CS 446 Machine Learning Fall 2016 OCT 11, 2016

Computational Learning Theory

Professor: Dan Roth *Scribe: Ben Zhou, C. Cervantes*

1 PAC Learning

We want to develop a theory to relate the probability of successful learning, the number of training examples, the complexity of the hypothesis space, the accuracy to which the target concept is approximated, and the manner in which training examples are presented.

1.1 Prototypical Concept Learning

Consider *instance space* X, the set of examples, and *concept space* C, the set of target functions that could have generated the examples such that there exists a $f \in C$ that is the hidden target function. For example, C could be all n-conjunctions, all n-dimensional linear functions, etc.

The *hypothesis space* is the set of all possible hypotheses that our learning algorithm can choose from, where H is not necessarily equal to C. We consider our *training instances* to be $S \times \{0, 1\}$ – including both positive and negative examples of the target concept – such that training instances are generated by a fixed unknown probability distribution D over X. Each training instance can be thought of as a (data, label) tuple, as below

$$S = [(x_1, f(x_1)), (x_2, f(x_2))...(x_n, f(x_n))]$$

In this setting, our goal is to determine a hypothesis $h \in H$ that estimates f, evaluated by its performance on subsequent instances $x \in X$ drawn according to D.

Note the assumption that both training and testing instances are drawn from from the same distribution D, as it is important in the analysis that follows.

1.2 Intuition

Consider Figure 1, showing the space predicted by target function f and hypothesis function h, where points inside the circle are positive, points outside are negative, and the functions are given by

$$h = x_1 \wedge x_2 \wedge x_3 \wedge x_4 \wedge x_5 \wedge x_{100}$$

$$f = x_2 \wedge x_3 \wedge x_4 \wedge x_5 \wedge x_{100}$$

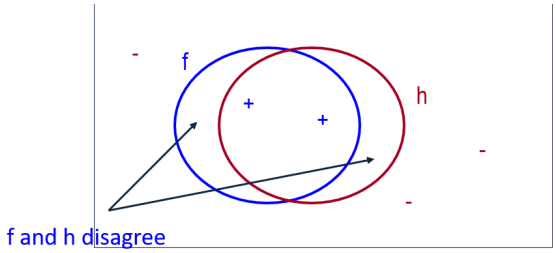

Figure 1: f not equal h

In this example, we have seen x_1 in all positive training instances (even though it is not active in f). Therefore, it is very likely that it will be active in future positive examples. If not, it is active in only a small percentage of examples, so the error should be small.

We can therefore consider the error as the probability of an example having different labels according to the hypothesis and the target function, given by

$$Error_D = Pr_{x \in D}[f(x) \neq h(x)]$$

2 Conjunctions

Consider z to be a literal in a conjunction. Let $p(z)$ be the probability that a D-sampled example is positive and z is false in it. In the example above, $z = x_1$.

2.1 Error Bounds

Claim: $Error(h) \leq \sum_{z \in h} p(z)$

Proof
Consider that $p(z)$ is the probability that a randomly chosen example is positive and z is deleted from h[1].

If z is in the target concept, then $p(z) = 0$; f is a conjunction and thus z can never be false in a positive example if $z \in f$.

h will make mistakes only on positive examples. A mistake is made only if z that is in h but not in f. In such cases, when $z = 0$, h will predict a negative example while f indicates a positive example.

[1]Recall that h is a conjunction, and z is a literal, such that if – during training – we see a positive example where $z = 0$, z is removed from h

Therefore $p(z)$ is also the probability that z causes h to make a mistake on a randomly drawn example from D. In a single example there could be multiple literals that are incorrect, but since in the worst case they are seen one-by-one, the sum of z bounds the error of h.

We can also consider literal $z \in h$ to be *bad* when $p(z) > \frac{\epsilon}{n}$. A *bad literal*, therefore, is a literal that is not in the target concept but has a significant probability to appear false in a positive example.

Claim: It there are no bad literals, then $Error(h) < \epsilon$

Proof
We have already stated that $Error(h) \leq \sum_{z \in h} p(z)$. Given that $|h| \leq n$ – the hypothesis has at most the same number of literals as there are features – then we know that the error cannot exceed $\sum^{n} \frac{\epsilon}{n} = \epsilon$.

Therefore $Error(h) < \epsilon$.

2.2 Example Bounds

Let z be a bad literal. We want to determine the probability that z has not been eliminated from h after seeing a given number of examples.

$$\begin{aligned} P(z \textit{ survives one example}) &= 1 - P(z \textit{ eliminated by one example}) \\ &\leq 1 - p(z) \\ &< 1 - \frac{\epsilon}{n} \end{aligned} \tag{1}$$

We can intuit that as we see more examples, the probability that bad literal z survives decreases exponentially, given by

$$p(z \textit{ survives } m \textit{ independent examples}) = (1 - p(z))^m < (1 - \frac{\epsilon}{n})^m$$

This is for one literal z. There are at most n bad literals, thus the probability that some bad literal survives m examples is bounded by $n(1 - \frac{\epsilon}{n})^m$

We want this probability to be bounded by δ, and thus we must choose m to be sufficiently large. Consider that we want

$$n(1 - \frac{\epsilon}{n})^m < \delta$$

Using $1 - x < e^{-x}$, it is sufficient to require

$$ne^{\frac{-m\epsilon}{n}} < \delta$$

Therefore, we need

$$m > \frac{n}{\epsilon}\{\ln(n) + \ln(\frac{1}{\delta})\}$$

to guarantee the probability of failure ($Error > \epsilon$) is less than δ

With probability $> 1 - \delta$, there are no bad literals, i.e., $Error(h) < \epsilon$

3 Formulating Prediction Theory

Given

- Instance space X
- Output space $Y = \{-1, +1\}$
- Distribution D, which is unknown over $X \times Y$
- Training examples S, where each is drawn independently from D ($|S| = m$

we can define the following

- True Error: $Error_D = Pr_{(x,y)\in D}[h(x)\neg = y]$
- Empirical Error: $Error_S = Pr_{(x,y)\in S}[h(x)\neg = y] = \sum_{1,m}[h(x_i)\neg = y_i]$
- Function space C, or set of possible target concepts, where $f \in C : X \rightarrow Y$
- Set of possible hypotheses H

We cannot expect a learner to learn a concept exactly. There may be many concepts consistent with the available data, and unseen examples may have any label. Thus we must agree to misclassify uncommon examples that were not seen during training.

Further, we cannot always expect a learner to learn a close approximation to the target concept, since sometimes the training set does not represent unseen examples.

Therefore, the only realistic expectation of a good learner is that it will learn a close approximation to the target concept with high probability.

3.1 Probably Approximately Correct Learning

In Probably Approximately Correct (PAC) learning, one requires that given small parameters ϵ and δ – with probability at least $(1 - \delta)$ – a learner produces a hypothesis with error at most ϵ.

This notion relies on the Consistent Distribution Assumption: there is one probability distribution D that governs both training and testing examples.

3.2 PAC Learnability

Consider a concept class C defined over an instance space X, and a learner L using a hypothesis space H.

C is PAC learnable by L using H
if $\forall f \in C$, $\forall D$ *over* X, and fixed $0 < \epsilon, \delta < 1$, L – given a collection of m examples sampled independently according to D – produces with probability at least $1 - \delta$ a hypothesis $h \in H$ with error at most ϵ where m is polynomial in $\frac{1}{\epsilon}$, $\frac{1}{\delta}$, n and $|H|$.

C is efficiently PAC learnable
if L can produce the hypothesis in time polynomial in $\frac{1}{\epsilon}$, $\frac{1}{\delta}$, n and $size(H)$.

Two Limitations

- Polynomial sample complexity, which is also called information theoretic constraint, governs if there is enough information in the sample to distinguish a hypothesis h that approximate f.
- Polynomial time complexity, also called computational complexity, which tells if there is an efficient algorithm that can process the sample and produce a good hypothesis h.

To be PAC Learnable, there must be a hypothesis $h \in H$ with arbitrary small error for every $f \in C$. We generally assume H is a super set of C.

The worst definition is that the algorithm must meet its accuracy for every distribution and every target function $f \in C$.

3.3 Occam's Razor

We want to prove the general claim that smaller hypothesis spaces are better.

Claim: The probability that there exists a hypothesis $h \in H$ that is consistent with m examples and satisfies $Error(h) > \epsilon$ is less then $|H|(1 - \epsilon)^m$.

Proof
Let h be a bad hypothesis. The probability that h is consistent with on examples is less than $1-\epsilon$. Since the m examples are independently drawn, the probability that h is consistent with m examples is less than $(1 - \epsilon)^m$.

The probability that any one of the hypothesis in H is consistent with m examples is less than $|H|(1 - \epsilon)^m$.

Given this fact, we now want this probability to be smaller than δ, that is

$$|H|(1 - \epsilon)^m < \delta$$

$$\ln(|H|) + m\ln(1-\epsilon) < \ln(\delta)$$

With the fact that $e^{-x} > 1 - x$, we have

$$m > \frac{1}{\epsilon}\{\ln(|H|) + \ln(\frac{1}{\delta})\}$$

This is called Occam's razor, because it indicates a preference towards small hypothesis space, i.e., if you have small hypothesis space, you do not have to see too many examples.

There is also a trade-off of the hypothesis space. If the space is small, then it generalizes well, but it many not be expressive enough.

4 Consistent Learners

Using the results from the previous section, we can get this general scheme for PAC learning:

Given a sample of m examples, find some $h \in H$ that is consistent with all m examples. If m if large enough, a consistent hypothesis will be sufficiently close to f. We can then check that m scales polynomially in the relevant parameters (i.e. m is not too large). "Closeness" guarantees

$$m > \frac{1}{\epsilon}(\ln|H| + \ln\frac{1}{\delta})$$

In the case of conjunctions, we used the elimination algorithm, which results in a hypothesis h that is consistent with the training set, and we showed directly that if we have sufficiently many examples (polynomial in relevant parameters), then h is close to the target function.

4.1 Examples

Conjunctions
For conjunctions, the size of the hypothesis space is 3^n, since there are 3 possible values for each of the n features (appear negative, positive, or not at all. Therefore, the number of examples we need according to the PAC learning framework, m, is given by

$$m > \frac{1}{\epsilon}\{\ln(3^n) + \ln(\frac{1}{\delta})\} = \frac{1}{\epsilon}\{n\ln 3 + \ln(\frac{1}{\delta})\}$$

Thus, if we want to guarantee a 95% chance of learning a hypothesis $(1-\delta)$ of at least 90% accuracy $(1-\epsilon)$, with $n = 10$ boolean variables, $m > 140$.

If we change to $n = 100$, $m > 1130$, which shows m is linear with n.

However, changing the confidence $(1 - \delta)$ to 99% makes $m > 1145$, shows m is logarithmic with δ.

k-CNF
Consider Conjunctive Normal Form functions (CNFs), which can express any boolean function. Recall that CNFs are conjuctions of disjunctions. A subset of this class is that of k-CNF, where each disjunction contains k terms, as in

$$f = \wedge_{i=1}^{m}(l_{i_1} \vee l_{i_2} \vee ... \vee l_{i_k})$$

To determine if we can learn such a class of functions, we must know the size of this hypothesis space. In this case, the hypothesis space is given by $2^{(2n)^k}$, corresponding to the number of ways to choose subsets from among the k literals, including negations. Thus, the sample complexity is given by

$$\ln(|k - CNF|) = O(n^k)$$

Since k is fixed, we have an order polynomial in the number of examples and thus h is guaranteed to be PAC learnable. Next step is to learn a consistent hypothesis.

Now we must consider how to learn such a hypothesis. Using what we know now, we cannot learn this directly. We can learn k-CNFs, however, if we move to a new space.

Consider the example in which $n = 4$ and $k = 2$ for a monotone k-CNF. Here, there are six disjunctions for which we can create a new mapping from the orig-

$$\begin{array}{ll} y_1 = x_1 \vee x_2 & y_2 = x_1 \vee x_3 \\ y_3 = x_1 \vee x_4 & y_4 = x_2 \vee x_3 \\ y_5 = x_2 \vee x_4 & y_6 = x_3 \vee x_4 \end{array}$$

inal space to a new space with six features: $(0000, 1) \rightarrow (000000, 1)$
$(1010, 1) \rightarrow (111101, 1)$
$(1110, 1) \rightarrow (111111, 1)$
$(1111, 1) \rightarrow (111111, 1)$
Now we can apply a standard algorithm for learning monotone conjunctions.

Unbiased Learning
Consider the hypothesis space of all boolean function on n features. There are 2^{2^n} different functions and the bound $(\ln(|H|)$ is therefore exponential in 2^n, which means that in general the set of all boolean functions are not PAC learnable.

k-Clause CNF
Conjunctions of at most k disjunctive clauses.

$$f = C_1 \wedge C_2 \wedge ...C_k; C_i = l_1 \vee l_2... \vee l_m$$

The size of the hypothesis space $\ln(|H|) = O(kn)$ is linear in n and thus PAC learnable.

k-DNF
Disjunctions of any number of terms where each conjunctive term has at most k literals.

$$f = T_1 \vee T_2 ... \vee T_m; T_i = l_1 \wedge l_2 \wedge ... \wedge l_m$$

4.2 k-term DNF Computational Complexity

Consider the class of k-term DNFs, or disjunctions of at most k conjunctive terms. From the sample complexity perspective, we should be able to learn in the same way as with k-CNFs, but computational complexity is challenging.

Consider a 2-term DNF consistent with a set of training data is NP-hard. Thus, even though 2-term DNFs are PAC learnable, they are not efficiently PAC learnable.

We can address this by enlarging the hypothesis space. If the hypothesis we wish to learn can be represented in a larger hypothesis space that we know is learnable, we can learn our desired hypothesis. In this case, we can represent k-term DNFs as k-CNFs, since k-CNF is a superset of k-term DNF. Consider a 3-term DNF (left) and its equivalent 3-CNF (right).

$$T_1 \vee T_2 \vee T_3 = \prod_{x \in T_1, y \in T_2, z \in T_3} \{x \vee y \vee z\}$$

Representation is important. Concepts that cannot be leaned using one representation may be learned using another more expressive representation.

However, this leaves us with two problems:

> How can we learn when data is not completely consistent with training data?
>
> How can we learn in an infinite hypothesis space?

5 Agnostic Learning

Assume we are trying to learn concept f using hypothesis space H, but $f \notin H$. We therefore cannot learn a a completely consistent hypothesis, and thus our goal is to find a hypothesis $h \in H$ that has as small training error as possible.

$$Err_{TR} = \frac{1}{m} \sum_{i}^{m} f(x_i) \neq h(x_i)$$

where x_i is the i^{th} training example.

We want to guarantee that a hypothesis with a small training error will have good accuracy on unseen examples, and one way to do so is with Hoeffding bounds. This characterizes the deviation between the true probability of some event and its observed frequency over m independent trails.

$$Pr[p > \hat{p} + \epsilon] < e^{-2m\epsilon^2}$$

To understand the intuition, consider tossing a biased coin. The more tosses, the more likely the observed result will correspond with the expected result. Similarly, the probability that an element in H will have training error which is off by more than ϵ can be bounded as follows:

$$Pr[Err_D(h) > Err_{TR}(h) + \epsilon] < e^{-2m\epsilon^2}$$

If we consider $\delta = |H|e^{-2m\epsilon^2}$, we can get a generalization bound, or how much will the true error E_D deviate from the observed (training) error E_{TR}.

For any distribution D, generating training and test instances with probability at least $1 - \delta$ over the choice of the training set of size m, (drawn i.i.d.), for all $h \in H$

$$Error_D(h) < Error_{TR}(h) + \sqrt{\frac{\log |H| + \log(\frac{1}{\delta})}{2m}}$$

An agnostic learner which makes no commitment to whether f is in H returns the hypothesis with least training error over at least the following number of examples m can guarantee with probability at least $1 - \delta$ that its training error is not off by more than ϵ from the true error. We therefore require a number of examples given by

$$m > \frac{1}{2\epsilon^2}\{\ln(|H|) + \ln(\frac{1}{\delta})\}$$

Learnability depends on the log of the size of the hypothesis space.

6 VC Dimension

For both consistent and agnostic learners, we assumed finite hypothesis spaces. We know consider an infinite hypothesis space.

6.1 Learning Rectangles

Consider a target concept as an axis parallel rectangle (positive points inside, negative outside, as given by Figure 2.

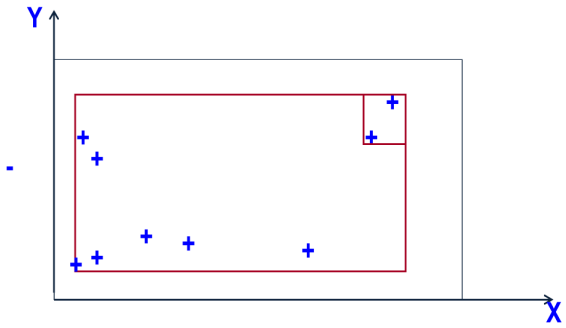

Figure 2: Learning Rectangles

We can simply choose the maximum x and y values as well as the minimum x and y values of the positive examples as the boundary for the rectangle. This is generally a good algorithm because it learns efficiently, but we cannot use the theorem from before to derive a bound because the hypothesis space is infinitely large.

Therefore, we need to find out how to derive a bound given an infinitely large hypothesis space.

6.2 Infinite Hypothesis Space

Just as before, where we discussed small hypothesis spaces (conjunctions) and large hypothesis spaces (DNFs), some infinite hypothesis spaces are larger (more expressive) than others; rectangles and general convex polygons have different levels of expressiveness.

We therefore need to measure the expressiveness of an infinite hypothesis space. The Vapnik-Chervonenkis dimension – or VC dimension – provides such a measure. Analogous to $|H|$, there are bounds for sample complexity using $VC(H)$.

6.3 Shattering

The key notion behind VC-dimension is that of *shattering*. Assume a set of points. We want to use a function to separate all possible labelings of the set of points. In Figure 3, a linear function (green) can separate the two points, regardless of how they're labeled.

For two points on a plane, there are two different ways of labeling the two points. A line can separate those two points no matter how they are labeled.

Figure 3: A linear function can shatter 2 points

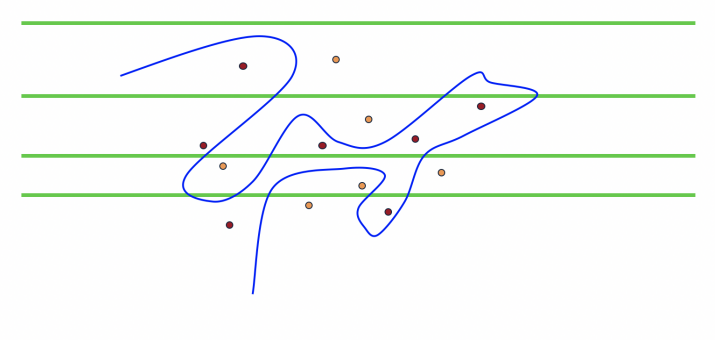

Figure 4: A set of multiple points

However, if there are many points in the set on a plane, as in Figure 4, no straight line can separate any labeling of these points. Linear functions are not expressive enough to shatter 13 points, but more expressive functions are.

We say that a set S of examples is shattered by a set of functions H if – for every partition of the examples in S into positive and negative examples – there is a function in H that gives exactly there labels to the examples. The intuition is that a rich set of functions shatters a large set of points.

(a) A function example

(b) function fails to shatter

Consider the function in which left bounded intervals on the real axis for some number is positive ($[0, a)$, for some $a > 0$).

It is trivial for this function to shatter a single point. However, in any set of two points on the line, the left can be labeled negative and the right positive, which

this H cannot label correctly. Thus, left bounded intervals cannot shatter two points.

Similarly, if we consider the class of functions for which real numbers $b > a$ and points within $[a, b]$ is positive, all sets of one or two points are shatterable, but no set of three points can be shattered.

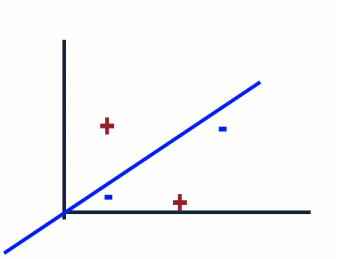

Figure 6: Half space

As a final example, consider half-spaces on the plane. We can trivially shatter all sets of one point and two points. We can shatter some sets of three points, but cannot shatter any set of four points. If the 4 points form a convex polygon, then by labeling each point different from its neighbors, the four points cannot be shattered. If, on the other hand, they do not form a convex polygon, then one point is inside the convex hull defined by the other three, and if that point is negative, there is no way to shatter them.

6.4 Definition

An unbiased hypothesis space H shatters the entire instance space X if it is able to induce every possible partition on the set of all possible instances.

The larger the subset X that can be shattered, the more expressive a hypothesis space is (i.e. less biased). The VC dimension of hypothesis space H over instance space X is the size of the largest finite subset X (even if there is only one subset) that is shattered by H.

If there exists a subset of size d that can be shattered, then $VC(H) \geq d$.

If no subset of size d can be shattered, then $VC(H) < d$.

- $VC(\textit{Half intervals}) = 1$; No subset of size 2 can be shattered
- $VC(\textit{Intervals}) = 2$; No subset of size 3 can be shattered
- $VC(\textit{Half-spaces in the plane}) = 3$; No subset of size 4 can be shattered

6.5 Sample complexity and VC dimension

VC dimension serves the same role as the size of the hypothesis space. Using VC dimension as a measure of expressiveness, we can give an Occam algorithm for infinite hypothesis spaces.

Given a sample D of m examples we will find $h \in H$ that is consistent with all m examples, if

$$m > \frac{1}{\epsilon}\{8VC(H)\log\frac{13}{\epsilon} + 4\log(\frac{2}{\delta})\}$$

then with probability at least $1-\delta$, h has error less then ϵ.

We consider that the hypothesis space has to be infinite if we want to use this bound. If we want to shatter m points, then H has to be at least 2^m in order to shatter any configurations of those m examples.

Thus $|H| > 2^m$, $\log(|H|) \geq VC(H)$.

6.6 Axis-Parallel Rectangles, Continued

Consider again the problem of learning axis-parallel rectangles. To determine if axis-parallel rectangles are PAC-learnable, we must determine sample complexity. Here, we show that $VC(H) \geq 4$. In Figure 7a, it is trivial to find an

(a) A set that can be shattered (b) a set that cannot be shattered

axis-parallel rectangle to shatter the points, regardless of their labeling. Though Figure 7b illustrates a set of four that is not shatterable, it is sufficient to find any set of four to prove the VC dimension is greater or equal to 4.

Next we need to argue that no set of five points can be shattered. For any layout of five points, we just need to show one kind of labeling that cannot be shattered. In this case, we can say that – of the five points – there must be a minimum and maximum x and y. There must by definition be a point within those minimum and maximum bounds, and thus by labeling the extreme four points as positive but the fifth, internal point as negative, axis-parallel rectangles cannot shatter five points.

Thus, from sample complexity perspective, axis-parallel rectangles are PAC learnable.

To determine if this hypothesis class is efficiently PAC learnable, we need an efficient algorithm to find the rectangle. Here, we can find the smallest example rectangle that contains all positive examples. This is likely not the best rectangle – it cannot generalize to new positive examples – so in the ideal case we would like a margin, that is, a rectangle slightly larger than the minimum one we've seen during training.

Given such an algorithm, we have shown that axis-parallel rectangles are efficiently PAC learnable.

6.7 Sample Complexity Lower Bound

We've discussed upper bounds on the number of examples; that is, if we have seen m examples, we can be reasonably sure our algorithm will perform well on new examples. There is also a general lower bound on the minimum number of examples necessary for PAC learning.

Consider any concept class C such that $VC(C) > 2$. For any learner L and small enough ϵ, δ, there exists a distribution D and a target function in C such that if L observes less than

$$m = max[\frac{1}{\epsilon}\log(\frac{1}{\delta}), \frac{VC(C)-1}{32\epsilon}]$$

examples, then with probability at least δ, L outputs a hypothesis having $error(h) > \epsilon$.

This is the inverse of the bound algorithm we have seen before.

Ignoring constant factors, the lower bound is the same as the upper bound, except for the extra $\log(\frac{1}{\epsilon})$ factor in the upper bound.

7 Conclusion

The PAC framework provides a reasonable model for theoretically analyzing the effectiveness of learning algorithms.

We discussed that the sample complexity for any consistent learner using the hypothesis space, H, can be determined from a measure of Hs expressiveness $(|H|, VC(H))$.

We discussed consistent and agnostic learners, showing that the log of the size of a finite hypothesis space is most important, and then extended this notion to the infinite hypothesis space.

We also discussed sample and computational complexity, showing that if sample complexity is tractable, the computational complexity of finding a consistent hypothesis governs the complexity of the learning problem.

Many additional models have been studied as extensions of the basic one: learning with noisy data, learning under specific distributions, learning probabilistic representations, etc..

An important extension is PAC-Bayesians theory, where the idea is that – rather than simply assume that training and test are governed by the same distribution – assumptions are also made about the prior distribution over the hypothesis space.

It's important to note that though the bounds we compute are loose, they can still guide model selection. A lot of recent work is on data dependent bounds.

The impact COLT has had on practical learning system in the last few years has been very significant: SVMs, Winnow (Sparsity), Boosting, and Regularization, to name a few.