CSE522, Winter 2011, Learning Theory Lecture 12 - 02/08/2011

VC Theory Conclusion / PAC-Bayes Intro

Lecturer: Ofer Dekel *Scribe: Galen Andrew*

1 VC Theory (Conclusion)

PAC learning ("Probably Approximately Correct" learning) is a way to define "learnable". An algorithm for selecting a hypothesis h from a class H should be probably approximately correct in the sense that with high probability (with probability $1-\delta$) it should be a good approximation to (achieve risk within ϵ of) the best hypothesis in the class.

The first works on PAC learning discussed the *realizable* case, in which there exists some $h^* \in H$ that corresponds to the true relation between x and y, that is, $\mathcal{D}$ draws x from some marginal distribution $\mathcal{D}_x$ over x and then the observed sample is $(x, h^*(x))$. In the context of linear classification, realizability is equivalent to separability. In the realizable case, the loss of the best hypothesis is zero, so bounding the *excess* risk of some hypothesis h is equivalent to bounding its actual risk.

Definition 1. *Let $\mathcal{X} \subseteq \mathbb{R}^d$ and let H be a set of functions $h : \mathcal{X} \mapsto \pm 1$. We say $\mathcal{A}$ is a* PAC-learning algorithm *for H if*

1. *$\mathcal{A}$ outputs a hypothesis $h \in H$ in time polynomial in the size of the training set S,*
2. *There exists a polynomial $m(\cdot,\cdot,\cdot)$ such that $\forall \mathcal{D} : \forall \epsilon > 0 : \forall \delta > 0 :$ with probability at least $(1-\delta)$ over samples of size $m(d, 1/\epsilon, 1/\delta)$, $\mathcal{A}$ returns some h with $\ell(h; \mathcal{D}) \leq \epsilon$.*

In the *agnostic model* we don't assume that any $h \in H$ corresponds to the true $\mathcal{D}$, so instead of bounding $\ell(h; \mathcal{D})$ we bound the excess risk $\ell(h; \mathcal{D}) - \min_{h' \in H} \ell(h'; \mathcal{D})$.

Definition 2. *We say that H is* PAC-learnable *if there exists a PAC-learning algorithm for H.*

We've already seen that if VC-dim$(H) < \infty$ then ERM is a PAC-learning algorithm for H (if it is poly-time). Now we will show a strong converse.

Proposition 3. *If VC-dim$(H) = \infty$ then H is not PAC-learnable.*

Proof. Assume VC-dim$(H) = \infty$. Then there exists some set $X = \{x_i\}_{i=1}^{\infty}$. For any set of i labels $y_{1:i}$, let $\mathcal{D}_i(y_{1:i})$ be the distribution that selects x uniformly from $x_1, \ldots, x_i$ and assigns the labels $y_1, \ldots, y_i$. Suppose there were some $\mathcal{A}$ that PAC-learns H. Given $\delta > 0$ and $\epsilon > 0$ with $\epsilon < 1/4$, let m be the sample size needed by $\mathcal{A}$ to achieve ϵ risk with probability $1-\delta$. Consider the distributions $\mathcal{D}_i(y_{1:i})$ for $i = 2m$. There exists some $y_{1:i}$ such that given $S \sim \mathcal{D}_i(y_{1:i})^m$, on expectation $h = \mathcal{A}(S)$ is correct on at most half of the remaining m points. Therefore its total risk is at least $1/4 > \epsilon$. □

2 PAC-Bayes (Introduction)

Recall: In Structural Risk Minimization (SRM) we have a nested sequence of hypothesis classes $H_1 \subseteq H_2 \subseteq H_3 \ldots$. Assume that the loss $\ell \in [0, c]$, and define the Rademacher complexity of the i^{th} hypothesis class to be $R_m(\ell \circ H_i) = \rho_i$ so $\rho_i \leq \rho_{i+1}$.

Proposition 4. *For all $\delta > 0$, with probability $(1-\delta)$ with respect to the choice of $S \sim \mathcal{D}^m$, for all i, for all $h \in H_i$,*

$$\ell(h; \mathcal{D}) \leq \ell(h; S) + \rho_i + c\sqrt{\frac{\log \frac{1}{\delta} + 2\log(1+i)}{2m}}$$

Proof. For each i, the bound holds with probability at least $1 - \frac{\delta}{i(i+1)}$ so using the union bound, it holds simultaneously for all i with probability $(1 - \delta)$. □

This gives us a way to encode prior information into our learning algorithm: put hypotheses that are more likely *a priori* into the earlier H_i in the sequence. If the data do not contradict our prior information, then some hypothesis $h \in H_i$ for small i achieves a small loss and the bound is tighter. Otherwise, we need to use h from a higher H_i, and the bound is weaker. The goal of PAC-Bayes is to extend this idea to more general forms of prior information, e.g., a smooth prior distribution over hypothesis, not "onion peels" of consecutive hypothesis classes.

In Bayesian learning, we start with a prior distribution $P(H)$ encoding our beliefs about how likely each hypothesis is prior to observing any data. Then we observe a sample $S \sim \mathcal{D}^m$ and use Bayes' rule to determine the posterior distribution $Q(H)$. Overloading the ℓ notation once again, we will write

$$\ell(Q; \mathcal{D}) \triangleq \mathsf{E}_{h \sim Q}[\ell(h; \mathcal{D})] = \mathsf{E}_{h \sim Q}[\mathsf{E}_{(x,y) \sim \mathcal{D}}[\ell(h; (x, y))]].$$

The natural question is, what use is a posterior distribution? I.e., how can we make predictions when our algorithm provides a distribution over hypotheses, and not a concrete h? The easiest solution is to use a randomized hypothesis called the Gibbs hypothesis: For each new input x, we sample an independent $h \sim Q$ and use it to predict $h(x)$. We will see that it is easiest to use PAC-Bayes to prove bounds on the risk of the Gibbs hypothesis. Another solution is to sample many $h_i \sim Q$ i.i.d. and output the majority vote. The majority vote classifier seems more reasonable and works better in practice, although all we can say theoretically is that it is not likely to be much worse than the Gibbs hypothesis.

Example: For $w \in \mathbb{R}^n$, define

$$h_w(x) = \begin{cases} +1 & \text{with probability } \frac{1}{Z} e^{\langle w, x \rangle} \\ -1 & \text{with probability } \frac{1}{Z} e^{-\langle w, x \rangle} \end{cases} \quad \text{where } Z = e^{\langle w, x \rangle} + e^{\langle -w, x \rangle}.$$

The prior P is a zero-mean Gaussian distribution over w with covariance $\sigma^2 I$: $P(h_w) \propto \exp(-||w||^2/\sigma^2)$. Think of the prior as encoding that "my model of the world" is that $\mathcal{D}$ samples $x_{1:m}$ i.i.d. from some marginal distribution over x, then samples $h \sim P$ and outputs $S = \{(x_i, h(x_i)\}_{i=1}^m$. Then the *likelihood* is

$$\Pr[y_{1:m} | h_w, x_{1:m}] = \prod_i \frac{1}{Z_i} e^{y_i \langle w, x_i \rangle} \propto \exp \sum_i y_i \langle w, x_i \rangle.$$

Using Bayes' rule

$$\Pr[A|B, C] = \frac{\Pr[B|A, C] \cdot \Pr[A|C]}{\Pr[B|C]},$$

we can form the posterior (note that the *evidence* $\Pr[y_{1:m}|x_{1:m}]$ is absorbed into the proportionality constant because it does not depend on h_w)

$$\begin{aligned} \Pr[h_w | y_{1:m}, x_{1:m}] &= \frac{\Pr[y_{1:m}|h_w, x_{1:m}] \cdot \Pr[h_w|x_{1:m}]}{\Pr[y_{1:m}|x_{1:m}]} \\ &\propto \left(\exp \sum_i y_i \langle w, x_i \rangle \right) \cdot \left(\exp - \frac{||w||^2}{\sigma^2} \right) \\ &\propto \exp \left(\sum_i y_i \langle w, x_i \rangle - \frac{||w||^2}{\sigma^2} \right) \end{aligned}$$

In the next lecture, we will see that the critical factor determining the complexity of the learning algorithm will become $\mathrm{KL}(Q||P)$, the Kullback-Liebler divergence from Q to P instead of the Rademacher complexity.