Universal Prediction [Merhav and Feder, 1998]

Can the future of a sequence be predicted based on its past?

If so, how well?

Knowledge of how the past and future are related is unavailable or inaccurate

so need methods of universal prediction.

A universal predictor is one that doesn't depend on the unknown underlying (statistical) model and yet performs essentially as well as if the model were known in advance.

Information-theoretic view of prediction goes back to Shannon who related prediction to estimating entropy of printed English.

Kelly showed equivalence between gambling and information [gambling is form of prediction].

There is strong relation between universal lossless source coding and universal prediction.

An observer sequentially receives a sequence of observations $x_1, x_2, \ldots, x_t, \ldots$ over an alphabet $\mathcal{X}$. At each time t, having seen $x^{t-1} = (x_1, \ldots, x_{t-1})$ but not x_t, the observer predicts the next outcome x_t, or more generally makes a decision b_t based on observed past x^{t-1}.

There is a loss function $\ell(b_t, x_t)$ that measures quality as compared to what actually happened, x_t.

The predictive objective could be to minimize instantaneous loss, its time-average, or expected value of either of these.

$\ell(b_t = \hat{x}_t, x_t) = (x_t - \hat{x}_t)^2$ is a common loss function in continuous case, Hamming in discrete case.

Another common case is to have confidence/reliability associated, so b_t is conditional probability assignment for x_t given x^{t-1}, i.e. $b_t(\cdot \mid x^{t-1})$ that sums to one for each x^{t-1}.

Here the loss function ℓ should decrease monotonically with probability assigned to actual outcome $b_t(x_t \mid x^{t-1})$

An example is the self-information loss (also called log-loss).

For every probability assignment $b = \{b(x), x \in \mathcal{X}\}$ over $\mathcal{X}$ and every $x \in \mathcal{X}$, this function is defined as

$$\ell(b, x) = -\log b(x).$$

Classical statistical signal processing assumes a known source distribution P, so to minimize expected loss: b_t^* chosen to minimize

$$\mathbb{E}\left[\ell(b, X_t) \mid X^{t-1} = x^{t-1}\right] = \int_{\mathcal{X}} dP(x \mid x^{t-1})\, \ell(b, x).$$

Given $X^{t-1} = x^{t-1}$,

$$U(x^{t-1}) = \inf_b \int dP(x | x^{t-1}) \, \ell(b, x)$$

is called the conditional Bayes envelope given x^{t-1}!

If $\{X_t\}$ is binary source, $b_t = \hat{x}_t$ and $\ell(\cdot,\cdot)$ is Hamming distance,

$$b_t^* = \begin{cases} 0 & \text{if } P(0 | x^{t-1}) \geq P(1 | x^{t-1}) \\ 1 & \text{else} \end{cases}$$

and conditional Bayes envelope is

$$U(x^{t-1}) = \min \{ P(0 | x^{t-1}), P(1 | x^{t-1}) \}.$$

If $\ell(b, x) = (b - x)^2$, then

$$b_t^* = \mathbb{E}[X_t | X^{t-1} = x^{t-1}]$$

$$U(x^{t-1}) = \mathrm{var}[X_t | X^{t-1} = x^{t-1}].$$

In log-loss case,

$b_t^*(\cdot | x^{t-1}) = P(\cdot | x^{t-1})$ minimizes $\mathbb{E}[-\log b(X_t | X^{t-1} = x^{t-1})]$ so

best probability assignment is true one, and

$U(x^{t-1}) = -\mathbb{E} \log P(X_t | X^{t-1} = x^{t-1})$ is (differential) entropy of X_t given X^{t-1}.

Now what if P unknown or non-existent.

Probabilistic setting.   deterministic setting

Probabilistic Setting.

A universal predictor $\{b_t^u(x^{t-1})\}$ does not depend on P and keeps the difference between

$$E\left[\frac{1}{n}\sum_{t=1}^{n} \ell(b_t^u, X_t)\right] \quad \text{and}$$

$$\bar{U}_n(P) = \frac{1}{n}\sum_{t=1}^{n} E[U(X^{t-1})] = \frac{1}{n}\sum_{t=1}^{n} E\left[\inf_b E[\ell(b,X_t) \mid X^{t-1}]\right]$$

vanishingly small for large n.

performance of optimal predictor tuned to P.

$\bar{U}(P) = \lim_{n\to\infty} \bar{U}_n(P)$ exists for stationary, ergodic source.

In log-loss case, $\bar{U}(P)$ is entropy rate of P, so goal of universal prediction is equivalent to universal coding.

- Universality with respect to Indexed Class of Sources:

  $\{P_\theta, \theta \in \Lambda\}$ e.g. k-th order Markov Sources or AR(p) Gaussian source

- Universality with respect to very large class of sources

  e.g. Markov of unknown finite order

  stationary and ergodic.

Deterministic Setting

observed sequence not assumed to be drawn randomly by some probability law but rather an individual, deterministic sequence.

Two difficulties in even defining universal prediction problem.

① goal: for a given sequence $x_1, \ldots$ there is always a prediction function $b_t(x^{t-1}) = x_t$ so seemingly trivial.

② adversary: for a given deterministic predictor $\{b_t(\cdot)\}_{t \geq 1}$, there is always the adversary sequence where at each time t, x_t is chosen to maximize $\ell(b_t, x_t)$.

To avoid the fundamental overfitting problem in ①, must limit class B of allowed predictors $\{b_t(\cdot)\}_{t \geq 1}$.

For example B could be Markov-structured predictors
$b_t(x^{t-1}) = b(x_{t-k}, \ldots, x_{t-1})$.

so want universal predictor that is independent of future, so

average loss $\frac{1}{n} \sum_{t=1}^{n} \ell(b_t^u, x_t)$ is asymptotically same as

$$\min_{B} \frac{1}{n} \sum_{t=1}^{n} \ell(b_t, x_t) \quad \text{for every } x^n.$$

universal predictor need not be in B but must be causal.
reference predictor can depend on future too.

To alleviate ②, allow randomization.