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

Lecture 10: Universal coding and prediction

Lecturer: Aarti Singh *Scribes: Georg M. Goerg*

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.*

10.1 Universal coding and prediction

We want an encoder that works well for any distribution $p \in \mathcal{P}$. For example, to gzip a text file we would like an encoder that works for several different languages, rather than designing a new zip code for each single language.

Shannon/Huffman codes could achieve this, but their disadvantage is that generally we have to wait for the whole sequence to arrive before beginning to encode/decode. We would like to design a code that can be encoded and decoded immediately on a symbol per symbol basis.

In this lecture we will focus on the duality between optimal coding and optimal prediction. For example, it is natural to require a code C to have short length $L_C(x)$ for the symbol x; analogously in prediction we often want to minimize a loss function $loss_q(x)$ for a predictor q. Table 10.1 gives a general overview of several dual concepts and notions of a code C and a predictor q that we will use in this (and upcoming) lectures.

10.1.1 Weak and strong universality

A code C is universal if

$$\overline{R}_C = \sup_{p \in \mathcal{P}} \mathbb{E}_p\left(R_{p,C}\right) \overset{n \to \infty}{\to} 0. \tag{10.1}$$

A predictor q is universally consistent if

$$\overline{R}_q = \sup_{p \in \mathcal{P}} \mathbb{E}_p\left(R_{p,q}\right) \overset{n \to \infty}{\to} 0. \tag{10.2}$$

Just as standard calculus has pointwise and uniform convergence (of e.g. functions), we can also differentiate between weak and strong universality:

weakly universal if the convergence rate depends on $p \in \mathcal{P}$, e.g. if $\overline{R}_C = o(n^{-\gamma_p})$ and γ_p changes for every $p \in \mathcal{P}$.[1]

strong universal if the convergence rate is the same for all $p \in \mathcal{P}$, e.g. $\overline{R}_C = o(n^{-\gamma}) \, \forall p \in \mathcal{P}$.

By Kraft's inequality we know that for prefix codes

$$\sum_{x^n \in \mathcal{X}^n} 2^{-L_C(x^n)} \leq 1, \tag{10.3}$$

[1] The rate need not necessarily be polynomial; it just serves as an example.

data $x^n, x^n \sim p \in \mathcal{P}$			
code C		predictor q	
length of symbol x using code C	$L_C(x)$	$loss_q(x) = -\log q(x)$	negative log-likelihood or self-information loss
Properties per symbol		Average	
code length	$\frac{L_C(x^n)}{n}$	$-\frac{1}{n}\log q(x^n) \overset{\text{if } x^n \text{ iid}}{=} -\frac{1}{n}\sum_{i=1}^n \log q(x_i)$	empirical log-likelihood of the data
expected length	$\mathbb{E}_p\left(\frac{L_C(x)}{n}\right)$	$-\frac{1}{n}\mathbb{E}_p \log q(x^n)$	$\mathbb{E}$-loss = risk
ideal code length	$-\frac{1}{n}\log p(x^n)$ (Shannon information content)	$-\frac{1}{n}\log p(x^n)$	likelihood under the true model p
redundancy	$R_{p,C} = \frac{L_C(x^n)}{n} - \left(-\frac{1}{n}\log p(x^n)\right)$	$R_{p,q} = -\frac{1}{n}\log q(x^n) - \left(-\frac{1}{n}\log p(x^n)\right)$	excess loss of predictor q wrt true p
expected redundancy	$\mathbb{E}_p\left(R_{p,C}\right) = \mathbb{E}_p\left(\frac{L_C(x^n)}{n}\right) - \frac{H(x^n)}{n}$ ≥ 0 for uniquely decodable codes	$\mathbb{E}_p\left(R_{p,q}\right) = \frac{1}{n}\mathbb{E}_p \log \frac{p(x^n)}{q(x^n)} = \frac{1}{n}D_n(p\|\|q) \geq 0$	excess risk = $\mathbb{E}$ excess loss
minimax expected redundancy [a]	$\underbrace{\min_C \sup_{p\in\mathcal{P}} \mathbb{E}_p\left(R_{p,C}\right)}_{:=\overline{R}_C}$	$\underbrace{\min_q \sup_{p\in\mathcal{P}} \frac{1}{n} D_n(p\|\|q)}_{=:\overline{R}_q}$	minimax excess risk

Table 10.1: Universal coding and prediction: per symbol properties of a code C and a predictor q.

[a] We want that the code C works well for all $p \in \mathcal{P}$.

and we can construct the corresponding predictor by

$$q(x^n) = \frac{2^{-L_C(x^n)}}{\sum_{x^n \in \mathcal{X}^n} 2^{-L_C(x^n)}} = \underbrace{k}_{\geq 1} \cdot 2^{-L_C(x^n)} \Rightarrow L_c(x^n) \geq -\log q(x^n). \tag{10.4}$$

Similarly, for any distribution $q(x^n)$ we can define a corresponding prefix code that satisfies

$$L_c(x^n) \leq -\log q(x^n) + 1. \tag{10.5}$$

This must be true since we know that at least the Shannon code $\left(\left\lceil \log_2 \frac{1}{q(x^n)} \right\rceil\right)$ can achieve this.

For prefix codes (see (10.1)) we have

$$\overline{R}_C \geq \underbrace{\min_q \sup_{p \in \mathcal{P}} \frac{1}{n} D_n(p||q)}_{\arg\min \rightsquigarrow \overline{q}} =: \overline{R}. \tag{10.6}$$

Let $\overline{q}$ be the predictor that achieves the above minimum. We can then construct a Shannon code C^* from this predictor. By (10.5) we know that $\overline{R}_{C^*}$ will be within $1/n$ bit of $\overline{R}_C$ (or within 1 bit for the entire sequence).

10.1.2 Prediction problem

We have data x^n from $p \in \mathcal{P}$. How much loss do we suffer from using $q \neq p$ instead of the true p?

Here q can be any distribution; however, typically q is an estimate of p depending on the data x^n.

In general, we have to impose some restrictions on the class of distributions $\mathcal{P}$ to get universally consistent codes C / predictors q, i.e. to achieve error rates $\to 0$. For example, we often assume i) iid, ii) Markov chains, ...

For a specific q consider $** \geq \overline{R}_q \geq \overline{R} \geq *$. Typically we try to bound $**$ and $*$ to get control over the error rates.

Note that $\overline{q} = \arg\min_q \overline{R}_q$, where $\overline{R}_q$ is the worst excess risk for a particular q. $\overline{q}$ is the model/estimate q that minimizes the expected worst case scenario.

We will show later in the course that, in general, the optimal $\overline{q}$ is a mixture distribution over the class $p \in \mathcal{P}$. In other words, for any q, $\exists$ a mixture distribution p_{mix} such that the excess risk of q is always greater or equal to the excess risk of p_{mix}, i.e.

$$D_n(p||q) \geq D_n(p||p_{mix}). \tag{10.7}$$

10.1.3 Maximum loss instead of expected loss

Now instead of expected loss, consider the maximum loss (maximum over all possible sequences x^n)

$$R^* = \min_q \sup_{p \in \mathcal{P}} \max_{x^n} \frac{1}{n} \log \frac{p(x^n)}{q(x^n)} \tag{10.8}$$

Let $P_{ML}(x^n) = \sup_{p \in \mathcal{P}} p(x^n)$ (the MLE). Define the normalized ML as

$$NML(x^n) = \frac{P_{ML}(x^n)}{\sum_{x^n} P_{ML}(x^n)} = q^*, \tag{10.9}$$

under maximum-loss (instead of $\mathbb{E}$-loss).

The normalized maximum likelihood distribution is the best universal predictor under maximum loss.

Theorem 10.1 *For any class $\mathcal{P}$ of processes with finite alphabet*

$$q^* = NML(x^n) \text{ and } R^* = \log \sum_{x^n} P_{ML}(x^n). \tag{10.10}$$

For a proof, see pg 480 of Csiszar and Shield's tutorial.

10.1.3.1 Problems of NML and maximum loss for arithmetic coding

For arithmetic coding we need the conditional distribution $q(x^n \mid x^{n-1})$. But the NML distribution is not consistent in the sense that

$$q^*(x^n) \neq \sum_{x_{n+1}} q^*(x^{n+1}) \tag{10.11}$$

or equivalently

$$q^*(x_1, \ldots, x_n) \neq \sum_{x_{n+1}} q^*(x_1, \ldots, x_n, x_{n+1}) \tag{10.12}$$

Remark: The right hand side in the above expressions yields a valid distribution, but it is not the distribution of $x_1, \ldots, x_n$ under q^*. This can be seen by recalling the definition of q^*.
Thus it is not possible to define a corresponding arithmetic code (Shannon and Huffman codes are possible though).

Thus we return to consider $\mathbb{E}$-loss as in this case we know that q is a mixture distribution - and this is consistent in the sense that

$$q(x_1, \ldots, x_n) = \sum_{x^{n+1}} q(x_1, \ldots, x_n, x_{n+1}). \tag{10.13}$$

Examples of model classes and their optimal codes/predictors

1. $\mathcal{P}$ is the class of iid processes with finite alphabet $\mathcal{X}$. It can be shown that the optimal predictor is given by

$$q(x^n) = \prod_{i=1}^{n} \frac{n(x_i \mid x^{i-1}) + \frac{1}{2}}{i - 1 + \frac{|\mathcal{X}|}{2}}, \tag{10.14}$$

$$n(x_i \mid x^{i-1}) = \text{ \# of ocurrences of symbol } x_i \text{ in past } x^{i-1} \, . \tag{10.15}$$

   The term $\frac{n(x_i|x^{i-1})}{i-1}$ is simply the frequency of symbols observed before time t; the additional $\frac{+\frac{1}{2}}{+\frac{|\mathcal{X}|}{2}}$ smoothes out the ML estimate. Thus it avoids assigning 0 probability to symbols that have not occurred yet (but may occur in the future).

Let n_x be the number of times the symbol x occurred in the entire length n sequence. Then (10.14) can be rewritten as (see next lecture)

$$q(x^n) = \frac{\prod_{x\in\mathcal{X}}(n_x - \frac{1}{2})(n_x - \frac{3}{2})\cdots\frac{1}{2}}{\left(n-1+\frac{|\mathcal{X}|}{2}\right)\left(n-2+\frac{|\mathcal{X}|}{2}\right)\cdots\frac{|\mathcal{X}|}{2}} \sim \sum_{p\in\mathcal{P}} \pi(p)\cdot p(x^n), \tag{10.16}$$

where $\pi(p) \sim \mathit{Dirichlet}\left(\frac{1}{2},\ldots,\frac{1}{2}\right)$.[2]

One can show that (for proof see next lecture)

$$\overline{R}_q \le R_q^* \le \underbrace{\frac{|\mathcal{X}|-1}{2}\frac{\log n}{n}}_{\text{best possible bound}} + \frac{\mathit{constant}}{n}. \tag{10.17}$$

2. $\mathcal{P}$ is the class of Markov processes of order 1.

   Let $n_{i-1}(k,j)$ be the count of how many times the sequence (k,j) appeared in the first $i-1$ symbols $(x_1,\ldots,x_{i-1})$; also let $n_{i-1}(k) = \sum_j n_{i-1}(k,j)$ be the total number of times the symbol k occurred in the first $i-1$ symbols.

$$q(x^n) = \prod_{i=1}^{n} q(j \mid x^{i-1}), \quad q(j \mid x^{i-1}) = \frac{n_{i-1}(k,j)+\frac{1}{2}}{n_{i-1}(k)+\frac{|\mathcal{X}|}{2}}. \tag{10.18}$$

   For a Markov process of order $m=1$ one can show (see next lecture)

$$\overline{R}_q \le R_q^* \le \underbrace{\frac{|\mathcal{X}|\,(|\mathcal{X}|-1)}{2}\frac{\log n}{n}}_{\text{best possible bound}} + \frac{\mathit{constant}}{n}. \tag{10.19}$$

3. $\mathcal{P}$ is the class of Markov processes of order m, i.e. x_i depends on previous m steps. Then

$$q(j \mid x^{i-1}) = \frac{\#\text{ times j occured preceeded by } x_{i-m}^{i-1} + \frac{1}{2}}{\#\text{ times } x_{i-m}^{i-1} \text{ occured} + \frac{|\mathcal{X}|}{2}}. \tag{10.20}$$

   Here it holds

$$\overline{R}_q \le R_q^* \le \underbrace{\frac{|\mathcal{X}|^m\,(|\mathcal{X}|-1)}{2}\frac{\log n}{n}}_{\text{best possible bound}} + \frac{\mathit{constant}_m}{n}. \tag{10.21}$$

   Again, see the next lecture for detailed derivations.

[2] https://en.wikipedia.org/wiki/Dirichlet_distribution