CSE522, Winter 2011, Learning Theory Lecture 3 - 01/11/2011

Finite Hypothesis Classes

Lecturer: Ofer Dekel *Scribe: Galen Andrew*

1 Corrected Definition of Expectation

The following definition of the expectation of a random variable subsumes the two separate definitions of expectation for discrete and continuous R.V.s given in introductory courses. It also applies to R.V.s that are neither discrete nor continuous.

Definition 1. *A random variable X is called* simple *if it takes a finite number of non-negative values.*

Definition 2. *If X is a simple R.V., and takes values $x_1, \ldots, x_k$, its expectation is defined as*

$$\mathsf{E}[X] = \sum_{i=1}^{k} Pr[X = x_i] x_i .$$

Definition 3. *If Z is any R.V. with $Z \geq 0$, we define*

$$\mathsf{E}[Z] = \sup_{X \leq Z} \mathsf{E}[X]$$

where the supremum is over all simple R.V.s $X \leq Z$. If the set of values on the r.h.s. is unbounded, we write $\mathsf{E}[Z] = \infty$.

And finally the definition for a general R.V.:

Definition 4. *If Z is any R.V., we define*

$$\mathsf{E}[Z] = \mathsf{E}[Z_+] - \mathsf{E}[-Z_-]$$

where $Z_+ = Z\mathbb{I}_{\{Z \geq 0\}}$ and $Z_- = Z\mathbb{I}_{\{Z \leq 0\}}$, provided $\mathsf{E}[Z_+] < \infty$ or $\mathsf{E}[-Z_-] < \infty$. Otherwise $\mathsf{E}[Z]$ is undefined.

2 Recap: Estimation

The empirical risk of hypothesis h on a sample $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ for a loss function ℓ is written as $\ell(h; S) \triangleq \frac{1}{m} \sum_{i=1}^{m} \ell(h; (x_i, y_i))$. It is easy to see that the empirical risk (on a sample of any size) is an unbiased estimator of the true risk:

$$\mathsf{E}[\ell(h; S)] = \frac{1}{m} \sum_{i=1}^{m} \mathsf{E}[\ell(h; (x_i, y_i)] = \frac{1}{m} \sum_{i=1}^{m} \ell(h; \mathcal{D}) = \ell(h; \mathcal{D}).$$

But unbiasedness doesn't tell us much about how accurate the estimator is: after all, the empirical risk on a sample of size 1 is an unbiased estimator, but we don't it expect to be very accurate in that case. For that, we need to look at the variance.

Definition 5. *The* variance *of an R.V. Z is defined as* $\operatorname{Var}(Z) = \mathsf{E}[(Z - \mathsf{E}[Z])^2] = \mathsf{E}[Z^2] - \mathsf{E}[Z]^2$.

Observation 6. *For any R.V. Z and constant α,* $\operatorname{Var} \alpha Z = \alpha^2 \operatorname{Var} Z$.

If m R.V.s are dependent, the variance of their sum might grow as severely as m^2; for example if $X_1 = X_2 = \cdots = X_n$ (which is as dependent as you can get) with $\forall i : \mathrm{Var}(X_i) = \sigma^2$, then $\mathrm{Var}(\sum_{i=1}^m X_i) = \mathrm{Var}(mX_1) = m^2 \mathrm{Var}(X_1) = m^2\sigma^2$, and the variance of their average does not decrease with m: $\mathrm{Var}(\frac{1}{m}\sum_{i=1}^m X_i) = \frac{1}{m^2}\mathrm{Var}(\sum_{i=1}^m X_i) = \sigma^2$. On the other hand, the variance of a sum m of independent R.V.s grows only *linearly* with m, due to the following property:

Proposition 7. *If $X_1, \ldots, X_n$ are independent,* $\mathrm{Var}(\sum_{i=1}^m X_i) = \sum_{i=1}^m \mathrm{Var}(X_i)$.

In particular, if the X_i are i.i.d. with variance σ^2, then $\mathrm{Var}(\sum_{i=1}^m X_i) = m\sigma^2$. Therefore by averaging independent samples, we can get a linear decrease in variance: $\mathrm{Var}(\frac{1}{m}\sum_{i=1}^m X_i) = \frac{1}{m^2}\mathrm{Var}(\sum_{i=1}^m X_i) = \frac{\sigma^2}{m}$. Now we can apply Chebyshev's inequality to show that the probability of a large deviation of the empirical risk from the true risk decreases linearly with the sample size when the samples are independent:

$$\Pr\left[|\ell(h;S) - \ell(h;\mathcal{D})| > \epsilon\right] \le \frac{\mathrm{Var}(\ell(h;S))}{\epsilon^2} = \frac{\sigma^2}{m\epsilon^2},$$

where σ^2 is the variance of the i.i.d. values $\ell(h;(x_i,y_i))$. Defining $\delta = \frac{\sigma^2}{m\epsilon^2}$, we often see this type of result written in the following equivalent form.

Proposition 8. *For any $\delta > 0$, with probability at least $(1-\delta)$, $|\ell(h;S) - \ell(h;\mathcal{D})| \le \sqrt{\frac{\sigma^2}{m\delta}}$.*

Remember that the sample space here is the space of possible test sets with m examples. The estimation is non-random, so the probabilities are with respect to the random draw of possible test sets. Note also that h is assumed fixed in this analysis, and does not vary with S.

3 McDiarmid and Hoeffding

With further assumptions on the distribution of $\ell(h;S)$ we can derive much stronger bounds on the probability of significant estimation error.

Definition 9. *If $A \subseteq \mathbb{R}$, a function $f : A^m \mapsto \mathbb{R}$ is said to have* d-bounded differences *if for any $z = (z_1, \ldots, z_m) \in A^m$, any $i \in 1\ldots m$ and any $z'_i \in A$, we have $|f(z) = f(z')| < d$, where $z' = (z_1, \ldots, z'_i, \ldots, z_m)$.*

Theorem 10 (McDiarmid)**.** *If $Z \in A^m \subseteq \mathbb{R}^m$ is a vector of independent R.V.s and $f : A^m \mapsto \mathbb{R}$ has d-bounded differences, then for any $\epsilon > 0$,*

$$Pr[f(Z) - \mathsf{E}[f(Z)] > \epsilon] \le \exp\left(-\frac{2\epsilon^2}{md^2}\right).$$

Corollary 11. *Under the same assumptions as in Theorem 10,*

$$Pr[|f(Z) - \mathsf{E}[f(Z)]| > \epsilon] \le 2\exp\left(-\frac{2\epsilon^2}{md^2}\right).$$

Proof. Apply the union bound to the events $\{f(Z) - \mathsf{E}[f(Z)] > \epsilon\}$ and $\{\mathsf{E}[f(Z)] - f(Z) > \epsilon\}$. □

Intuition about this theorem can be gained by imagining that the Z_i's are agents that try to trick the estimator by choosing their values to make $f(Z)$ far from $\mathsf{E}[f(Z)]$. If each individual Z_i can shift the value of f by at most d (d-bounded differences), and furthermore the Z_i's cannot collude (independence), then the probability that they might actually be able to produce a "bad" sample is very small.

A special case of McDiarmid's inequality is Hoeffding's inequality, in which we specify that $f(Z)$ is the empirical average of the $f(Z_i)$'s—$f(Z) = \frac{1}{m}\sum_{i=1}^m Z_i$—and A is the interval $[a,b]$ where $b - a = c$. Then f has $\frac{c}{m}$-bounded differences, so

$$\Pr\left[|f(Z) - \mathsf{E}[f(Z)]| > \epsilon\right] \le 2\exp\left(-\frac{2m\epsilon^2}{c^2}\right).$$

Define $\delta = 2\exp\left(-2m\epsilon^2/c^2\right)$. The theorem then relates three important quantities: the sample size m, the maximal deviation ϵ, and the maximal probability of error δ. We have seen for a given sample size m and deviation ϵ a bound on the probability of error δ. Equivalently, we can say that for a given m and δ, the best bound on the deviation we can prove with these tools is

$$\epsilon \geq c\sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

We could also discuss how many samples m are necessary to achieve maximal deviation ϵ with probability of error at most δ:

$$m \geq \frac{c^2 \log \frac{2}{\delta}}{2\epsilon^2}.$$

Note that the dependence on δ here is only logarithmic, meaning the required data size to achieve a given bound on the probability of error grows very slowly as the bound tightens. We can put these together in a table giving each of the three quantities in terms of the other two:

independent quantity	m	ϵ	δ
minimal value	$\frac{c^2 \log \frac{2}{\delta}}{2\epsilon^2}$	$c\sqrt{\frac{\log \frac{2}{\delta}}{2m}}$	$2\exp\left(-\frac{2m\epsilon^2}{c^2}\right)$

4 Finite hypothesis classes and ERM

We'd like to be able to say how close is the empirical risk of the ERM hypothesis to its true risk. First let's look at the *wrong* (read "incorrect") way to do that. Suppose we have a set of hypotheses H and a training set S, and we compute the empirical risk minimizer $h_{\text{ERM}} = \operatorname{argmin}_{h \in H} \ell(h; S)$. Then we might naively think that it follows from Theorem 10 that

$$|\ell(h_{\text{ERM}}; S) - \ell(h_{\text{ERM}}; \mathcal{D})| \leq c\sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

But there are two problems with this analysis. First, h_{ERM} is not a fixed hypothesis; it depends on S. The emprical risk of h_{ERM} is not even an unbiased estimator of its true risk. Second, the quantities $\ell(h_{\text{ERM}}; (x_i, y_i))$ and $\ell(h_{\text{ERM}}; (x_j, y_j))$ are no longer independent—consider that the sample (x_i, y_i) may affect the choice of h_{ERM} which then influences the value of $\ell(h_{\text{ERM}}; (x_j, y_j))$.

The simplest correct solution is to prove a *uniform* bound on H, that is, find a δ such that

$$\Pr\left[\forall h \in H\colon |\ell(h; S) - \ell(h; \mathcal{D})| \leq \epsilon\right] > 1 - \delta$$

or equivalently

$$\Pr\left[\exists h \in H\colon |\ell(h; S) - \ell(h; \mathcal{D})| > \epsilon\right] \leq \delta.$$

Then it doesn't matter that h_{ERM} depends on S, because our bound holds for all $h \in H$ uniformly; in particular it holds for whichever h ERM chooses.

If H is finite, we can do this quite simply by applying the union bound to the "bad" events $E_h = \Pr\left[|\ell(h; S) - \ell(h; \mathcal{D})| > \epsilon\right]$. If $|H| = k$, then for any $\epsilon > 0$ we have

$$\Pr\left[\exists h \in H\colon |\ell(h; S) - \ell(h; \mathcal{D})| > \epsilon\right] \leq \sum_{h \in H} \Pr\left[|\ell(h; S) - \ell(h; \mathcal{D})| > \epsilon\right] \leq 2k\exp\left(-\frac{2m\epsilon^2}{c^2}\right).$$

Making a table relating the three quantities as before, we have

independent quantity	m	ϵ	δ
minimal value	$\frac{c^2 \log(2k/\delta)}{2\epsilon^2}$	$c\sqrt{\frac{\log(2k/\delta)}{2m}}$	$2k\exp\left(-\frac{2m\epsilon^2}{c^2}\right)$

Note that m depends only logarithmically on k, so even with potentially very large finite hypothesis classes we can get an accurate estimate of $\ell(h_{\text{ERM}}; \mathcal{D})$ with high probability without huge amounts of data.

Definition 12. *For a hypothesis class H and some $h \in H$, the* excess risk *of h with respect to H is $\ell(h; \mathcal{D}) - \min_{h' \in H} \ell(h'; \mathcal{D})$.*

Theorem 13. *For hypothesis class $|H| = k$, loss function ℓ taking values in the interval $[0, c]$, and dataset $S \sim \mathcal{D}^m$, for all $\delta > 0$, with probability at least $(1 - \delta)$ the excess risk of the ERM hypothesis h_{ERM} is at most 2ϵ where*

$$\epsilon = c\sqrt{\frac{\log(2k/\delta)}{2m}}$$

Proof. Let $h^* = \operatorname{argmin}_{h \in H} \ell(h, \mathcal{D})$. Then,

$$\begin{aligned}
\ell(h^*; \mathcal{D}) &\geq \ell(h^*; S) - \epsilon && \text{(uniform bound applied to } h^*\text{)} \\
&\geq \ell(h_{\text{ERM}}; S) - \epsilon && (h_{\text{ERM}} \text{ minimizes } \ell(h_{\text{ERM}}; S)) \\
&\geq \ell(h_{\text{ERM}}; \mathcal{D}) - 2\epsilon && \text{(uniform bound applied to } h_{\text{ERM}}\text{)}
\end{aligned}$$

and therefore $\ell(h_{\text{ERM}}; \mathcal{D}) - \ell(h^*; \mathcal{D}) \leq 2\epsilon$. □

Note that Theorem 13 does not depend on the finiteness of H, only on the fact that a uniform bound has been proved over H. Later, when we prove uniform bounds for infinite H, we will have a similar bound on the excess risk.