10-704: Information Processing and Learning **Spring 2012**

Lecture 14: Minimum Complexity Penalized Density Estimation

Lecturer: Aarti Singh *Scribes: Evangelos Papalexakis*

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.*

14.1 Review

Model Class Selection

Last time, we talked about Model Class Selection. We introduced a two-stage coding scheme where we 1) encode the class and 2) we encode the data given that class.

The method in order to do this is defined by the following minimization:

$$\hat{\gamma} = \arg\min_{\gamma \in \Gamma} \{L(\gamma) + L_{\gamma}(x^n)\}$$

where $L_{\gamma}(x^n)$ is the code-length computed with a representative distribution in Q_{γ} that minimizes univeral redundancy within the class Q_{γ}.

Result: Suppose the true (or best approximating) class is γ^*. If we knew the true class, the length of the universal code for true (or best approximating) class will be $L_{\gamma^*}(x^n)$. How far from this optimal value are we, if we use the two-stage process described above to select the model class? By definition, it follows that:

$$L(\hat{\gamma}) + L_{\hat{\gamma}}(x^n) \leq L(\gamma^*) + L_{\gamma^*}(x^n)$$

This result is interesting because it intuitively tells us that even if we don't know the true class, doing a two-step scheme, yields a negligible overhead of $L(\gamma^*)$ which is $O(\log\log\gamma^*)$, the length of a prefix code for encoding the integer γ^*.

Universal Model Selection, using a two-stage code

The aim here is to find the best model in a *given* class. This is done by solving the following minimization:

$$q_{\gamma} = \arg\min_{q \in Q_{\gamma}} \{L(q) + L_q(x^n)\} \tag{14.1}$$

A natural choice of $L_q(x^n) = \log 1/q(x^n)$, the Shannon information content or negative log likelihood loss suffered by using model q to represent data x^n. We now need to decide how to pick a prefix code $L(q)$, and the analysis will tell us how.

14.2 Minimum Complexity Density Estimation

A result by Barron and Cover [1], that we proved in last class, states that for the two-stage model selection procedure in Equation 14.1 and all $p \in \mathcal{P}$:

$$\mathbb{E}[R_{p,c^{\gamma}_{\text{2-stage}}}] \leq \min_{q \in Q_\gamma} \frac{1}{n}[L(q) + D_n(p||q)]$$

where $c^{\gamma}_{\text{2-stage}}$ denotes a two stage code designed for class γ. We are going to use this bound to come up with a good $L(q)$.

Instead of the set Q_γ (which may have infinitely many elements), we will consider the set $\overline{Q}_\gamma$ which is a *countable subset* of Q_γ. The bound presented above still holds, because we restrict ourselves to a smaller set, namely:

$$\min_{q \in Q_\gamma} \frac{1}{n}[L(q) + D_n(p||q)] \leq \min_{q \in \overline{Q}_\gamma} \frac{1}{n}[L(q) + D_n(p||q)]$$

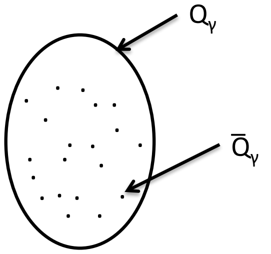

Figure 14.1: Here we see a pictorial example of the sets Q_γ and $\overline{Q}_\gamma$. The set Q_γ might be infinite, so we restrict ourselves to the dots inside, which form the countable subset of Q_γ, $\overline{Q}_\gamma$.

We also require for $\overline{Q}_\gamma$ that $\forall q \in \overline{Q}_\gamma$

$$L(q) = \log |\overline{Q}_\gamma|$$

i.e., all elements are equally likely and we can assign the same length code to all models. Thus $L(q)$ will be determined by the size of the subset $\overline{Q}_\gamma$ we choose. Therefore, the minimization problem of interest becomes:

$$\min_{g \in \overline{Q}_\gamma} \frac{1}{n}[L(q) + D_n(p||q)] = \frac{\log |\overline{Q}_\gamma|}{n} + \min_{q \in \overline{Q}_\gamma} \frac{1}{n} D_n(p||q)$$

In the above expression, the quantity $\frac{\log |\overline{Q}_\gamma|}{n}$ is the estimation error whereas $\min_{q \in \overline{Q}_\gamma} \frac{1}{n} D_n(p||q)$ is the approximation error. We are thus presented with a trade-off between these two quantities - choosing a sufficiently large class $\overline{Q}_\gamma$ allows us to approximate the original distribution $\mathcal{P}$ well but increases the estimation error (price for searching in a large subset), and vice versa. A good choice of $\overline{Q}_\gamma$ will balance these two terms and lead us to a good $L(q)$.

At this point, we will introduce two useful facts for the KL-divergence that simplify our analysis:

1. By definition,

$$D_n(p||q) = \mathbb{E}_p[\log \frac{p(x^n)}{q(x^n)}]$$

   Furthermore, if our process is iid, it follows that

$$D_n(p||q) = \sum_{i=1}^{n} D(p||q) = nD(p||q)$$

2. Any distributions $p(x)$ and $q(x)$, may be written in the following, exponential form:

$$p(x) \propto e^{f_p(x)}$$

   and

$$q(x) \propto e^{f_q(x)}$$

   where f_q, f_p are either parameters or functions, depending on whether our distributions are parametric or not. Then, we may bound the KL-divergence as follows:

$$D(p||q) \leq O(\|f_p - f_q\|^2)$$

   For a proof, see Lemma 1 of [2]. This relates the distance between distributions to distance between their parameters.

Using these properties, we are now going to provide two examples of selecting an optimal $\overline{Q}_\gamma$, one for the parametric case and one for the non-parametric case.

Examples

1. **Parametric case & connection to Regularized Maximum Likelihood**

   Let $\mathcal{P} = Q_\gamma$, Markov chain of order γ. First let $\gamma = 0$ (iid case). We are going to encode Q_γ. Notice that any model in Q_γ is specified by $\mathcal{X} - 1$ parameters corresponding to the probability of symbols (one less since probabilities must sum to 1). Parameters $q_\gamma = (p_1, p_2, \cdots, p_{|\mathcal{X}|-1})$. We quantize these parameters to levels of $\frac{1}{n}$, as in Figure 14.2.

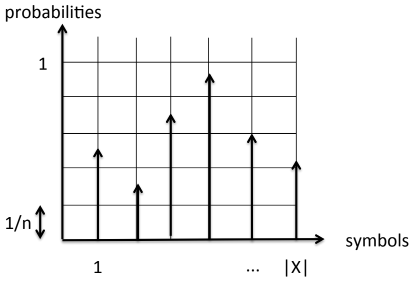

Figure 14.2: An example of quantized probabilities, into levels of $\frac{1}{n}$

   The set $\overline{Q}_\gamma$ is the collection of models with parameters quantized to $1/n$ and

$$|\overline{Q}_\gamma| = (n+1)^{|\mathcal{X}|-1}$$

(Note that if we had a general Markov chain with $\gamma > 0$, the size would be $(n+1)^{|\mathcal{X}|-1}(n+1)^{|\mathcal{X}|^\gamma}$)

Then, the worst-case expected redundancy is

$$\sup_{p\in\mathcal{P}} \text{Exp. red} \le O\left(\frac{|\mathcal{X}|^\gamma(|\mathcal{X}|-1)\log n}{n}\right) + O\left(\left(\frac{1}{n}\right)^2\right)$$

This is due to the fact that

$$D(p||q) \le O(\|f_p - f_q\|^2)$$

and the true parameters specifying p are $\frac{1}{n}$ apart from the parameters of the best approximating model in $\overline{Q}_\gamma$ due to the quantization.

Notice that the expected redundancy bound is of the same order as the minimax redundancy of a universal code for Markov-γ processes (as shown in last few lectures, where we used a universal predictor based on a mixture of models in $\mathcal{P}$ instead). Thus, using a two-stage code for model selection as given by Equation 14.1 with $L(q) = \log((n+1)^{|\mathcal{X}|-1}(n+1)^{|\mathcal{X}|^\gamma})$, we have been able to obtain a universal model for the Markov-γ process.

We now go back and refine our procedure to select both the model class as well as universal model within the model class automatically. Notice that for the universal model chosen using Equation 14.1, we have:

$$L_\gamma(x^n) = \log|\overline{Q}_\gamma| + \min_{q\in Q_\gamma} \log\frac{1}{q(x^n)}$$

Plugging this into the model selection procedure,

$$\hat{\gamma} = \arg\min_{\gamma\in\Gamma}\{\underbrace{L(\gamma) + \log|\overline{Q}_\gamma|}_{(1)} + \underbrace{\min_{q\in Q_\gamma}\log\frac{1}{q(x^n)}}_{(2)}\}$$

The above cost function is divided into two distinct parts:

(1) Encoding the class and its index. This can be viewed as a penalty or regularization.

(2) A Maximum Likelihood problem.

It turns out that we are essentially doing *regularized maximum likelihood.*

Having solved for $\hat{\gamma}$, we pick a model class as follows:

$$\hat{q}_{\hat{\gamma}} = \arg\min_{q\in Q_{\hat{\gamma}}} \log\frac{1}{q(x^n)}$$

In general, if we are dealing with parametric classes, the regularization penalty is going to be

$$\#\ \text{parameters}\cdot\log n$$

(as demonstrated above for Markov-γ processes).

2. **Non-parametric case**

In this example we are dealing with continuous alphabets. This means that instead of the symbols $x \in \mathcal{X}$, we have that (without loss of generality) $x \in [0,1]$. A motivating application for this example is encoding of measurements of a sensor.

We are going to use a histogram with γ bins, i.e. our $\overline{Q}_\gamma$ is going to be a histogram of γ bins, and each bin value will be quantized to levels of size $\frac{1}{n}$ as in the previous example. A pictorial example of $\overline{Q}_\gamma$ follows, in Figure 14.3.

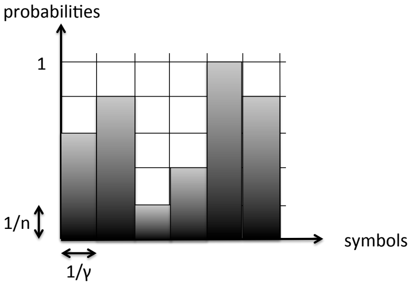

Figure 14.3: An example of $\overline{Q}_\gamma$ as a histogram, where we have quantized both the probabilities, into levels of $\frac{1}{n}$ and the symbols, into γ bins.

The size of $\overline{Q}_\gamma$ will now be

$$|\overline{Q}_\gamma| = \# \text{ values a param. can take}^{\#bins} = O\left((n+1)^\gamma\right)$$

Now, let's see how good the histogram is able to approximate our distribution. For this example, we will consider as $\mathcal{P}$ the class of smooth densities, i.e.

$$\mathcal{P} = \{p : |p(x) - p(y)| \leq L(|x-y|)\}$$

which intuitively means that the densities in $\mathcal{P}$ can't change dramatically (have a bounded gradient). This class is called *Lipschitz* class.

Remember that we have now quantized in both dimensions - symbol space and probabilites, and we expect this to be reflected in the error $\min_{q \in \overline{Q}_\gamma} D(p||q)$. In fact, we have that

$$\min_{q \in \overline{Q}_\gamma} D(p||q) = O(\|f_p - f_q\|^2) = O(\frac{L^2}{\gamma^2} + \frac{1}{n^2})$$

Then, the maximum expected redundancy is

$$\sup_{p \in \mathcal{P}} \text{Exp. red} \preceq \frac{\gamma \log n}{n} + \frac{1}{\gamma^2} + \frac{1}{n^2}$$

where the inequality $\preceq$ denotes that we are ignoring constants.

We now observe that the overhead $\frac{1}{\gamma^2} + \frac{1}{n^2}$ is no longer negligible, and in fact, there is a trade-off between the number of bins and the approximation error, as we would expect.

We would ideally need $\frac{\gamma \log n}{n} \approx \frac{1}{\gamma^2}$. Thus, the best number of bins we can use is $\gamma \approx (n/\log n)^{\frac{1}{3}}$, but in reality, we don't get to choose this; it is chosen automatically by our procedure. Therefore, our error scales as $\frac{1}{\gamma^2} + \frac{1}{n^2} \approx (n/\log n)^{-\frac{2}{3}}$, which is in fact the best convergence rate for estimating densities in this class.

In the case where we have more general classes than Lipschitz (where, for instance, we require smoothness for α derivatives, not just the first one), we can show that the error scales as $n^{\frac{-2\alpha}{2\alpha+d}}$, where d is the dimension and α is the number of derivatives we require to be "smooth". This requires considering models with polynomials of order α in each bin (instead of a constant value per bin as for histograms).

References

[1] A.R Barron. and T.M. Cover, ”Minimum complexity density estimation”, *Information Theory, IEEE Transactions on*, vol. 37, no. 4, pp 1034–1054, 1991, IEEE

[2] A. R. Barron and C.-H. Sheu, ”Approximation of Density Functions by Sequences of Exponential Families”, *The Annals of Statistics*, Vol. 19, No. 3, Sep., 1991, pp 1347-1369