Agnostic Learning and Structural Risk Minimization

These notes are slightly edited from scribe notes in previous years.

1 Introduction

In this lecture we address some limitations of the analysis of Occam algorithms that limit their applicability. We first discuss the case where the target concept $c$ is not in $H$ which is known as the non-realizable case and as agnostic PAC learning. We then turn to the case where the number of examples $m$ is fixed (i.e., we cannot ask for more examples as in the standard PAC model) and consider how we can handle an infinite size hypothesis class. This is the classical model selection problem that can be solved by structural risk minimization.

Although the results are presented for (unions of) finite classes the same arguments translate directly to the infinite case by replacing $\ln|H|$ with the VC dimension.

2 Target concept $c \notin H$

If $c \notin H$ it is not certain that we can find a hypothesis consistent with the example set. In this case we must refine our goal to do the best we can with $H$.

For any fixed concept $c$, distribution $D$, and hypothesis $h$ let $err(h) = err_{c,D}(h) = Pr_{D}[h(x) \neq c(x)]$. Define $h^* = h^*(c,D) \in H$ to be $\arg\min_{h \in H}(\hat{err}(h))$, so $h^*$ is the best hypothesis in $H$ when learning $c$ under $D$.

Definition: Agnostic PAC Learning An algorithm $A$ Agnostically PAC learns a concept class $H$ if $\forall D, \forall \delta, \forall \epsilon$, and $\forall c$, $A$ uses IID examples labeled by $c$ and, with probability $1 - \delta$, produces a hypothesis $h$ which has $err(h) \leq err(h^*) + \epsilon$. If $A$ runs in time polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$ then we say $A$ efficiently agnostically PAC learns $A$.

Assuming a finite size of the hypothesis class $H$, we can define an algorithm analogous to Occam Algorithms. This is known in the literature as “minimizing disagreements” or as “empirical risk minimization” (ERM). In the following explanation we use the notation $\hat{err}(h)$ to indicate $\frac{\# of mistakes on S}{|S|}$ where $S$ is the sample set used in learning.

Definition: ERM Algorithm An ERM algorithm for Agnostic PAC learning operates by taking a sample and finding $h = \arg\min_{h \in H} (\hat{err}(h))$.

Theorem: if $A$ is an ERM algorithm for $H$ and it uses $m \geq \frac{2}{\epsilon^2} \ln \frac{2|H|}{\delta}$ IID examples, then $A$ agnostically PAC learns $H$.
Proof: We will prove this theorem through the use of the following lemma.

Lemma 1: if \( \forall h \in H \; |err(h) - e\hat{r}(h)| \leq \frac{\epsilon}{2} \) then the hypothesis \( h \) output by \( A \) satisfies the condition \( err(h) \leq err(h^*) + \epsilon \)

Pick any bad hypothesis \( h_{bad} \) such that \( err(h_{bad}) > err(h^*) + \epsilon \). From the antecedent of Lemma 1 we have that \( e\hat{r}(h_{bad}) \geq err(h_{bad}) - \frac{\epsilon}{2} \) and \( err(h^*) \geq e\hat{r}(h^*) - \frac{\epsilon}{2} \) implying that \( e\hat{r}(h_{bad}) \geq err(h_{bad}) - \frac{\epsilon}{2} \geq err(h^*) + \epsilon - \frac{\epsilon}{2} \geq e\hat{r}(h^*) \). Since the ERM algorithm picks the hypothesis with the smallest empirical error, it would have then chosen \( h^* \) over \( h_{bad} \), and therefore no \( h \) defying the error bounds of agnostic PAC learning could be chosen.

All that is required to complete the proof is then to show that the condition in Lemma 1 is satisfied with probability \( \geq 1 - \delta \). The probability that a single hypothesis violates the condition of Lemma 1 can be bounded using the Chernoff bound:

\[
Pr[|err(h) - e\hat{r}(h)| \geq \frac{\epsilon}{2}] < 2e^{-\frac{m(\frac{\epsilon}{2})^2}{2}}
\]

Using \( m \geq \frac{\epsilon^2}{2} \ln \frac{2H^2}{\delta} \) gives \( Pr < \frac{\delta}{|H|} \) and by the union bound over all \( h \in H \) we have a probability of failure \( \leq \delta \).

The requirement that \( A \) produces the hypothesis which minimizes the empirical error, \( e\hat{r}(h) \), may be relaxed slightly without losing Agnostic PAC learnability.

Definition: Let \( h_{best} \) be the hypothesis from \( H \) which has minimum empirical error on a given sample \( S \). An algorithm \( A \) is an \( \alpha \)-Almost ERM algorithm for a concept class \( H \) if it produces a hypothesis \( h \in H \) which has \( e\hat{r}(h) \leq e\hat{r}(h_{best}) + \alpha \).

Theorem: if \( A \) is an Almost ERM algorithm then \( A \) is an Agnostic PAC learner.

The proof follows the same lines as that for ERM algorithms.

3 Agnostic Learning for General Bounded Loss Functions

The analysis of the previous part is quite general and can be applied for tasks other than classification. To illustrate this point we define the notions of loss and risk.

For hypothesis \( h \) and (example,label) pair \( (x,y) \) the Loss function \( \ell() \) abstractly captures the loss of \( h \) on the pair. For classification, the 0-1 loss \( \ell_{0-1}(h,(x,y)) \) is equal to 0 if \( h(x) = y \) and to 1 if \( h(x) \neq y \). The Risk is the expected loss when \( (x,y) \) pairs are drawn according to \( D \), that is, \( R(h) = E_{(x,y) \sim D}[\ell(h,(x,y))] \). For 0-1 loss the risk coincides with our notion of error \( R(h) = E_{(x,y) \sim D}[\ell_{0-1}(h,(x,y))] = E_{(x,y) \sim D}[h(x) \neq y] = err(h) \).

For other tasks both our predictions and the loss may not be binary. For example, for regression problems where we predict numerical values, a popular loss function is the square loss \( \ell_{sq}(h,(x,y)) = (h(x) - y)^2 \). For classification with logistic regression models, the standard optimization criterion aims to maximize likelihood, or equivalently minimize the negative log likelihood. The corresponding loss function is \( \ell_{sq}(h,(x,y)) = \ln(1 + e^{yh(x)}) \), where in this case the label \( y \) is in \( \{-1,1\} \). Here too the risk is the expectation of the loss.
The criterion for learnability requires a bound on the risk of the hypothesis of the algorithm to hold with high probability. Reformulating the agnostic case we would require: with probability $\geq 1 - \delta$, $R(h_{\text{hyp}}) \leq R(h^*) + \epsilon$.

The natural generic algorithm is ERM - that is where the name comes from. The proof of the previous section holds with very minor modification in this more general case. The only requirement is that the loss function has to be bounded. If this holds then we can apply Hoeffding’s bound and obtain a variant of the condition of lemma 1 replacing error by risk. The rest of the analysis is unchanged.

This provides an agnostic PAC guarantee for logistic regression. Unfortunately the square loss is not bounded so such a bound is not implied in general. For linear regression where $h(x) = w^T x$ for weight version $w$, we can bound the loss if both $\|w\|$ and $\|x\|$ are bounded. But this limits the range of $w$ allowed to be output by the algorithm. We will revisit the square loss later in the course.

4 Structural Risk Minimization

Another requirement of the Occam Theorem is that the size of the hypothesis class be finite or have finite VC dimension. In this section we consider the case where the hypothesis class is of potentially infinite size and the number of examples is fixed.

First, let us define our hypothesis class as consisting of a hierarchy of hypotheses classes $H_i$ indexed such that $H_0 \subseteq H_1 \subseteq H_2 \subseteq \ldots$ and $H = \bigcup_{i=0}^{\infty} H_i$. Our goal is to pick an index $i$, a class $H_i$ and a hypothesis $h \in H_i$ as the output of the learning algorithm. This classical problem is known as the model selection problem.

Simply selecting the smallest indexed hypothesis with zero empirical error will favor large models, however, and lead to overfitting. The idea, as in our previous analyses, is to provide a bound on the quality of $e\hat{r}(H)$ that holds simultaneously for all hypotheses. However, this is not possible (with fixed $\epsilon$) because $m$ is fixed and the hierarchy is infinite. The solution is to provide a different tolerance $\epsilon_i$ for each $i$ and adjust our choice of hypothesis accordingly. This algorithmic approach can be applied in different contexts and is known as Structural Risk Minimization (SRM).

The SRM algorithm is as follows, where, in this lecture we use $\epsilon_i = \sqrt{\frac{1}{2m} \ln \frac{4|H_i|^2}{\delta}}$.

- For each $i$ let $\hat{h}_i = \arg\min_{h \in H_i} e\hat{r}(h)$. For future reference note that this is the same as $\hat{h}_i = \arg\min_{h \in H_i} e\hat{r}(h) + \epsilon_i$.

- We then pick the $i$ that minimizes the penalized error estimate. That is we return $\hat{h}_i$ such that $i = \arg\min_j e\hat{r}(\hat{h}_j) + \epsilon_j$.

We can analyse this algorithm as follows:

- Fix any $i$ and any $h_i \in H_i$. Using Chernoff bound, we can show that $\Pr[|err(h_i) - e\hat{r}(h_i)| > \epsilon_i] \leq \frac{\delta}{2^{|H_i|}}$.

- Using a union bound, we can bound the probability that the error estimate for any $h_i$ in any $H_i$ is off by more than $\epsilon_i$. Thus we sum over $h_i$ in $H_i$, and over $i$ to yield $\Pr \leq \sum_i \frac{\delta}{2^{|H_i|}} \leq \delta$.

- Therefore, with probability $\geq 1 - \delta$, for all $i$, and for all $h_i \in H_i$, $|e\hat{r}(h_i) - err(h_i)| \leq \epsilon_i$. 


Assuming this condition holds, the overall quality of the chosen hypothesis can be shown to be close to optimal. Recall that $h^*$ is the actual best hypothesis, and let the least of the $H_i$’s containing it be $H_M$. Recall that $\hat{h}_i$ is the hypothesis with the smallest empirical error in $H_i$ and assume that the algorithm chooses index $k$ and returns $\hat{h}_k$.

- Assuming $\forall h \in H, |err(h) - e\hat{r}(h)| \leq \epsilon_i$ we have that $err(h^*) > e\hat{r}(h^*) - \epsilon_M$.
- Note that by definition of $\hat{h}_M$, we have $\forall h \in H_M, e\hat{r}(\hat{h}_M) \leq e\hat{r}(h)$. This implies that $err(h^*) \geq e\hat{r}(\hat{h}_M) - \epsilon_M$.
- We next note that $[e\hat{r}(\hat{h}_k) + \epsilon_k] \leq [e\hat{r}(\hat{h}_M) + \epsilon_M]$ because SRM chooses the hypothesis which minimizes $e\hat{r}(h_i) + \epsilon_i$.
- Therefore $err(h^*) \geq e\hat{r}(\hat{h}_M) - \epsilon_M = [e\hat{r}(\hat{h}_M) + \epsilon_M] - 2\epsilon_M \geq [e\hat{r}(\hat{h}_k) + \epsilon_k] - 2\epsilon_M \geq err(\hat{h}_k) - 2\epsilon_M$ where in the last step we used again the approximation guarantee (this time for $\hat{h}_k$).

We have therefore shown the following result.

**Theorem:** Fix sample size $m$. $\forall D, \forall \epsilon, \forall \delta$, and $\forall c$, given a hypothesis class $H$ let $h^*$ be the best hypothesis in $H$ and let $M$ be the index of the least set $H_M$ that includes $h^*$. Then with probability $1 - \delta$ the output $hyp$ of the SRM algorithm satisfies $err(hyp) \leq err(h^*) + 2\epsilon_M$.

The statement of the previous theorem does not provide a concrete error bound $\epsilon$ since $M$ and $\epsilon_M$ depend on $h^*$ and therefore on $D$. Following [SB] we can rephrase the guarantee in a slightly more pleasing manner as follows.

Fix any $h \in H$ and let $i$ be the least index where $h \in H_i$. We next require that $2\epsilon_i = 2\sqrt{\frac{1}{2m} \ln \frac{4|H_i|^2}{\delta^2}} \leq \epsilon$ which is equivalent to

\[
(*) \quad m \geq \frac{2}{\epsilon^2} \ln \frac{4|H_i|i^2}{\delta}
\]

We therefore get that if $(*)$ holds then $err(hyp) \leq err(h) + \epsilon$. We can still view this as a bound for fixed $m$. This time the guarantee is limited. With sample size $m$ the output of SRM is guaranteed to complete with all $h$ for which $(*)$ holds.

**Theorem:** Fix sample size $m$ and hypothesis class $H$. $\forall D, \forall \epsilon, \forall \delta$, and $\forall c$, with probability $1 - \delta$, for all $h \in H$ that satisfy $(*)$, the output $hyp$ of the SRM algorithm satisfies $err(hyp) \leq err(h) + \epsilon$.

As $m$ increases SRM competes with larger and larger portions of $H$. The algorithm remains the same. The guarantee improves with increasing sample size.