**PAC Results**

**Occam’s Razor**

*Blumer*

* ‘Given two explanations, the simpler one is preferred’ – Occam’s Razor
  + Goal of machine learning is to find the simplest hypothesis consistent with the sample data
* Blumer shows that ‘under very general conditions’, Occam’s razor produces hypotheses that can predict future observations with low error
  + When hypotheses of minimum complexity are produced efficiently, we have a polynomial learning algorithm
* Applying the razor
  + Let **H** be the class of all Boolean functions representing regular language over **X**
    - Hypothesis class
  + Let **X** be the set of all finite strings of 0 and 1
    - Domain
  + Given **f** ∈ **H**, an observation of **f** is a point **x** ∈ **X** along with **f(x)**
  + **m** observations is a sample of size **m** (of **f**)
  + The problem is to recover **f** from a sample
  + Assumptions
    - Observations of **f** are independently made according to a fixed distribution, **P**
    - Selection of **m** samples is made using **Pm** on **Xm**
  + Error of a hypothesis is the probability it disagrees with **f** on a particular observation
  + Given **f** in **|H|=r**, the probability of a hypothesis with error larger than **e** is consistent with a sample of **f** of size **m** is less than **(1-e)mr**
* Bounds on sample size are independent of **f** and the probability distribution
* For an infinite hypothesis class size, let the complexity of the class be the number of bits needed to represent each hypothesis in a fixed encoding of the class
* A learning algorithm should choose a hypothesis that is consistent and shows minimum complexity
* This is not always practical for cases like DNF

**Role of Occam’s Razor in Knowledge Discovery**

*Domingos*

* The difficulty or simplicity of an ML model is usually equated with the syntactic size of the model
  + Nodes in a tree, parametres, conditions…
* Definition of error
  + Generalization error
    - Error on unseen examples
  + Training set error
    - Error on examples learnt from
* Definitions of the razor
  + First Razor
    - Given 2 models with the same generalization error, the simpler one should be preferred
  + Second Razor (**Domingos proves this false in the paper**)
    - Given 2 models with the same training set error, the simple one will have less generalization error
* Note, when comparing between models with differing dimensions (i.e. quadratic vs. linear), the higher dimensioned model is penalized because there are simply ‘more’ models in the class making the higher dimensioned model more prone to overfitting rather than because a higher dimension implies complexity.
  + Higher dimensioned model spaces can be narrowed
* No free lunch Theorem
  + For every domain where a simpler model performs better, there exist domains where complex models perform better (assuming same training set error)
* Conclusion
  + With the same generalization error, comprehensibility should be preferred over simplicity

**Probably Approximately Correct Learning**

*Haussler*

* Learning problems are feasible if the solution algorithm is in P-time
* PAC model
  + Concept class **C** is PAC learnable by hypothesis space **H** if there is a polynomial time algorithm **A** and polynomial **p(-,-,-)** such that for all **n>=1**, all **c** ∈ **C**, all probability distributions **D** on **{0,1}n** (instance space) and all **e, δ,** if **A** is given at least **p(n, 1/e, 1/δ)** independent random examples of **c** drawn by **D**, then with probability of at least **1-δ**, **A** returns a hypothesis **h** ∈ **H** with error less than **e**.
  + If concept class is equal to the hypothesis class, the concept class is properly PAC learnable
* Most difficulties in proper PAC learning are due to computational complexity of finding a hypothesis in a form specified by the target class
  + The greater the **H** space is, the easier the computational problem is
  + As the hypothesis space increases, it becomes easier to find a consistent hypothesis, but more random training samples are required
  + There is a tradeoff between computational and sample complexity
    - Restricting the hypothesis space is a form of bias to facilitate learning
* Sample complexity bounds
  + Same as Occam’s Razor bounds (at most) for finite hypothesis spaces
  + For infinite hypothesis spaces, the VC dimension of the class is a substitute
    - Alternate bound exists involving VC
* Criticisms of PAC
  + It is unusable because of the worst case emphasis
    - Error is overestimated as a result in the computations
    - Not good for predicting learning curves
    - Extensions to PAC learning like Bayesian PAC address some of these issues
  + It is too restrictive because of the concepts of ‘target concept’ and ‘noise free training’
    - The advantage of PAC is that negative results can be stated strongly
    - However, positive learnability results need to be strengthened
      * A method of fixing this is by making false positives more expensive than false negatives in loss terms