10-704: Information Processing and Learning **Fall 2016**

Lecture 5: Sept 14

Lecturer: Aarti Singh

Note: *These notes are based on scribed notes from Spring15 offering of this course. LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

5.1 Estimation of information theoretic quantities

Information theory often assumes the source distribution or probability models are known (as they can often be designed in classical signal processing applications), however in modern signal processing and machine learning applications, these need to be estimated from data i.e. samples from the probability models. We saw an example of this in the clustering application where we estimated the conditional entropy by plugging-in estimates of the mean and variance. Here we revisit that idea and few other estimators. We start by discussing some estimations of entropy in discrete and continuous settings.

5.2 Discrete Setting

We have a distribution P supported on a finite alphabet $\{1, \dots, d\}$ with $P(X = j) = p_j$ ($\sum_{j=1}^d p_j = 1$). We observe independent samples $\{X\}_{i=1}^n \sim P$ and would like to estimate some functional of P, say the entropy:

$$H(P) = -\sum_{j=1}^{d} p_j \log_2(p_j) \tag{5.1}$$

5.2.1 Plugin Estimator

The classical estimator here is the **plugin** estimator. We use the sample to estimate the frequencies, $\hat{p}_j = \frac{1}{n}\sum_{i=1}^n \mathbf{1}[X_i = j]$ and simply plug in these frequencies to obtain:

$$\hat{H}_n = -\sum_{j=1}^{d} \hat{p}_j \log_2(\hat{p}_j) \tag{5.2}$$

Let us characterize the error rate of this estimator. Some of the relevant papers here are Basharin (1959); Antos et al. (2001).

Theorem 1.

$$\mathbb{E}\left(\hat{H}_n - H\right)^2 = O\left(\frac{\sigma^2}{n} + \frac{d^2}{n^2}\right) \tag{5.3}$$

where $\sigma^2 = \mathrm{Var}(-\log_2 p(X))$. In particular the plugin estimator achieves the parametric rate of $O(1/n)$ for low-dimensional (fixed d) settings.

Proof. For the whole proof, we'll actually work with the entropy defined using the natural log. At the end, multiply everything through by $\log_2(e)$ to change the base back to base two.

To bound the mean-squared error of the plug-in entropy estimator, we use the bias-variance decomposition.

$$\begin{aligned}\mathbb{E}\left(\hat{H}_n - H\right)^2 &= \mathbb{E}[(\hat{H}_n - \mathbb{E}\hat{H}_n)^2] + \left[\mathbb{E}\hat{H}_n - H\right]^2 + 2\mathbb{E}[(\hat{H}_n - \mathbb{E}\hat{H}_n)(\mathbb{E}\hat{H}_n - H)] \\ &= \mathbb{E}[(\hat{H}_n - \mathbb{E}\hat{H}_n)^2] + \left[\mathbb{E}\hat{H}_n - H\right]^2\end{aligned}$$

The first term is the **variance** of the estimator and characterizes the squared deviation of the estimator around its mean, the second term is square of the **bias** which characterizes how well the mean of the estimator approximates the true entropy. The third term vanishes since $\mathbb{E}\hat{H}_n$ and H are not random quantities, and hence moving the outer expectation inside yields $\mathbb{E}\hat{H}_n - \mathbb{E}\hat{H}_n = 0$.

We will show that:

$$|\mathbb{E}\hat{H}_n - H| = O(\frac{d}{n}) \qquad \mathrm{Var}(\hat{H}_n) = O(\frac{\sigma^2}{n} + \frac{d^2}{n^2})$$

The result then follows by plugging these bounds into the bias-variance decomposition shown above. Both of these bounds will be obtained by Taylor expanding $\hat{H}_n = H(\hat{p})$ around $H(p)$ and using standard facts about how $\hat{p}_j$ converges to p_j. Notice that $n\hat{p}_j$ is Binomial(n, p_j). We will use the following facts about the moments of binomial random variables (these are fairly easy to prove):

$$\begin{aligned}\mathbb{E}\hat{p}_j - p_j &= 0 \\ \mathbb{E}\left(\hat{p}_j - p_j\right)^2 &= \frac{p_j(1-p_j)}{n} \\ \mathbb{E}\left(\hat{p}_j - p_j\right)\left(\hat{p}_k - p_k\right) &= \frac{-p_j p_k}{n} \qquad j \neq k \\ \mathbb{E}\left(\hat{p}_j - p_j\right)^3 &= \frac{p_j - 3p_j^2 + 3p_j^3}{n^2}\end{aligned}$$

The fourth moment is also $\Theta(1/n^2)$. Recall that, for a function f, Taylor's expansion of degree m with remainder is given as: There exists a c between a and x such that

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^k(a)}{k!}(x-a)^k + \frac{f^{k+1}(c)}{(k+1)!}(x-a)^{k+1}$$

Using this, we have

$$\begin{aligned}\hat{H}_n = H(\hat{p}) = H(p) - \sum_{j=1}^{d}(\hat{p}_j - p_j)(1 + \log p_j) - \frac{1}{2}\sum_{j=1}^{d}\frac{(\hat{p}_j - p_j)^2}{p_j} + \frac{1}{6}\sum_{j=1}^{d}\frac{(\hat{p}_j - p_j)^3}{p_j^2} \\ - \frac{1}{12}\sum_{j=1}^{d}\frac{(\hat{p}_j - p_j)^4}{((1-\theta)p_j + \theta\hat{p}_j)^3}\end{aligned}$$

where $\theta \in [0, 1]$. For the bias, the first order term is annihilated. The second and third order terms are:

$$\begin{aligned}\frac{1}{2}\sum_{j=1}^{d}\mathbb{E}\frac{(\hat{p}_j - p_j)^2}{p_j} &= \frac{1}{2}\sum_{j=1}^{d}\frac{(1-p_j)}{n} = \frac{d-1}{2n} \\ \frac{1}{6}\sum_{j=1}^{d}\mathbb{E}\frac{(\hat{p}_j - p_j)^3}{p_j^2} &= \frac{1}{6n^2}\sum_{j=1}^{d}3p_j - 3 + \frac{1}{p_j} = O(\frac{d}{n^2})\end{aligned}$$

By the fact that $\frac{1}{((1-\theta)p_j+\theta\hat{p}_j)^3} \leq \frac{1}{(1-\theta)^3 p_j^3}$, the remainder term is also $O(d/n^2)$.

The variance analysis (or direct analysis of Mean Square Error using Taylor expansion) is quite tedious but it follows along similar lines. We present a rough argument below. Write:

$$\mathbb{E}\left[(\hat{H}_n - \mathbb{E}\hat{H}_n)^2\right] = \mathbb{E}\left[(\hat{H}_n - H + \frac{d-1}{2n} + O(d/n^2))^2\right]$$

and then take a third order taylor expansion of $\hat{H}_n$ around H to get:

$$\mathbb{E}\left[\left(-\sum_{j=1}^{d}(\hat{p}_j - p_j)(1+\log p_j) + \frac{d-1}{2n} - \frac{1}{2}\sum_{j=1}^{d}\frac{(\hat{p}_j-p_j)^2}{p_j} + \frac{1}{6}\sum_{j=1}^{d}\frac{(\hat{p}_j-p_j)^3}{(p_j(1-\theta)+\theta\hat{p}_j)^2} + O(d/n^2)\right)^2\right]$$

When you square, the key thing to note is that the term that matters is the first one (all other terms can be shown to be $O(d^2/n^2)$):

$$\begin{aligned}
\mathbb{E}\left[\left(-\sum_{j=1}^{d}(\hat{p}_j - p_j)(1+\log p_j)\right)^2\right] &= \sum_{j,k=1}^{d}(1+\log p_j)(1+\log p_k)\mathbb{E}(\hat{p}_j-p_j)(\hat{p}_k-p_k) \\
&= \frac{1}{n}\sum_{j=1}^{d}(1+\log p_j)^2 p_j(1-p_j) - \frac{1}{n}\sum_{j\neq k}(1+\log p_j)(1+\log p_j)p_j p_k \\
&= \frac{1}{n}\sum_{j=1}^{d}p_j \log^2 p_j - H^2 = \frac{1}{n}\sigma^2
\end{aligned}$$

All of the other terms are $O(d^2/n^2)$, and hence the variance and MSE bounds follow. □

5.3 Continuous Setting

Here, we assume the distribution P is continuous with density $p = dP/d\mu$, where μ is the Lebesgue measure. As before we obtain a sample $\{X_i\}_{i=1}^n \sim P$. We would like to estimate the differential entropy:

$$H(p) = -\int p(x)\log p(x) d\mu(x)$$

We discuss a few approaches.

5.3.1 Plugin Estimator

The plugin estimator is conceptually straightforward: estimate the density $\hat{p}$ using a parametric estimator such as a Gaussian with estimated means and variances, as discussed in last class, or a non-parametric estimator such as a kernel density estimator (KDE), nearest-neighbor estimator, or histogram [1] and plug this into the functional. Specifically with an estimator $\hat{p}$ we get:

$$\hat{H} = -\int \hat{p}(x)\log\hat{p}(x) d\mu(x)$$

[1] if you are unfamiliar with these estimators, see e.g. http://www.cs.cmu.edu/ aarti/Class/10701_Spring14/slides/nonparametric.pdf

We investigate rates of error convergence for the KDE based entropy estimator Liu et al. (2012). Recall that the density estimator $\hat{p}_h$ using a kernel K and bandwidth h is given by (e.g. $K(x) \propto \exp(-x^2)$, i.e. the Gaussian kernel.):

$$\hat{p}_h(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\|x - X_i\|}{h}\right)$$

The main assumption we will use is a **smoothness assumption** on the density, since an arbitrary non-smooth density cannot be estimated well from finite samples.

Definition 2. *The class of functions $\Sigma(\beta, L)$ is called the **Holder class**. A function $f : [0,1]^d \to \mathbb{R}$ is in $\Sigma(\beta, L)$ if for all tuples $(r_1, \ldots, r_d)$ where r_i is a non-negative integer with $\sum_i r_i \leq \lfloor \beta \rfloor$ (the largest integer strictly smaller than β) we have:*

$$|D^r f(x+u) - D^r f(x)| \leq L\|u\|^{\beta - |r|}$$

where $|r| = \sum_j r_j$ and D^r is the r^{th} derivative of f.

For example the class $\Sigma(2, L)$ is the class of functions with Lipschitz continuous derivatives. In this case, the assumption essentially says that small changes in the input only cause small changes in the function.

It is a classical result that if p has smoothness β (i.e. belongs to Holder, or Sobolev classes) then:

$$\mathbb{E}\hat{p}_h(x) - p(x) \asymp h^\beta \qquad \mathrm{Var}(\hat{p}_h(x)) \asymp \frac{1}{nh^d}$$

And the mean squared error is equal to the squared bias plus the variance. The MSE is minimized when the squared bias and variance are balanced, i.e. we choose $h \asymp n^{\frac{-1}{2\beta+d}}$ in which case the squared bias and variance, and hence the MSE, are all $O(n^{\frac{-2\beta}{2\beta+d}})$.

We decompose the error of the entropy estimate as

$$|H(\hat{p}_h) - H(p)| \leq \underbrace{|H(\hat{p}_h) - \mathbb{E}H(\hat{p}_h)|}_{\text{Variance}} + \underbrace{|\mathbb{E}H(\hat{p}_h) - H(p)|}_{\text{bias}}$$

Liu, Lafferty, and Wasserman Liu et al. (2012) show (and you will do it in Homework) that functional:

$$\mathbb{P}(|H(\hat{p}_h) - \mathbb{E}H(\hat{p}_h)| > \epsilon) \leq 2\exp\left(\frac{n\epsilon^2}{32\kappa^2}\right)$$
$$|\mathbb{E}H(\hat{p}_h) - H(p)| \leq c_1 h^\beta + \frac{c_3}{nh^d}$$

where κ, c_1, c_3 are constants. The bias term is optimized by setting $h \asymp n^{\frac{-1}{\beta+d}}$, so that the bias is $O(n^{\frac{-\beta}{\beta+d}})$. The concentration bound on the variance shows that variance-like term is $O(1/\sqrt{n})$ and putting these together gives the rate:

$$\mathbb{E}[|H(\hat{p}_h) - H(p)|] \leq O_P\left(\frac{1}{\sqrt{n}} + n^{\frac{-\beta}{\beta+d}}\right)$$

Notice that when $\beta > d$, the second term is asymptotically dominated by the first, leading to a parametric $O(1/\sqrt{n})$ convergence rate in this smooth regime.

The main point here is that we chose the bandwidth $h \asymp n^{\frac{-1}{\beta+d}}$ which is suboptimal for density estimation and infact *undersmooths* the density estimate (leading to higher variance than bias for density estimation).

However, integration in the entropy functional removes some of this variance, making a smaller bandwidth suitable for entropy estimation.

Some comments:

1. You can also use other forms of plugin estimators. For example, Singh and Poczos Singh et al. (2016) use and analyze k-NN based estimators. Interestingly, they show that you can consistently estimate these divergences by keeping k fixed with n. This is essentially undersmoothing, when k is fixed, you have lower bias than usual but higher variance. The k-NN estimator is:
$$\hat{p}_k(x) = \frac{1}{n}\frac{k}{\text{vol}(B_k(x))}$$
where B_k is the Euclidean ball centered at x that contains k samples nearest to x.

2. One downside is that tuning the hyperparameter is tough here. It is not clear what loss to use for doing something like cross-validation. The problem is that the hyperparameter is not the same one you would use for density estimation.

3. Another downside is that the existing analyses are highly specialized. The proofs for the kNN method are quite different from the techniques outlined here. This contrasts with Von-Mises analysis we will discuss next time, where we get to borrow a lot from existing density estimation results.

4. Yet another downside is computing the integral which requires numerical integration, we discuss alternatives below.

There are few other plug-in estimators that can be used. There are three different ways to obtain a plug-in estimator using a density estimate $\hat{p}(x)$:

1. **Integral Estimate**: $\hat{H}(x) = -\int \hat{p}(x)\log\hat{p}(x)dx$ (we discussed this above)

2. **Re-substitution Estimate**: $\hat{H}(x) = -\frac{1}{n}\sum_{i=1}^{n}\log\hat{p}(X_i)$ where $\hat{p}$ is obtained using the samples $\{X_1, ..., X_n\}$.

3. **Splitting Data Estimate**: $\hat{H}(x) = -\frac{1}{m}\sum_{i=1}^{m}\log\hat{p}(X_i)$ where $\hat{p}$ is obtained using the samples $\{X_{m+1}, ..., X_n\}$. A cross-validation estimate can be defined similarly, e.g. the leave-one-out estimate is given as $\hat{H}(x) = -\frac{1}{n}\sum_{i=1}^{n}\log\hat{p}_i(X_i)$ where $\hat{p}_i$ is obtained using all samples except X_i.

The key difference between the Re-substitution estimate and the Splitting Data Estimate is that the splitting estimate sums over different samples than the ones used for estimating the density $\hat{p}$.

We will analyze the resubstitution estimator next time using Von-Mises analysis.

References

Basharin, GP (1959). On a statistical estimate for the entropy of a sequence of independent random variables. *Theory of Probability & its Applications.*

Antos, A. & and Kontoyiannis, I. (2001). Convergence properties of functional estimates for discrete distributions. *Random Structures & Algorithms.*

Liu, Han and Wasserman, Larry and Lafferty, John D (2012). Exponential concentration for mutual information estimation with application to forests. *Advances in Neural Information Processing Systems.*

Singh, S. and Poczos, B (2016). Finite-Sample Analysis of Fixed-k Nearest Neighbor Density Functionals Estimators. *Advances in Neural Information Processing Systems.*