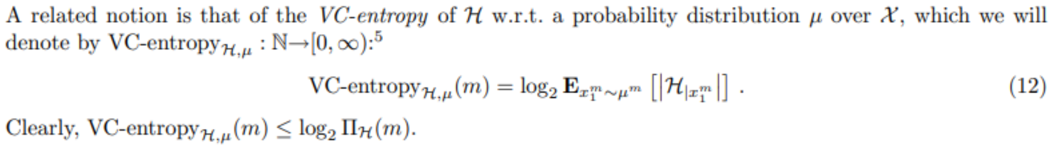
Afterward to Estimation of Dependences Based on Empirical Data (Vapnik) (2006)

- Vapnik's bio (paraphrasing Wikipedia and parts of reading)
  - From Jewish family in Soviet Union
  - Got PhD there (not without some drama)
  - Moved to USA in 1990 and started working at Bell Labs (alongside some notable researchers)
  - Currently affiliated with Columbia University and Facebook AI Research
- Relevant terminology (or "what is VC-dimension?")
  (some content shamelessly stolen from Siva's/Larry's 10/36-705 Intermediate Statistics course)
  - Suppose we have a class of functions as well as a set of *n* points
  - **n-th shatter coefficient**: the maximum number of subsets that can be selected from this set of *n* points
    - Can also think of this as number of ways functions from class can correctly classify any labeling of *n* points
    - At most 2^*n* for any set of *n* points
  - **VC-dimension**: the largest integer *d* for which shatter coefficient is 2^*d*
    - Does not have to be *all* possible combinations of *d* points (more on this soon)
  - **VC-entropy**: log_2 of expected value of subsets that can be selected from *n* points
    - Upper-bounded by log_2 of *n*-th shatter coefficient (see https://www.shivani-agarwal.net/Teaching/E0370/Aug-2011/Lectures/3.pdf)



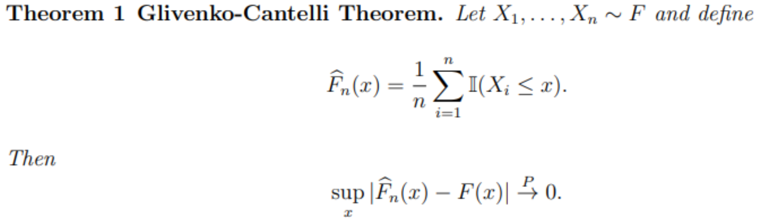
A related notion is that of the *VC-entropy* of $\mathcal{H}$ w.r.t. a probability distribution μ over $\mathcal{X}$, which we will denote by $\text{VC-entropy}_{\mathcal{H},\mu} : \mathbb{N} \rightarrow [0, \infty)$:[5]

$$\text{VC-entropy}_{\mathcal{H},\mu}(m) = \log_2 \mathbf{E}_{x_1^m \sim \mu^m} \left[\left| \mathcal{H}_{|x_1^m} \right| \right] . \tag{12}$$

Clearly, $\text{VC-entropy}_{\mathcal{H},\mu}(m) \leq \log_2 \Pi_{\mathcal{H}}(m)$.

    - May help to just think of shatter coefficient and VC-dimension, but listing here since Vapnik's definition on one of the early pages uses entropy instead of shatter coefficient
  - **Uniform convergence**: for a class of functions, stronger convergence towards a function than pointwise convergence
    - Almost like a Law of Large Numbers except at all points along a function

  - **Glivenko-Cantelli Theorem**: empirical CDF converges to true CDF

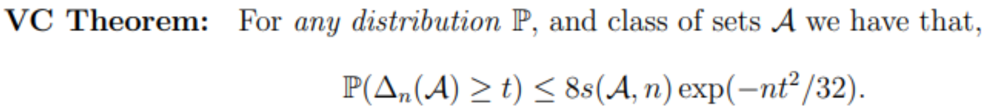
    

    Theorem 1 Glivenko-Cantelli Theorem. *Let $X_1, \ldots, X_n \sim F$ and define*

    $$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \le x).$$

    Then

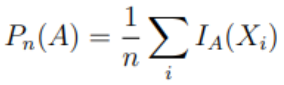
    $$\sup_x |\widehat{F}_n(x) - F(x)| \xrightarrow{P} 0.$$

  - *Note that uniform convergence is* ***not*** *guaranteed for all classes of functions; VC-dimension is a useful indicator here*
  - Quantitative version of VC Theorem:
    - 

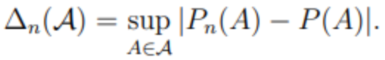
      VC Theorem: For *any distribution* $\mathbb{P}$, and class of sets $\mathcal{A}$ we have that,

      $$\mathbb{P}(\Delta_n(\mathcal{A}) \ge t) \le 8s(\mathcal{A}, n)\exp(-nt^2/32).$$

    - where s(A,*n*) is the n-th shatter coefficient and

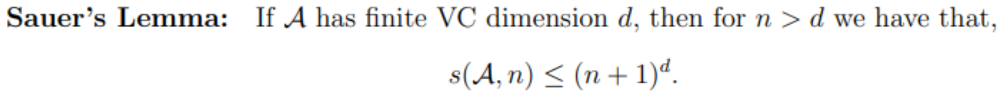
      

      $$P_n(A) = \frac{1}{n}\sum_i I_A(X_i)$$

      $$\Delta_n(\mathcal{A}) = \sup_{A\in\mathcal{A}} |P_n(A) - P(A)|.$$

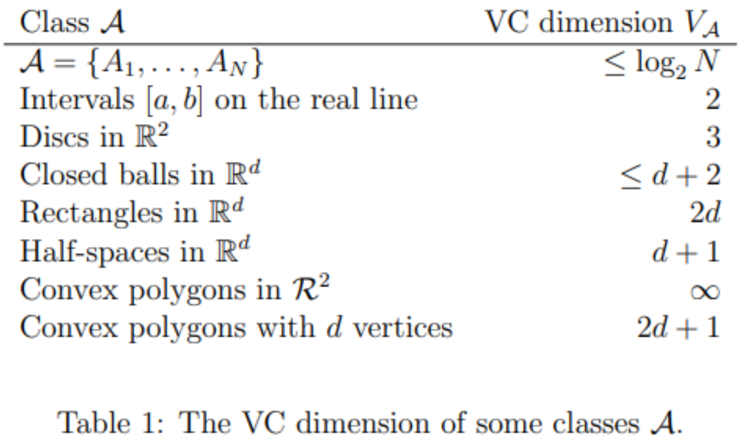
    - Key takeaway here is that if shatter coefficient does not grow exponentially, there is uniform convergence, but if it does, there is not uniform convergence
  - **Sauer's Lemma**:
    - (looks like Vapnik lost the battle here)

      Sauer's Lemma: If $\mathcal{A}$ has finite VC dimension d, then for $n > d$ we have that,

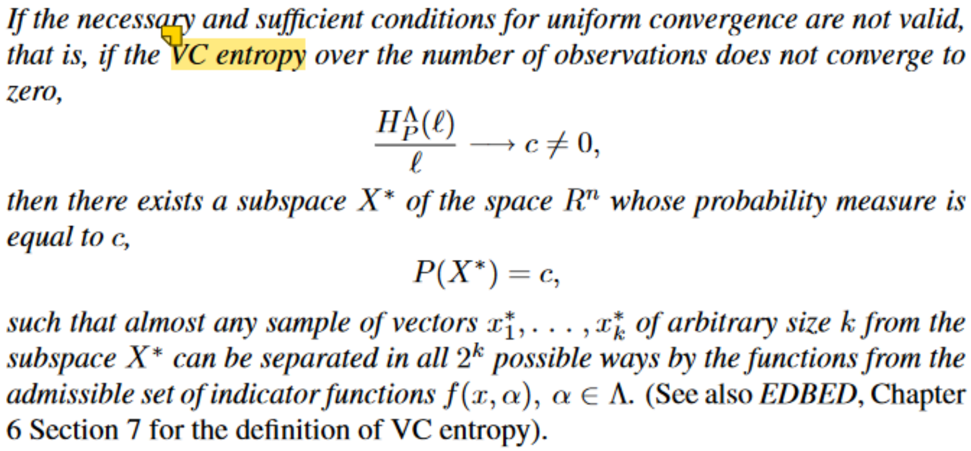
      $$s(\mathcal{A}, n) \le (n+1)^d.$$

    -
    - So if VC dimension is finite, shatter coefficient eventually grows sub-exponentially, in which case we have uniform convergence
- Examples to make all of this concrete
  - Points in 2D vs hyperplanes anchored at the origin
    - Can choose two points such that any line through the origin properly classifies them regardless of how they are labeled (e.g. (1,0) and (0,1))
    - *Cannot* do the same with three points
    - Therefore VC-dimension is 2
  - Points in 2D vs hyperplanes with offset
    - Possible to shatter a set of 3 points
    - Run into XOR problem with 4 points
    - Therefore VC-dimension is 3
  - Empirical CDF and half-spaces (-inf,t] in R^1
    - Can always define half-space that contains or does not contain 1 point
    - Fails with 2 points

        - VC-dimension is 1
        - According to Sauer’s lemma, shatter-coefficient is n+1 with n points
        - Combining this with quantitative version of VC theorem proves Glivenko-Cantelli theorem (uniform convergence of CDF)
    - Other examples
        - 

Class $\mathcal{A}$	VC dimension $V_{\mathcal{A}}$
$\mathcal{A} = \{A_1, \ldots, A_N\}$	$\leq \log_2 N$
Intervals $[a, b]$ on the real line	2
Discs in $\mathbb{R}^2$	3
Closed balls in $\mathbb{R}^d$	$\leq d + 2$
Rectangles in $\mathbb{R}^d$	$2d$
Half-spaces in $\mathbb{R}^d$	$d + 1$
Convex polygons in $\mathcal{R}^2$	∞
Convex polygons with d vertices	$2d + 1$

Table 1: The VC dimension of some classes $\mathcal{A}$.

- Chapter 1
    - Central question: “when do we generalize based on training data?”
    - 

If the necessary and sufficient conditions for uniform convergence are not valid, that is, if the VC entropy over the number of observations does not converge to zero,

$$\frac{H_P^\Lambda(\ell)}{\ell} \longrightarrow c \neq 0,$$

then there exists a subspace X^* of the space R^n whose probability measure is equal to c,

$$P(X^*) = c,$$

such that almost any sample of vectors $x_1^*, \ldots, x_k^*$ of arbitrary size k from the subspace X^* can be separated in all 2^k possible ways by the functions from the admissible set of indicator functions $f(x, \alpha)$, $\alpha \in \Lambda$. (See also *EDBED*, Chapter 6 Section 7 for the definition of VC entropy).

    - “This means that if uniform convergence does not take place then any algorithm that does not use additional prior information and picks up one function from the set of admissible functions cannot generalize”
        - So if VC dimension also grows with number of points (e.g. exponentially), then will fail to converge to something that generalizes
    - “If, however, the conditions for uniform convergence are valid then (as shown in Chapter 6 of EDBED) for any fixed number of observations one can obtain a bound that defines the guaranteed risk of error for the chosen function.”
        - Can use probability bounds (e.g. Chernoff, Chebyshev) to bound risk if there is uniform convergence but you only have a finite number of examples

    - VC theory constructed in response to how few parameters Perceptron used (in comparison to classical techniques developed by Fisher's line of work) to perform well
    - 2 ways to approximate a black-box model: approximate underlying model or approximate its error rate
        - Latter requires fewer parameters
    - Regularization vs structural risk minimization
        - Regularization for solving ill-posed problems to attempt to learn underlying black-box model
        - Structural risk minimization to learn model that converges to one that approximates error rate of black-box model
    - Split between statistical learning theory and classical statistics
        - Reminiscent of Breiman reading
    - "The Story Behind This Book"
        - Intense back-and-forth regarding publishing of work and graduation
- Chapter 2
    - PAC learning and simplification of theory
    - Working at Bell Labs with Neural Nets and Yann LeCun
        - Aside about how neural nets can converge to a local minimum but not generalize
    - Lots of math about SVMs
        - Basically solve a quadratic program to maximize margin between separator and "support vectors"
        - Can also provide custom kernels so not limited to linear separators
            - "One can even use kernels in the situation when input vectors belong to nonvectorial spaces. For example, the inputs may be sequences of symbols of different size (as in problems of bioinformatics or text classification). Therefore SVMs form a universal generalization engine that can be used for different problems of interest."
            - Provide radial basis function kernels as an example
        - Also lend well to structural risk minimization through being able to control VC dimension
    - Extensions of SVMs
        - SVM+, SVM Regression, …
    - Shout-out to "the third generation" of Statistical Learning Theory researchers for democratizing and spreading their techniques
        - (some familiar names here)
    - "Relation to the Philosophy of Science" (or "Popper Slanderfest")
        - "By the end of the 1990s it became clear that there were strong ties between machine learning research and research conducted in the classical philosophy of induction."
        - Starting with Occam's Razor
        - Defining VC dimension and VC falsifiability

Human Induction in Machine Learning: A Survey of the Nexus

Petr Spelda and Vit Stritecky

The human-ML nexus has two tiers:

1. *First tier.* Humans commit to assumptions (inductive predictions or generalizations).
2. *Second tier.* ML models learn generalizations from the available data under these assumptions.

Inductive prediction. $r\%$ of all so far observed Fs have been Gs. Therefore, with a (subjective) probability of approximately $r\%$, the next F will be a G—and thus will be predicted to be a G, provided r is greater than 1/2 and F is the total evidence regarding the next observed individual.

Inductive generalization. $r\%$ of all so far observed Fs have been Gs. Therefore, with high (subjective) probability, approximately $r\%$ of all Fs are Gs.

In the first tier,

- $F \rightarrow$ target environments
- $G \rightarrow$ target environments which preserve the uniformity in the training / evaluation environments
- $r \rightarrow$ likelihood of distribution shifts

In the second tier,

- $F \rightarrow$ predictions of the ML model
- $G \rightarrow$ correct predictions
- $r \rightarrow$ approximate generalization performance of the ML model

Proposition 1. For the second-tier r (of any trained ML model) to hold for targets past the holdout, the first-tier r must hold as well.

Informally, for the inductive inference of the ML model to be justified, the inductive inference of the human needs to be justified.

Environments

1. *Elsewheres in space and time.*
   - Using ML to reconstruct unobservable phenomena, distant in space or time. *e.g.,* Very Long Baseline Interferometry (VLBI) – distant celestial body imaging, using generalizations acquired from synthetic or local data.
   - Data is sparse and noisy. The reconstruction task is *underdetermined.*
   - ML model provides a reconstruction which is likely under some prior assumptions.
   - These assumptions are made by the human based on acquaintance with the present. *What validates this assumption?*
   - Uniformitarianism – assuming a uniformity between the training and target environments. But this cannot be justified in terms of reliability, only *optimality.*
2. *Non-benign environments.*
   - In the previous case, inductive inferences to uniformity lacked access to *ground-truths.* Here, we lack guarantees regarding their *stability.*
   - Non-robust regularities imperceptible to humans can be exploited by the attacker.
   - This makes human inductive inference about uniformity unreliable.

Assumptions in Machine Learning

1. *Empirical Risk Minimization (ERM)*
   - Assumes an unknown but fixed *distribution* from which training, evaluation and target samples are drawn independently.
   - This is a strong assumption which fails under dynamic environments.
   - Further, this invites *adversarial attacks* – models can learn highly predictive yet non-robust features which can be flipped at deployment.
2. *Invariant Risk Minimization (IRM)*
   - Relaxes the IID assumption with "exchangability" – ground truths must be invariant to permutation of samples.
   - Equivalently for classification, assumes invariance of the phenomenon-class label dependency.
   - The causal nature of the dependency doesn't help, as induction cannot be justified by causality (Hume).
   - Further, the reliability of the invariance assumption itself depends on the human inductive generalization stemming from so-far observed variance.
   - Moreover, IRM is a local prediction strategy, so the dependencies need to be enumerated at the object-level to discen an invariance.

Both these assumptions are inductive inferences. Both themselves represent a ground-truth. Hence, non-circular justification becomes impossible.

The Optimality Justification

Schurz's exponential attractivity-weighted meta-induction (eMI).

- Weight candidates based on their absolute successes.
- In the limit, eMI is a universally optimal prediction strategy.
- The optimality is independent of ground truths or uniformity assumptions. This allows for a non-circular justification.
- The human-machine enterprise is seen as acting according to Schurz's strategy, thus giving a method that "achieves the best possible acquaintance with the available state of affairs."

Impact of in-accessible/unstable ground-truths.

1. Instead of the ground-truth, we use the next best thing – a reference frame of spatiotemporally situated observers – as success records.
2. Further the received updates are delayed.
3. Optimality of meta-induction still holds.

Takeaways

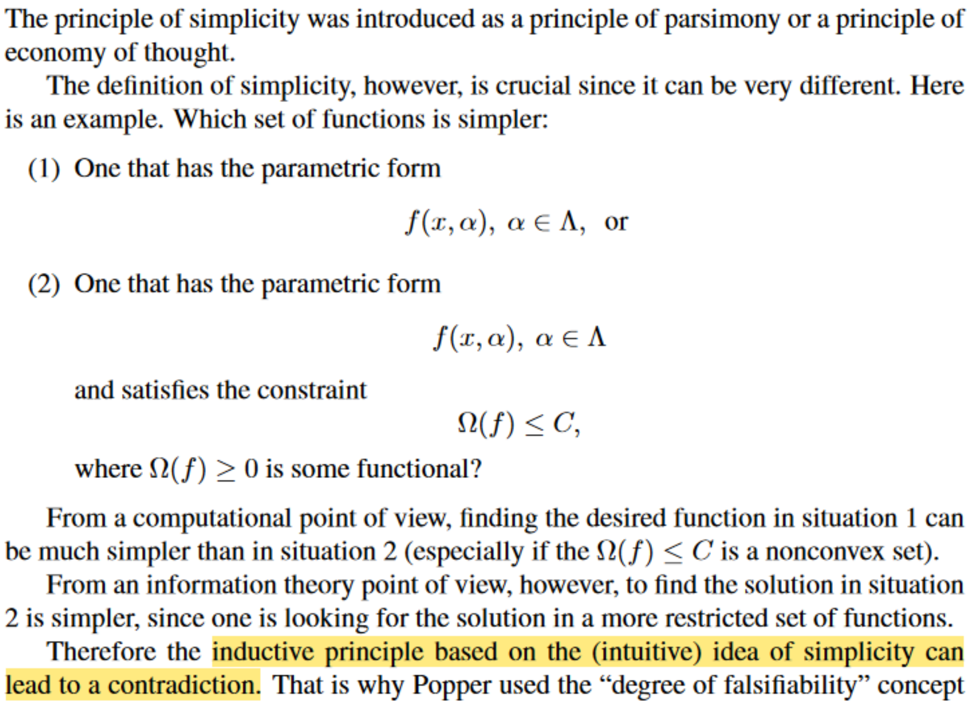
1. The human-ML contract is fragile. This should reflect in our epistemic ambitions.
2. The contract is based on optimality, rather than reliability.
3. For effective eMI, a constant inflow of updates is important. It is important to have leaderboards which are kept public and meticulously updated even on longer timescales.

4. Replacing inaccessible/unstable ground-truths with present information from our local environment is our best bet to extend our epistemic pursuits. This is fine as long as we don't lose sight of the fragile contract that justifies the human assumptions in ML.

Critiques

1. It is not clear to me how closely ML practice resembles eMI. How do you quantify the success records and the weights associated with various predictive methods / assumptions?
2. Different scenarios warrant different assumptions. How does eMI handle that? Are there separate weights for separate settings?
3. This paper downplays the role of humans (domain experts) in coming up with domain-specific assumptions.
4. Thinking in terms of Norton's theory, it might be possible to talk of some reliability contract. The assumptions are inductive schemas inferred locally from facts, and facilitate induction under some inductive risk.
5. One can also try to verify / falsify the assumptions in practice, in a Popper-esque manner, rather than solely relying on the evolution of the eMI weights.

  - Set of functions of VC falsifiable if dimension is finite, otherwise nonfalsifiable
- Popper dimension is almost inverse of VC dimension
  - “(1) any h vectors cannot falsify it and (2) there exist h +1 vectors that can falsify this set.”
- VC dimension lends better to notions of simplicity than Popper dimension and related writing
  - E.g. Popper classified sinusoidal functions as “simple” yet they have infinite VC dimension
- “It is surprising that the mathematical correctness of Popper’s claims has never been discussed in the literature.”
- Turns out that simplicity is still difficult to define



The principle of simplicity was introduced as a principle of parsimony or a principle of economy of thought.

The definition of simplicity, however, is crucial since it can be very different. Here is an example. Which set of functions is simpler:

(1) One that has the parametric form

$$f(x, \alpha),\ \alpha \in \Lambda, \text{ or}$$

(2) One that has the parametric form

$$f(x, \alpha),\ \alpha \in \Lambda$$

and satisfies the constraint

$$\Omega(f) \leq C,$$

where $\Omega(f) \geq 0$ is some functional?

From a computational point of view, finding the desired function in situation 1 can be much simpler than in situation 2 (especially if the $\Omega(f) \leq C$ is a nonconvex set).

From an information theory point of view, however, to find the solution in situation 2 is simpler, since one is looking for the solution in a more restricted set of functions.

Therefore the inductive principle based on the (intuitive) idea of simplicity can lead to a contradiction. That is why Popper used the “degree of falsifiability” concept

- Describes experiments from training SVMs with background knowledge (Universum)
  - Involves artificially generating images as part of training data
- “In trying to find an interpretation of the role of the Universum in machine learning, it is natural to compare it to the role of culture in the learning of humans, where knowledge about real life is concentrated not only in examples of reality but also in artificial images that reflect this reality. To classify well, one uses inspiration from both sources.”

No free lunch theorems – Wolpert (1992, 1996)

Machine learning is a field devoted to concocting algorithms that are better able to learn from data

The no free lunch theorem is a result in learning theory that states that we can get no formal guarantees that any algorithm will perform better than any other

Authors relate this to global/universal versus local induction

Data-driven/model free = global induction

Model-dependent = local induction

Specification of a model or hypothesis class in a learning context is an *explicit choice of inductive bias*

Some interpreters of Hume's skeptical problem have understood it as saying that *in order to carry out induction and thus to successfully navigate the world*, *the mind-independent world we inhabit must be structured in such a way as to facilitate our induction over it*

Assuming a uniform distribution over learning scenarios is like assuming the world to be fundamentally unstructured

Such a "maximum-entropy universe" (Wolpert's phrase) does not permit the possibility of learning

If we cannot make assumptions about unseen instances, learning is impossible!

The "all algorithms are equal" interpretation of the NFL is a straw-man so I'm not going to get into that—same as interpretation of Hume as a "genuine inductive skeptic"

Rather Wolpert and Hume are both attempting to show how to be a good empiricist

Namely, being a good empiricist requires understanding that all induction "relies on a specific and local empirical assumption"

The high-level takeaway: machine learning, like any other scientific modelling or induction procedure, requires inductive biases, namely, it requires us to make assumptions about the target domain over which we are learning, assumptions about the ease of learning, assumptions about what we intend to learning, salience, etc.

All ML requires assumptions of learning-friendliness but not all ML requires *the same assumption of learning-friendliness*; it requires *local* assumptions

We can learn better if we make these (local) assumptions explicit!