Today: Universal Models/Codes

1. **Simple** MDL Model Selection
   – Three interpretations
2. The Fourth Type of Universal Code/Model: **prequential**
   – Fourth interpretation simple MDL mod.sel.
3. Questions/Feedback

MDL Model Selection

Select $\mathcal{M}_j$ minimizing $-\log \bar{p}_{\mathrm{nml}}(x^n \mid \mathcal{M}_j)$, i.e. minimizing

$$-\log p_{\hat{\theta}_j(x^n)}(x^n) + \log \sum_{x^n \in \mathcal{X}^n} p_{\hat{\theta}_j(x^n)}(x^n)$$

error (= minus fit) term **complexity** term ("log $|\Theta|$")

- select model that compresses data most, *treating all distributions within model on equal footing*;
- selected model detects most (non-spurious) regularity in data

Simple Refined MDL Model Selection

- Suppose we are given data $x^n = (x_1, \ldots, x_n)$
- We want to select between models $\mathcal{M}_1$ and $\mathcal{M}_2$ as explanations for the data. MDL tells us to pick the $\mathcal{M}_i$ for which the associated optimal universal model $\bar{p}_{\mathrm{nml}}(\cdot \mid \mathcal{M}_i)$ assigns the largest probability to the data:

$$\mathcal{M}_{\mathsf{mdl}} = \arg \sup_{j \in \{1,2\}} \bar{p}_{\mathsf{nml}}(x^n \mid \mathcal{M}_j)$$

MDL Model Selection

Select $\mathcal{M}_j$ minimizing $-\log \bar{p}_{\text{nml}}(x^n \mid \mathcal{M}_j)$, i.e. minimizing

$$- \log p_{\hat{\theta}_j(x^n)}(x^n) + \log \sum_{x^n \in \mathcal{X}^n} p_{\hat{\theta}_j(x^n)}(x^n)$$

error (= minus fit) term **complexity** term (“log |Θ|”)

(this is just ‘MDL model selection between two simple models’; it is not ‘the MDL Principle’)

Four Interpretations

- Compression interpretation
- Counting/Geometric interpretation
- Bayesian interpretation
- Predictive interpretation

MDL Model Selection, Regular Parametric Models

Select $\mathcal{M}_j$ minimizing $-\log \bar{p}_{\text{nml}}(x^n \mid \mathcal{M}_j)$, i.e. minimizing

$$-\log p_{\hat{\theta}_j(x^n)}(x^n) + \frac{k}{2} \log \frac{n}{2\pi} + \log \int_{\Theta_j} \sqrt{\det I(\theta)}\, d\theta + o(1)$$

error (= minus fit) term

Generic complexity term

Geometric, model specific complexity term

Regular Parametric Models

Select $\mathcal{M}_j$ minimizing $-\log \bar{p}_{\mathrm{nml}}(x^n \mid \mathcal{M}_j)$, i.e. minimizing

$$-\log p_{\hat{\theta}_j(x^n)}(x^n) + \frac{k}{2} \log \frac{n}{2\pi} + \log \int_{\Theta_j} \sqrt{\det I(\theta)}\, d\theta + o(1)$$

error (= minus fit) term

Generic complexity term

Geometric, model specific complexity term

compare to BIC/"old" MDL (Rissanen 1978):

$$\mathrm{BIC}(j) = -\log p_{\hat{\theta}_j(x^n)}(x^n) + \frac{k}{2} \log \frac{n}{2\pi}$$

Bayesian Model Selection

- Recall the Bayesian universal model

$$\bar{p}_{\text{Bayes}}(x^n|\mathcal{M}_j) = \int_{\theta\in\Theta_j} p(x^n|\theta)w(\theta)d\theta$$

- Bayesian model selection between $\mathcal{M}_1$ and $\mathcal{M}_2$ tells us to select the $\mathcal{M}_i$ maximizing

$$w(j \mid x^n) = \frac{\bar{p}_{\text{Bayes}}(x^n \mid \mathcal{M}_j)w(j)}{\sum_{k\in\{0,1\}} \bar{p}_{\text{Bayes}}(x^n \mid \mathcal{M}_k)w(k)}$$

- with uniform prior W this is the $\mathcal{M}_j$ maximizing

$$\bar{p}_{\text{Bayes}}(x^n \mid \mathcal{M}_j)$$

MDL vs Bayesian Model Selection, Regular Parametric Models

MDL: select $\mathcal{M}_j$ minimizing $-\log \bar{p}_{\text{nml}}(x^n \mid \mathcal{M}_j) =$

$$-\log p_{\hat{\theta}_j(x^n)}(x^n) + \frac{k}{2} \log \frac{n}{2\pi} + \log \int_{\Theta_j} \sqrt{\det I(\theta)}\, d\theta + o(1)$$

BAYES: select $\mathcal{M}_j$ minimizing $-\log p_{\text{Bayes}}(x^n \mid \mathcal{M}_j) =$

$$-\log p_{\hat{\theta}_j(x^n)}(x^n) + \frac{k}{2} \log \frac{n}{2\pi} - \log w(\hat{\theta}_j(x^n)) + \log \sqrt{\det I(\hat{\theta}(x^n))} + o(1)$$

- Always within O(1) ; hence, for large enough n, Bayes and MDL (and BIC) select the same model
- For Jeffreys' prior even within o(1)

Four Interpretations

- Compression interpretation
- Counting/Geometric interpretation
- Bayesian interpretation
- Predictive interpretation

Universal Prediction

- Suppose data arrives sequentially in time.
- Let $\mathcal{M}$ be a set of predictors. There exist prediction strategies that, for each data sequence that can possibly be realized, predict essentially as well as the predictor in $\mathcal{M}$ that turns out to be best for that sequence 'with hindsight'

On-Line "Probabilistic" Prediction

- Consider sequence $(x_1, y_1), (x_2, y_2), \ldots$ where all $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$

- Goal: sequentially predict y_i ,
  - given past $(x_1, y_1), \ldots, (x_{i-1}, y_{i-1})$
  - using 'probabilistic prediction' P_i (distribution on $\mathcal{Y}$)

On-Line Probabilistic Prediction

- Consider sequence $(x_1, y_1), (x_2, y_2), \ldots$ where all $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$
- Goal: sequentially predict y_i ,
  - given past $(x_1, y_1), \ldots, (x_{i-1}, y_{i-1})$
  - using 'probabilistic prediction' P_i (distribution on $\mathcal{Y}$)
- Example: **weather forecaster**

$\mathcal{Y} = \{0, 1\}$ (0 = no rain, 1 = rain)

$$\mathcal{X} = \left\{ \begin{array}{c} \text{gigantic vector indicating humidity,} \\ \text{air pressure temperature etc. at} \\ \text{various locations} \end{array} \right.$$

Prediction Strategies

- prediction strategy S is function mapping, for all i, histories $(x_1, y_1), \ldots, (x_{i-1}, y_{i-1})$ to distributions for i -th outcome

$$S : \cup_{n=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^n \rightarrow \text{set of distributions on } \mathcal{Y}$$

- Weather forecasting example:
  - Prediction strategy is simply the prediction algorithm used by the weather forecaster, hopefully designed by meteorologists
  - Prediction for y_i will depend on data observed on previous days $(x_{i-1}, y_{i-1}), (x_{i-2}, y_{i-2}), \ldots$

Universal Prediction

- Suppose we have two weather forecasters
  - Marjon de Hond (Dutch public TV)
  - Peter Timofeeff (Dutch commercial TV)
- On each i (day), Marjon and Peter announce the probability that $y_{i+1} = 1$,i.e. that it will rain on day $i+1$

Universal Prediction

- Suppose we have two weather forecasters
  - Marjon de Hond
  - Peter Timofeeff
- On each i (day), Marjon and Peter announce the probability that $y_{i+1} = 1$,i.e. that it will rain on day $i+1$
- We would like to combine their predictions in some way such that for every sequence $y_1, \ldots, y_n \in \{0,1\}^n$ we predict almost as well as whoever turns out to be the best forecaster for that sequence

Universal Prediction

- We would like to combine predictions such that for every sequence $y_1, \ldots, y_n \in \{0, 1\}^n$ we predict almost as well as the best forecaster for that sequence
- Surprisingly, there exist prediction strategies that achieve this. These are called **universal**
  - "universal" is really a misnomer
- To formalize this idea, we need to define how we measure prediction quality
  - i.e., what do we mean by "the best forecaster"

Logarithmic Loss

- To compare performance of different prediction strategies, we need a measure of prediction quality
- A standard quality measure is the log loss:

$$\mathrm{loss}(y, P) := -\log_2 P(y)$$

$$\mathrm{loss}(y_1 \ldots, y_n, S) := \sum_{i=1}^{n} \mathrm{loss}(y_i, S(y_1, \ldots, y_{i-1}))$$

- Why log-loss? Because…
  - …it's mathematically convenient
  - …it makes universal prediction equivalent to universal coding
  - …it has a gambling interpretation
  - … it's the only **local proper scoring rule**

Universal prediction with log loss

- We would like to combine predictions such that for every sequence $y_1, \ldots, y_n \in \{0, 1\}^n$ we predict almost as well as the best forecaster for that sequence
- It turns out that there exists a universal strategy $\bar{\mathrm{S}}$ such that, for all $n, y_1, \ldots, y_n \in \{0, 1\}^n$

$$\mathsf{loss}(y_1 \ldots, y_n, \bar{\mathrm{S}}) \leq$$

$$\min\{\mathsf{loss}(y_1 \ldots, y_n, S_{\mathsf{Marjon}}), \mathsf{loss}(y_1 \ldots, y_n, S_{\mathsf{Peter}})\} + 1.$$

Universal prediction with log loss

- We would like to combine predictions such that for every sequence $y_1, \ldots, y_n \in \{0,1\}^n$ we predict almost as well as the best forecaster for that sequence
- It turns out that there exists a universal strategy $\bar{\mathrm{S}}$ such that, for all $n, y_1, \ldots, y_n \in \{0,1\}^n$

$$\mathsf{loss}(y_1 \ldots, y_n, \bar{\mathrm{S}}) \leq$$

$$\min\{\mathsf{loss}(y_1 \ldots, y_n, S_{\mathsf{Marjon}}), \mathsf{loss}(y_1 \ldots, y_n, S_{\mathsf{Peter}})\} + 1.$$

- Losses increase linearly in n so this is very good!

$$\mathsf{loss}(y_1 \ldots, y_n, S) := \sum_{i=1}^{n} \mathsf{loss}(y_i, S(y_1, \ldots, y_{i-1}))$$

On-Line Probabilistic Prediction

- Consider sequence $y_1, y_2, \cdots$, all $y_i \in \mathcal{Y}$
- Goal: sequentially predict y_i given past $y_1, \ldots, y_{i-1}$ using a 'probabilistic prediction' P_i (distribution on $\mathcal{Y}$)
- prediction strategy S is function mapping, for all i, 'histories' $y_1, \ldots, y_{i-1}$ to distributions for i -th outcome

$$S : \cup_{n=1}^{\infty} \mathcal{Y}^n \rightarrow \text{set of distributions on } \mathcal{Y}$$

prediction strategy = distribution

- If we think that $Y_1, \ldots, Y_n \sim P$ (not necessarily i.i.d !) then we should predict Y_i using the conditional distribution

$$P(\cdot \mid y^{i-1}) := P(Y_i = \cdot \mid Y_1 = y_1, \ldots, Y_{i-1} = y_{i-1})$$

- Conversely, every prediction strategy S may be thought of as a distribution on $(Y_1, \ldots, Y_n)$, by defining:

$$P(\cdot \mid y^{i-1}) := S(y^{i-1})$$

$$P(y_1, \ldots, y_n) := \prod_{i=1}^{n} P(y_i \mid y^{i-1})$$

prediction strategy = distribution

- If we think that $Y_1, \ldots, Y_n \sim P$ (not necessarily i.i.d !) then we should predict Y_i using the conditional distribution

$$P(\cdot \mid y^{i-1}) := P(Y_i = \cdot \mid Y_1 = y_1, \ldots, Y_{i-1} = y_{i-1})$$

- Conversely, every prediction strategy S may be thought of as a distribution on $(Y_1, \ldots, Y_n)$, by defining:

$$P(\cdot \mid y^{i-1}) := S(y^{i-1})$$

$$P(y_1, \ldots, y_n) := \prod_{i=1}^{n} P(y_i \mid y^{i-1})$$

Log loss & likelihood

- For every "prediction strategy" P, all n,

$$\sum_{i=1}^{n} \text{loss}(y_i, P(\cdot \mid y^{i-1})) = \sum_{i=1}^{n} -\log P(y_i \mid y^{i-1}) = -\log P(y_1, \ldots, y_n)$$

$$\sum_{i=1}^{n} -\log P(y_i \mid y^{i-1}) = -\log \prod_{i=1}^{n} P(y_i \mid y^{i-1}) = -\log \prod \frac{P(y_i)}{P(y^{i-1})} = -\log P(y_1, \ldots, y_n)$$

Log loss & likelihood

- For every "prediction strategy" P, all n,

$$\sum_{i=1}^{n} \text{loss}(y_i, P(\cdot \mid y^{i-1})) = \sum_{i=1}^{n} -\log P(y_i \mid y^{i-1}) = -\log P(y_1, \ldots, y_n)$$

Accumulated log loss = minus log likelihood

Dawid '84, Rissanen '84

Universal Prediction

- Let $\mathcal{M} = \{P_1, P_2, \ldots\}$ be a for now, finite or countable set of predictors (identified with probability distributions on $\mathcal{Y}^\infty$)

- GOAL: given $\mathcal{M}$, construct a new predictor predicting data 'essentially as well' as any of the $P_\theta \in \mathcal{M}$

A Bayesian Strategy

- One possibility is to act Bayesian:

1. Put some prior W on (parameter space of) $\mathcal{M}$
2. Define Bayesian marginal distribution

$$P_{\text{Bayes}}(y_1, \ldots, y_n) := \sum_{\theta=1}^{\infty} P_\theta(y_1, \ldots, y_n) W(\theta)$$

3. Predict with Bayesian (posterior) predictive distribution

$$P_{\text{Bayes}}(y_{i+1} \mid y_1, \ldots, y_i) = \frac{P_{\text{Bayes}}(y_1, \ldots, y_{i+1})}{P_{\text{Bayes}}(y_1, \ldots, y_i)}$$

Evaluating Bayes

- For arbitrary strategies P :

$$\sum_{i=1}^{n} \mathsf{loss}(y_i, P(\cdot \mid y^{i-1})) = \sum_{i=1}^{n} - \log P(y_i \mid y^{i-1}) = - \log P(y_1, \ldots, y_n)$$

Evaluating Bayes

- For arbitrary strategies P :

$$\sum_{i=1}^{n} \text{loss}(y_i, P(\cdot \mid y^{i-1})) = \sum_{i=1}^{n} -\log P(y_i \mid y^{i-1}) = -\log P(y_1, \ldots, y_n)$$

- Moreover, for Bayes strategy P_{Bayes} ,for all n, y^n ,all θ_0:

$$\sum_{i=1}^{n} \text{loss}(y_i, P_{\text{Bayes}}(\cdot \mid y^{i-1})) = -\log P_{\text{Bayes}}(y_1, \ldots, y_n)$$

$$= -\log \sum_{\theta=1}^{\infty} P_\theta(y_1, \ldots, y_n) W(\theta) \leq \underbrace{-\log P_{\theta_0}(y_1, \ldots, y_n)}_{\text{linear increase in n}} \underbrace{-\log W(\theta_0)}_{\text{constant in n}}$$

Bayesian strategy is universal

- For all n, y^n, all θ :

$$\sum_{i=1}^{n} \text{loss}(y_i, P_{\text{Bayes}}(\cdot \mid y^{i-1})) \leq$$

$$-\log P_\theta(y_1, \ldots, y_n) + C_\theta = \sum_{i=1}^{n} \text{loss}(y_i, P_\theta(\cdot \mid y^{i-1})) + C_\theta$$

- For all sequences of each length n, total loss of Bayes strategy bounded by constant depending on θ, not on n (Marjon vs. Peter: $w(\theta) = \frac{1}{2}, C_\theta = -\log w(\theta) = 1$)

Prequential Interpretation of Universal "Coding"

- In the prequential view, the regret obtained by $\bar{p}$ on sequence x^n is just the difference between the cumulative prediction error (as measured by log-loss) made by sequentially predicting with $\bar{p}$ and sequentially predicting using p_θ with $\theta = \hat{\theta}(x^n)$, the $\theta \in \Theta$ that is 'prediction-optimal with hindsight'
- $-\log \bar{p}_{\mathrm{Bayes}}(x^n) = \sum_{i=1..n} \mathrm{loss}(x_i; \bar{p}_{\mathrm{Bayes}}(\cdot|x^{i-1}))$

(predict by smoothed maximum likelihood)

but also

$$-\log \bar{p}_{\mathrm{nml}}(x^n) = \sum_{i=1..n} \mathrm{loss}(x_i; \bar{p}_{\mathrm{nml}}(\cdot|x^{i-1}))$$

Simple Refined MDL Model Selection

- Suppose we are given data $x^n = (x_1, \dots, x_n)$
- We want to select between models $\mathcal{M}_1$ and $\mathcal{M}_2$ as explanations for the data. MDL tells us to pick the $\mathcal{M}_i$ for which the associated optimal universal model $p_{\text{nml}}(\cdot | \mathcal{M}_i)$ assigns the largest probability to the data:

$$\mathcal{M}_{\text{mdl}} = \arg \sup_{j \in \{1,2\}} p_{\text{nml}}(x^n \mid \mathcal{M}_j)$$

Simple Refined MDL Model Selection

- Suppose we are given data $x^n = (x_1, \ldots, x_n)$
- We want to select between models $\mathcal{M}_1$ and $\mathcal{M}_2$ as explanations for the data. MDL tells us to pick the $\mathcal{M}_i$ for which the associated optimal universal model $\bar{p}_{\mathrm{nml}}(\cdot \mid \mathcal{M}_i)$ assigns the largest probability to the data:

$$\mathcal{M}_{\mathsf{mdl}} = \arg \sup_{j \in \{1,2\}} \bar{p}_{\mathsf{nml}}(x^n \mid \mathcal{M}_j)$$

Simple Refined
MDL Model Selection

- Suppose we are given data $x^n = (x_1, \ldots, x_n)$
- We want to select between models $\mathcal{M}_1$ and $\mathcal{M}_2$ as explanations for the data. MDL tells us to pick the $\mathcal{M}_i$ for which the associated optimal universal model $\bar{p}_{\mathrm{nml}}(\cdot | \mathcal{M}_i)$ assigns the largest probability to the data:

$$\mathcal{M}_{\mathsf{mdl}} = \arg \sup_{j \in \{1,2\}} \bar{p}_{\mathsf{nml}}(x^n \mid \mathcal{M}_j)$$

Simple Refined
MDL Model Selection

- Suppose we are given data $x^n = (x_1, \dots, x_n)$
- We want to select between models $\mathcal{M}_1$ and $\mathcal{M}_2$ as explanations for the data. MDL tells us to pick the $\mathcal{M}_i$ for which the associated optimal universal model $p_{\text{nml}}(\cdot \mid \mathcal{M}_i)$ assigns the largest probability to the data:

$$\mathcal{M}_{\text{mdl}} = \arg \sup_{j \in \{1,2\}} p_{\text{nml}}(x^n \mid \mathcal{M}_j)$$

MDL Model Selection

Select $\mathcal{M}_j$ minimizing $-\log \bar{p}_{\mathrm{nml}}(x^n \mid \mathcal{M}_j)$, i.e. minimizing

$$- \log p_{\widehat{\theta}_j(x^n)}(x^n) + \log \sum_{x^n \in \mathcal{X}^n} p_{\widehat{\theta}_j(x^n)}(x^n)$$

error (= minus fit) term **complexity** term (“log |Θ|”)

(this is just ‘MDL model selection between two simple models’; it is not ‘the MDL Principle’)

Four Interpretations

- Compression interpretation
- Counting/Geometric interpretation
- Bayesian interpretation
- Predictive interpretation

MDL Model Selection: Prequential Interpretation

Select $\mathcal{M}_j$ minimizing $-\log \bar{p}_{\mathrm{nml}}(x^n \mid \mathcal{M}_j)$, i.e. minimizing

$$\sum_{j=1}^{n} \mathrm{loss}(x_i; \bar{p}_{\mathrm{nml}}(\cdot \mid x^{i-1}, \mathcal{M}_j))$$

- i.e. for each model, sequentially predict the data points based on past data using a prediction strategy based on the model. Then, select the model with the smallest cumulative (equivalently, average) loss.
- **Viewed in this way, MDL is quite similar to leave-one-out cross validation!**

  (MDL = "forward" rather than "cross" validation)

Four Interpretations

- Compression interpretation
- Counting/Geometric interpretation
- Bayesian interpretation
- Predictive interpretation

The Prequential Universal Model

- Now recall that with Bayesian universal model, predictive distr. given by "smoothed ML estimator"

$$\bar{p}_{\text{Bayes}}(x_i \mid x^{i-1}) = \int p_\theta(x_i) w(\theta \mid x^n) d\theta$$

- $w(\theta|x^n)$ approximately normal with mean $\hat{\theta}(x^n)$, variance $O(\frac{1}{\sqrt{n}})$. So predictions quite close to what you would get if you would directly predict with ML estimator based on the past!
  - Very visible in Bernoulli model with Jeffreys prior:

$$\bar{p}_{\text{Bayes}}(x_{n+1} \mid x^n) = \frac{n_1 + 1/2}{n + 1}$$

The Prequential Universal Model

$$\bar{p}_{\text{Bayes}}(x_i \mid x^{i-1}) = \int p_\theta(x_i) w(\theta \mid x^n) d\theta$$

- $w(\theta|x^n)$ approximately normal with mean $\hat{\theta}(x^n)$, variance $O(\frac{1}{\sqrt{n}})$. So predictions quite close to what you would get if you would directly predict with ML estimator based on the past!
- **IDEA:** define new distribution

$$\bar{p}_{\text{preq}}\left(x_i \mid x^{i-1} \right) \coloneqq p_{\hat{\theta}\left(x^{i-1}\right)}(x_i) \; ;$$

$$\bar{p}_{\text{preq}}(x^n) := \prod_{i=1..n} \bar{p}_{\text{preq}}\left(x_i \mid x^{i-1} \right)$$

The Prequential Universal Model

IDEA: define new distribution

$$\bar{p}_{\text{preq}}\left(x_i \mid x^{i-1} \right) := p_{\hat{\theta}(x^{i-1})}(x_i) \; ;$$

$$\bar{p}_{\text{preq}}(x^n) := \prod_{i=1..n} \bar{p}_{\text{preq}}\left(x_i \mid x^{i-1} \right)$$

This will behave essentially like a universal model

The Prequential Universal Model

$$\bar{p}_{\text{preq}}(x_i \mid x^{i-1}) := p_{\hat{\theta}(x^{i-1})}(x_i) \;\; ; \;\; \bar{p}_{\text{preq}}(x^n) := \prod_{i=1}^{n} \bar{p}_{\text{preq}}(x_i \mid x^{i-1})$$

Theorem: if $\hat{\theta}$ is 'slightly modified' ML estimator for "regular" k-dim. parametric model $\mathcal{M} = \{ p_\theta : \theta \in \Theta \}$, then for all θ in ineccsi subset of Θ,

$$\mathbf{E}_{X^n \sim P_\theta} \left[- \log \bar{p}_{\text{preq}}(X^n) - \left(- \log p_{\hat{\theta}(X^n)}(X^n) \right) \right] = \frac{k}{2} \log n + O(1)$$

… so the prequential distribution $\bar{p}_{\mathbf{preq}}$ is a "O(1)-universal model in expectation"

Prequential Universal Model

Advantage: often easier to calculate than $\bar{p}_{\mathrm{Bayes}}$

- aside: MDL model selection now becomes even more similar to cross-validation!

Prequential Universal Model

Disadvantages:

– does **not** achieve individual sequence minimax regret

–unnatural order dependence

– start-up problem: using the plain ML estimator often does not work, since it can result in infinite loss

- Solution 1: smooth it (add ‘virtual data’)

[gives same as $\bar{p}_{\text{Bayes}}$ for multinomial model, but not for other models]

- Solution 2: do a ‘late start’ (ignore data until smallest n such that this cannot happen any more)

Today: Universal Models/Codes

1. **Simple** MDL Model Selection
   – Three interpretations
2. The Fourth Type of Universal Code/Model: **prequential**
   – Fourth interpretation simple MDL mod.sel.
3. Questions/Feedback