...,m}} . The symbol r | | s {\displaystyle r||s} means the concatenation of r {\displaystyle r} and s {\displaystyle s} . [ citation needed ]

Reduction to a finite class: We can now restrict the function class H {\displaystyle H} to a fixed joint sample and hence, if H {\displaystyle H} has finite VC Dimension, it reduces to the problem to one involving a finite function class.

We present the technical details of the proof.

### Symmetrization

Lemma: Let V = { x ∈ X m : | Q P ( h ) − Q ^ x ( h ) | ≥ ε for some h ∈ H } {\displaystyle V=\{x\in X^{m}:|Q_{P}(h)-{\widehat {Q}}_{x}(h)|\geq \varepsilon {\text{ for some }}h\in H\}} and

Then for m ≥ 2 ε 2 {\displaystyle m\geq {\frac {2}{\varepsilon ^{2}}}} , P m ( V ) ≤ 2 P 2 m ( R ) {\displaystyle P^{m}(V)\leq 2P^{2m}(R)} .

Proof:

By the triangle inequality, if | Q P ( h ) − Q ^ r ( h ) | ≥ ε {\displaystyle |Q_{P}(h)-{\widehat {Q}}_{r}(h)|\geq \varepsilon } and | Q P ( h ) − Q ^ s ( h ) | ≤ ε / 2 {\displaystyle |Q_{P}(h)-{\widehat {Q}}_{s}(h)|\leq \varepsilon /2} then | Q ^ r ( h ) − Q ^ s ( h ) | ≥ ε / 2 {\displaystyle |{\widehat {Q}}_{r}(h)-{\widehat {Q}}_{s}(h)|\geq \varepsilon /2} .

Therefore,

since r {\displaystyle r} and s {\displaystyle s} are independent.

Now for r ∈ V {\displaystyle r\in V} fix an h ∈ H {\displaystyle h\in H} such that | Q P ( h ) − Q ^ r ( h ) | ≥ ε {\displaystyle |Q_{P}(h)-{\widehat {Q}}_{r}(h)|\geq \varepsilon } . For this h {\displaystyle h} , we shall show that

Thus for any r ∈ V {\displaystyle r\in V} , A ≥ P m ( V ) 2 {\displaystyle A\geq {\frac {P^{m}(V)}{2}}} and hence P 2 m ( R ) ≥ P m ( V ) 2 {\displaystyle P^{2m}(R)\geq {\frac {P^{m}(V)}{2}}} . And hence we perform the first step of our high level idea.

Notice, m · Q ^ s ( h ) {\displaystyle m\cdot {\widehat {Q}}_{s}(h)} is a binomial random variable with expectation m · Q P ( h ) {\displaystyle m\cdot Q_{P}(h)} and variance m · Q P ( h ) ( 1 − Q P ( h ) ) {\displaystyle m\cdot Q_{P}(h)(1-Q_{P}(h))} . By Chebyshev's inequality we get

for the mentioned bound on m {\displaystyle m} . Here we use the fact that x ( 1 − x ) ≤ 1 / 4 {\displaystyle x(1-x)\leq 1/4} for x {\displaystyle x} .

Permutations

Let Γ m {\displaystyle \Gamma _{m}} be the set of all permutations of { 1 , 2 , 3 , … , 2 m } {\displaystyle \{1,2,3,\dots ,2m\}} that swaps i {\displaystyle i} and m + i {\displaystyle m+i} ∀ i {\displaystyle \forall i} in some subset of { 1 , 2 , 3 , … , 2 m } {\displaystyle \{1,2,3,\ldots ,2m\}} .

Lemma: Let R {\displaystyle R} be any subset of X 2 m {\displaystyle X^{2m}} and P {\displaystyle P} any probability distribution on X {\displaystyle X} . Then,

where the expectation is over x {\displaystyle x} chosen according to P 2 m {\displaystyle P^{2m}} , and the probability is over σ {\displaystyle \sigma } chosen uniformly from Γ m {\displaystyle \Gamma _{m}} .

Proof:

For any σ ∈ Γ m , {\displaystyle \sigma \in \Gamma _{m},}

(since coordinate permutations preserve the product distribution P 2 m {\displaystyle P^{2m}} .)

The maximum is guaranteed to exist since there is only a finite set of values that probability under a random permutation can take.

Reduction to a finite class

Lemma: Basing on the previous lemma,

Proof:

Let us define x = ( x 1 , x 2 , … , x 2 m ) {\displaystyle x=(x_{1},x_{2},\ldots ,x_{2m})} and t = | H | x | {\displaystyle t=|H|_{x}|} which is at most Π H ( 2 m ) {\displaystyle \Pi _{H}(2m)} . This means there are functions h 1 , h 2 , … , h t ∈ H {\displaystyle h_{1},h_{2},\ldots ,h_{t}\in H} such that for any h ∈ H , ∃ i {\displaystyle h\in H,\exists i} between 1 {\displaystyle 1} and t {\displaystyle t} with h i ( x k ) = h ( x k ) {\displaystyle h_{i}(x_{k})=h(x_{k})} for 1 ≤ k ≤ 2 m . {\displaystyle 1\leq k\leq 2m.}

We see that σ ( x ) ∈ R {\displaystyle \sigma (x)\in R} iff for some h {\displaystyle h} in H {\displaystyle H} satisfies, | 1 m | { 1 ≤ i ≤ m : h ( x σ i ) = 1 } | − 1 m | { m + 1 ≤ i ≤ 2 m : h ( x σ i ) = 1 } | | ≥ ε 2 {\displaystyle |{\frac {1}{m}}|\{1\leq i\leq m:h(x_{\sigma _{i}})=1\}|-{\frac {1}{m}}|\{m+1\leq i\leq 2m:h(x_{\sigma _{i}})=1\}||\geq {\frac {\varepsilon }{2}}} .

Hence if we define w i j = 1 {\displaystyle w_{i}^{j}=1} if h j ( x i ) = 1 {\displaystyle h_{j}(x_{i})=1} and w i j = 0 {\displaystyle w_{i}^{j}=0} otherwise.

For 1 ≤ i ≤ m {\displaystyle 1\leq i\leq m} and 1 ≤ j ≤ t {\displaystyle 1\leq j\leq t} , we have that σ ( x ) ∈ R {\displaystyle \sigma (x)\in R} iff for some j {\displaystyle j} in 1 , … , t {\displaystyle {1,\ldots ,t}} satisfies | 1 m ( ∑ i w σ ( i ) j − ∑ i w σ ( m + i ) j ) | ≥ ε 2 {\displaystyle |{\frac {1}{m}}\left(\sum _{i}w_{\sigma (i)}^{j}-\sum _{i}w_{\sigma (m+i)}^{j}\right)|\geq {\frac {\varepsilon }{2}}} . By union bound we get

Since, the distribution over the permutations σ {\displaystyle \sigma } is uniform for each i {\displaystyle i} , so w σ i j − w σ m + i j {\displaystyle w_{\sigma _{i}}^{j}-w_{\sigma _{m+i}}^{j}} equals ± | w i j − w m + i j | {\displaystyle \pm |w_{i}^{j}-w_{m+i}^{j}|} , with equal probability.

Thus,

where the probability on the right is over β i {\displaystyle \beta _{i}} and both the possibilities are equally likely. By Hoeffding's inequality , this is at most 2 e − m ε 2 / 8 {\displaystyle 2e^{-m\varepsilon ^{2}/8}} .

Finally, combining all the three parts of the proof we get the Uniform Convergence Theorem .

# References