How to Predict with Bayes, MDL, and Experts

Marcus Hutter

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale
IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland
marcus@idsia.ch, http://www.idsia.ch/~marcus

MLSS-2005, January 24 – February 4

Overview

- Setup: Given (non)iid data $D = (x_1, ..., x_n)$, predict x_{n+1}
- Ultimate goal is to maximize profit or minimize loss
- Consider Models/Hypothesis $H_i \in \mathcal{M}$
- Max.Likelihood: $H_{best} = \arg\max_i p(D|H_i)$ (overfits if $\mathcal{M}$ large)
- Bayes: Posterior probability of H_i is $p(H_i|D) \propto p(D|H_i)p(H_i)$
- MDL: $H_{best} = \arg\min_{H_i}\{\text{CodeLength}(D|H_i) + \text{CodeLength}(H_i)\}$ (Complexity penalization)
- Bayes needs prior(H_i), MDL needs CodeLength(H_i)
- Occam+Epicurus: High prior for simple models with short codes.
- Kolmogorov/Solomonoff: Quantification of simplicity/complexity
- MDL & Bayes work if D is sampled from $H_{true} \in \mathcal{M}$
- Prediction with Expert Advice works w/o assumption on D.

Table of Contents

Philosophical Issues: Contents

- On the Foundations of Machine Learning
- Example 1: Probability of Sunrise Tomorrow
- Example 2: Digits of a Computable Number
- Example 3: Number Sequences
- Occam's Razor to the Rescue
- Foundations of Induction
- Dichotomies in Machine Learning
- Sequential/online Prediction – Setup

Philosophical Issues: Abstract

I start by considering the philosophical problems concerning machine learning in general and induction in particular. I illustrate the problems and their intuitive solution on various (classical) induction examples. The common principle to their solution is Occam's simplicity principle. Based on Occam's and Epicurus' principle, Bayesian probability theory, and Turing's universal machine, Solomonoff developed a formal theory of induction. I describe the sequential/online setup considered in this lecture series and place it into the wider machine learning context.

On the Foundations of Machine Learning

- Example: Algorithm/complexity theory: The goal is to find fast algorithms solving problems and to show lower bounds on their computation time. Everything is rigorously defined: algorithm, Turing machine, problem classes, computation time, ...
- Most disciplines start with an informal way of attacking a subject. With time they get more and more formalized often to a point where they are completely rigorous. Examples: set theory, logical reasoning, proof theory, probability theory, infinitesimal calculus, energy, temperature, quantum field theory, ...
- Machine learning: Tries to build and understand systems that learn from past data, make good prediction, are able to generalize, act intelligently, ... Many terms are only vaguely defined or there are many alternate definitions.

Example 1: Probability of Sunrise Tomorrow

What is the probability $p(1|1^d)$ that the sun will rise tomorrow?
(d = past # days sun rose, 1 =sun rises. 0 = sun will not rise)

- p is undefined, because there has never been an experiment that tested the existence of the sun $tomorrow$ (ref. class problem).
- The $p = 1$, because the sun rose in all past experiments.
- $p = 1 - \epsilon$, where ϵ is the proportion of stars that explode per day.
- $p = \frac{d+1}{d+2}$, which is Laplace rule derived from Bayes rule.
- Derive p from the type, age, size and temperature of the sun, even though we never observed another star with those exact properties.

Conclusion: We predict that the sun will rise tomorrow with high probability independent of the justification.

Example 2: Digits of a Computable Number

- Extend 14159265358979323846264338327950288419716939937?
- Looks random?!
- Frequency estimate: $n =$ length of sequence. $k_i =$ number of occured $i \Longrightarrow$ Probability of next digit being i is $\frac{i}{n}$. Asymptotically $\frac{i}{n} \to \frac{1}{10}$ (seems to be) true.
- But we have the strong feeling that (i.e. with high probability) the next digit will be 5 because the previous digits were the expansion of π.
- Conclusion: We prefer answer 5, since we see more structure in the sequence than just random digits.

Example 3: Number Sequences

Sequence: $\begin{array}{cccccc} x_1, & x_2, & x_3, & x_4, & x_5, & \ldots \\ 1, & 2, & 3, & 4, & ?, & \ldots \end{array}$

- $x_5 = 5$, since $x_i = i$ for $i = 1..4$.
- $x_5 = 29$, since $x_i = i^4 - 10i^3 + 35i^2 - 49i + 24$.

Conclusion: We prefer 5, since linear relation involves less arbitrary parameters than 4th-order polynomial.

Sequence: 2,3,5,7,11,13,17,19,23,29,31,37,41,43,47,53,59,?

- 61, since this is the next prime
- 60, since this is the order of the next simple group

Conclusion: We prefer answer 61, since primes are a more familiar concept than simple groups.

On-Line Encyclopedia of Integer Sequences:

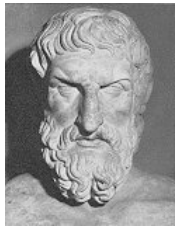
http://www.research.att.com/~njas/sequences/

Occam's Razor to the Rescue

- Is there a unique principle which allows us to formally arrive at a prediction which
  - coincides (always?) with our intuitive guess -or- even better,
  - which is (in some sense) most likely the best or correct answer?
- Yes! Occam's razor: Use the simplest explanation consistent with past data (and use it for prediction).
- Works! For examples presented and for many more.
- Actually Occam's razor can serve as a foundation of machine learning in general, and is even a fundamental principle (or maybe even the mere definition) of science.
- Problem: Not a formal/mathematical objective principle.
  What is simple for one may be complicated for another.

Foundations of Induction

Ockhams' razor (simplicity) principle
Entities should not be multiplied beyond necessity.



Epicurus' principle of multiple explanations
If more than one theory is consistent with the observations, keep all theories.

Bayes' rule for conditional probabilities
Given the prior belief/probability one can predict all future probabilities.

Turing's universal machine
Everything computable by a human using a fixed procedure can also be computed by a (universal) Turing machine.

Solomonoff's universal prior=Ockham+Epicurus+Bayes+Turing
Solves the question of how to choose the prior if nothing is known.
$\Rightarrow$ universal induction, formal Occam, AIT,MML,MDL,SRM,...

Problem Setup

- Every induction problem can be phrased as a sequence prediction task.
- Classification is a special case of sequence prediction.
  (With some tricks the other direction is also true)
- This lecture focusses on maximizing profit (minimizing loss).
  We're not (primarily) interested in finding a (true/predictive/causal) model.
- Separating noise from data is not necessary in this setting!

Dichotomies in Machine Learning

scope of my lecture	⇔	scope of other lectures
online learning	⇔	offline/batch learning
passive prediction	⇔	active learning
Bayes ⇔ MDL	⇔	Expert
noninformative and universal prior	⇔	informed or problem-specific prior
conceptual/mathematical issues	⇔	computational issues
exact/principled	⇔	heuristic
supervised learning	⇔	unsupervised ⇔ RL learning
exploitation	⇔	exploration

Sequential/online predictions

In sequential or online prediction, for times $t = 1, 2, 3, ...$,

our predictor p makes a prediction $y_t^p \in \mathcal{Y}$

based on past observations $x_1, ..., x_{t-1}$.

Thereafter $x_t \in \mathcal{X}$ is observed and p suffers $\text{Loss}(x_t, y_t^p)$.

The goal is to design predictors with small total loss or cumulative $\text{Loss}_{1:T}(p) := \sum_{t=1}^{T} \text{Loss}(x_t, y_t^p)$.

Applications are abundant, e.g. weather or stock market forecasting.

Example:

$\text{Loss}(x,y)$	$\mathcal{X} = \{\text{sunny}, \text{rainy}\}$	
$\mathcal{Y} = \left\{ \begin{matrix} \text{umbrella} \\ \text{sunglasses} \end{matrix} \right\}$	$\begin{matrix} 0.1 \\ 0.0 \end{matrix}$	$\begin{matrix} 0.3 \\ 1.0 \end{matrix}$

Setup also includes: Classification and Regression problems.

Bayesian Sequence Prediction: Contents

Bayesian Sequence Prediction: Abstract

The aim of probability theory is to describe uncertainty. There are various sources and interpretations of uncertainty. We compare the frequency, objective, and subjective probabilities, and show that they all respect the same rules. We derive Bayes' and Laplace's famous and fundamental rules and present two brain-teasing paradoxes. Then we concentrate on general sequence prediction tasks. We define the Bayes mixture distribution and show that the posterior converges rapidly to the true posterior by exploiting some bounds on the relative entropy. Finally we show that the mixture predictor is also optimal in a decision-theoretic sense w.r.t. any bounded loss function.

Uncertainty and Probability

The aim of probability theory is to describe uncertainty.

Sources/interpretations for uncertainty:

- Frequentist: probabilities are relative frequencies.
  (e.g. the relative frequency of tossing head.)
- Objectivist: probabilities are real aspects of the world.
  (e.g. the probability that some atom decays in the next hour)
- Subjectivist: probabilities describe an agent's degree of belief.
  (e.g. it is (im)plausible that extraterrestrians exist)

Frequency Interpretation: Counting

- The frequentist interprets probabilities as relative frequencies.
- If in a sequence of n independent identically distributed (i.i.d.) experiments (trials) an event occurs $k(n)$ times, the relative frequency of the event is $k(n)/n$.
- The limit $\lim_{n\to\infty} k(n)/n$ is defined as the probability of the event.
- For instance, the probability of the event head in a sequence of repeatedly tossing a fair coin is $\frac{1}{2}$.
- The frequentist position is the easiest to grasp, but it has several shortcomings:
- Problems: definition circular, limited to i.i.d, reference class problem.

Objective Interpretation: Uncertain Events

- For the objectivist probabilities are real aspects of the world.
- The outcome of an observation or an experiment is not deterministic, but involves physical random processes.
- The set Ω of all possible outcomes is called the sample space.
- It is said that an event $E \subset \Omega$ occurred if the outcome is in E.
- In the case of i.i.d. experiments the probabilities p assigned to events E should be interpretable as limiting frequencies, but the application is not limited to this case.
- (Some) probability axioms:
  $p(\Omega) = 1$ and $p(\{\}) = 0$ and $0 \leq p(E) \leq 1$.
  $p(A \cup B) = p(A) + p(B) - p(A \cap B)$.
  $p(B|A) = \frac{p(A \cap B)}{p(A)}$ is the probability of B given event A occurred.

Subjective Interpretation: Degrees of Belief

- The subjectivist uses probabilities to characterize an agent's degree of belief in something, rather than to characterize physical random processes.
- This is the most relevant interpretation of probabilities in AI.
- We define the plausibility of an event as the degree of belief in the event, or the subjective probability of the event.
- It is natural to assume that plausibilities/beliefs $\mathrm{Bel}(\cdot|\cdot)$ can be repr. by real numbers, that the rules qualitatively correspond to common sense, and that the rules are mathematically consistent. $\Rightarrow$
- Cox's theorem: $\mathrm{Bel}(\cdot|A)$ is isomorphic to a probability function $p(\cdot|\cdot)$ that satisfies the axioms of (objective) probabilities.
- Conclusion: Beliefs follow the same rules as probabilities

Bayes' Famous Rule

Let D be some possible data (i.e. D is event with $p(D) > 0$) and $\{H_i\}_{i \in I}$ be a countable complete class of mutually exclusive hypotheses (i.e. H_i are events with $H_i \cap H_j = \{\}$ $\forall i \neq j$ and $\bigcup_{i \in I} H_i = \Omega$).

Given: $p(H_i) =$ a priori plausibility of hypotheses H_i (subj. prob.)

Given: $p(D|H_i) =$ likelihood of data D under hypothesis H_i (obj. prob.)

Goal: $p(H_i|D) =$ a posteriori plausibility of hypothesis H_i (subj. prob.)

$$\text{Solution:} \quad p(H_i|D) = \frac{p(D|H_i)p(H_i)}{\sum_{i \in I} p(D|H_i)p(H_i)}$$

Proof: From the definition of conditional probability and

$$\sum_{i \in I} p(H_i|...) = 1 \quad \Rightarrow \quad \sum_{i \in I} p(D|H_i)p(H_i) = \sum_{i \in I} p(H_i|D)p(D) = p(D)$$

Example: Bayes' and Laplace's Rule

Assume data is generated by a biased coin with head probability θ, i.e. H_θ :=Bernoulli(θ) with $\theta \in \Theta := [0,1]$.

Finite sequence: $x = x_1 x_2 ... x_n$ with n_1 ones and n_0 zeros.

Sample infinite sequence: $\omega \in \Omega = \{0,1\}^\infty$

Basic event: $\Gamma_x = \{\omega : \omega_1 = x_1, ..., \omega_n = x_n\}$ = set of all sequences starting with x.

Data likelihood: $p_\theta(x) := p(\Gamma_x | H_\theta) = \theta^{n_1}(1-\theta)^{n_0}$.

Bayes (1763): Uniform prior plausibility: $p(\theta) := p(H_\theta) = 1$

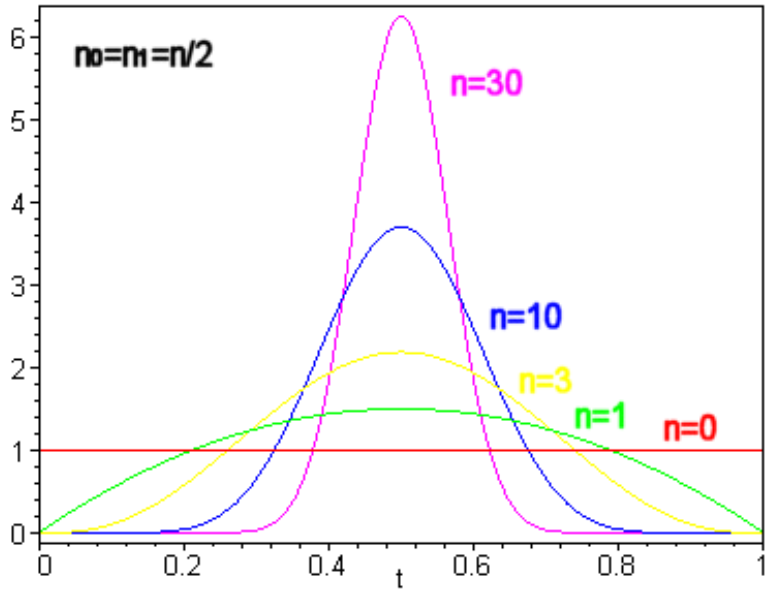
$$(\int_0^1 p(\theta)\, d\theta = 1 \text{ instead } \sum_{i \in I} p(H_i) = 1)$$

Evidence: $p(x) = \int_0^1 p_\theta(x) p(\theta)\, d\theta = \int_0^1 \theta^{n_1}(1-\theta)^{n_0}\, d\theta = \frac{n_1! n_0!}{(n_0+n_1+1)!}$

Example: Bayes' and Laplace's Rule

Bayes: Posterior plausibility of θ after seeing x is:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{(n+1)!}{n_1!n_0!}\theta^{n_1}(1-\theta)^{n_0}$$



Laplace: What is the probability of seeing 1 after having observed x?

$$p(x_{n+1} = 1|x_1...x_n) = \frac{p(x1)}{p(x)} = \frac{n_1+1}{n+2}$$

Exercise 1: Envelope Paradox

- I offer you two closed envelopes, one of them contains twice the amount of money than the other. You are allowed to pick one and open it. Now you have two options. Keep the money or decide for the other envelope (which could double or half your gain).
- Symmetry argument: It doesn't matter whether you switch, the expected gain is the same.
- Refutation: With probability $p = 1/2$, the other envelope contains twice/half the amount, i.e. if you switch your expected gain increases by a factor 1.25=1/2*2+1/2*1/2.
- Present a Bayesian solution.

Exercise 2: Confirmation Paradox

(i) $R \to B$ is confirmed by an R-instance with property B

(ii) $\neg B \to \neg R$ is confirmed by a $\neg B$-instance with property $\neg R$.

(iii) Since $R \to B$ and $\neg B \to \neg R$ are logically equivalent,
$R \to B$ is also confirmed by a $\neg B$-instance with property $\neg R$.

Example: Hypothesis (o): All ravens are black (R=Raven, B=Black).

(i) observing a Black Raven confirms Hypothesis (o).

(iii) observing a White Sock also confirms that all Ravens are Black,
since a White Sock is a non-Raven which is non-Black.

This conclusion sounds absurd.

Present a Bayesian solution.

Notation: Strings & Probabilities

Strings: $x = x_1 x_2 ... x_n$ with $x_t \in \mathcal{X}$ and $x_{1:n} := x_1 x_2 ... x_{n-1} x_n$ and $x_{<n} := x_1 ... x_{n-1}$.

Probabilities: $\rho(x_1 ... x_n)$ is the probability that an (infinite) sequence starts with $x_1 ... x_n$.

Conditional probability:

$$\rho_n := \rho(x_n | x_{<n}) = \rho(x_{1:n}) / \rho(x_{<n}),$$
$$\rho(x_1 ... x_n) = \rho(x_1) \cdot \rho(x_2 | x_1) \cdot ... \cdot \rho(x_n | x_1 ... x_{n-1}).$$

True data generating distribution: μ

The Bayes-Mixture Distribution ξ

- Assumption: The true (objective) environment μ is unknown.
- Bayesian approach: Replace true probability distribution μ by a Bayes-mixture ξ.
- Assumption: We know that the true environment μ is contained in some known countable (in)finite set $\mathcal{M}$ of environments.
- The Bayes-mixture ξ is defined as

$$\xi(x_{1:m}) := \sum_{\nu\in\mathcal{M}} w_\nu \nu(x_{1:m}) \quad \text{with} \quad \sum_{\nu\in\mathcal{M}} w_\nu = 1, \quad w_\nu > 0 \;\forall \nu$$

- The weights w_ν may be interpreted as the prior degree of belief that the true environment is ν, or $k^\nu = \ln w_\nu^{-1}$ as a complexity penalty (prefix code length) of environment ν.
- Then $\xi(x_{1:m})$ could be interpreted as the prior subjective belief probability in observing $x_{1:m}$.

Relative Entropy

Relative entropy: $D(\mathbf{p}||\mathbf{q}) := \sum_i p_i \ln \frac{p_i}{q_i}$

Properties: $D(\mathbf{p}||\mathbf{q}) \geq 0$ and $D(\mathbf{p}||\mathbf{q}) = 0 \Leftrightarrow \mathbf{p} = \mathbf{q}$

Instantaneous relative entropy: $d_t(x_{<t}) := \sum_{x_t \in \mathcal{X}} \mu(x_t|x_{<t}) \ln \frac{\mu(x_t|x_{<t})}{\xi(x_t|x_{<t})}$

Total relative entropy: $D_n := \sum_{t=1}^{n} \mathbf{E}[d_t] \leq \ln w_\mu^{-1}$

$\mathbf{E}[f]$ =Expectation of f w.r.t. the *true* distribution μ, e.g.

If $f : \mathcal{X}^n \to \mathbb{R}$, then $\mathbf{E}[f] := \sum_{x_{1:n}} \mu(x_{1:n}) f(x_{1:n})$.

Proof based on dominance or universality: $\xi(x) \geq w_\mu \mu(x)$.

Proof of the Entropy Bound

$$D_n \equiv \sum_{t=1}^{n}\sum_{x_{<t}} \mu(x_{<t})\cdot d_t(x_{<t}) \overset{(a)}{=} \sum_{t=1}^{n}\sum_{x_{1:t}} \mu(x_{1:t}) \ln \frac{\mu(x_t|x_{<t})}{\xi(x_t|x_{<t})} =$$

$$\overset{(b)}{=} \sum_{x_{1:n}} \mu(x_{1:n}) \ln \prod_{t=1}^{n} \frac{\mu(x_t|x_{<t})}{\xi(x_t|x_{<t})} \overset{(c)}{=} \sum_{x_{1:n}} \mu(x_{1:n}) \ln \frac{\mu(x_{1:n})}{\xi(x_{1:n})} \overset{(d)}{\leq} \ln w_\mu^{-1}$$

(a) Insert def. of d_t and used chain rule $\mu(x_{<t})\cdot\mu(x_t|x_{<t})=\mu(x_{1:t})$.

(b) $\sum_{x_{1:t}} \mu(x_{1:t}) = \sum_{x_{1:n}} \mu(x_{1:n})$ and argument of log is independent of $x_{t+1:n}$. The t sum can now be exchanged with the $x_{1:n}$ sum and transforms to a product inside the logarithm.

(c) Use chain rule again for μ and ξ.

(d) Use dominance $\xi(x) \geq w_\mu \mu(x)$.

Posterior Convergence

Theorem: $\xi(x_t|x_{<t}) \to \mu(x_t|x_{<t})$ rapid w.p.1 for $t \to \infty$

Proof: $D_\infty \equiv \sum_{t=1}^{\infty} \mathbf{E}[d_t] \leq \ln w_\mu^{-1}$ and $d_t \geq 0$

$$\Longrightarrow \quad d_t \stackrel{t\to\infty}{\longrightarrow} 0 \quad \Longleftrightarrow \quad \xi_t \to \mu_t .$$

Fazit: ξ is excellent universal predictor if unknown μ belongs to $\mathcal{M}$.

How to choose $\mathcal{M}$ and w_μ? Both as large as possible?! More later.

Sequential Decisions

A prediction is very often the basis for some decision. The decision results in an action, which itself leads to some reward or loss.

Let $\text{Loss}(x_t, y_t) \in [0,1]$ be the received loss when taking action $y_t \in \mathcal{Y}$ and $x_t \in \mathcal{X}$ is the t^{th} symbol of the sequence.

For instance, decision $\mathcal{Y} = \{\text{umbrella, sunglasses}\}$ based on weather forecasts $\mathcal{X} = \{\text{sunny, rainy}\}$.

Loss	sunny	rainy
umbrella	0.1	0.3
sunglasses	0.0	1.0

The goal is to minimize the μ-expected loss. More generally we define the Λ_ρ prediction scheme, which minimizes the ρ-expected loss:

$$y_t^{\Lambda_\rho} := \arg\min_{y_t \in \mathcal{Y}} \sum_{x_t} \rho(x_t | x_{<t}) \text{Loss}(x_t, y_t)$$

Loss Bounds

- Definition: μ-expected loss when Λ_ρ predicts the t^{th} symbol:
  $\text{Loss}_t(\Lambda_\rho)(x_{<t}) := \sum_{x_t} \mu(x_t|x_{<t})\text{Loss}(x_t, y_t^{\Lambda_\rho})$
- $\text{Loss}_t(\Lambda_{\mu/\xi})$ made by the informed/universal scheme $\Lambda_{\mu/\xi}$.
  $\text{Loss}_t(\Lambda_\mu) \leq \text{Loss}_t(\Lambda)\ \forall t, \Lambda$.
- Theorem: $0 \leq \text{Loss}_t(\Lambda_\xi) - \text{Loss}_t(\Lambda_\mu) \leq \sum_{x_t} |\xi_t - \mu_t| \leq \sqrt{2d_t} \stackrel{w.p.1}{\longrightarrow} 0$
- Total $\text{Loss}_{1:n}(\Lambda_\rho) := \sum_{t=1}^{n} \mathbf{E}[\text{Loss}_t(\Lambda_\rho)]$.
- Theorem: $\text{Loss}_{1:n}(\Lambda_\xi) - \text{Loss}_{1:n}(\Lambda_\mu) \leq 2D_n + 2\sqrt{\text{Loss}_{1:n}(\Lambda_\mu) D_n}$
- Corollary: If $\text{Loss}_{1:\infty}(\Lambda_\mu)$ is finite, then $\text{Loss}_{1:\infty}(\Lambda_\xi)$ is finite, and $\text{Loss}_{1:n}(\Lambda_\xi)/\text{Loss}_{1:\infty}(\Lambda_\mu) \to 1$ if $\text{Loss}_{1:\infty}(\Lambda_\mu) \to \infty$.
- Remark: Holds for any loss function $\in [0,1]$ with no assumptions (like i.i.d., Markovian, stationary, ergodic, ...) on $\mu \in \mathcal{M}$.

Proof of Instantaneous Loss Bounds

Abbreviations: $\mathcal{X}=\{1,...,N\}$, $N=|\mathcal{X}|$, $i=x_t$, $y_i=\mu(x_t|x_{<t})$, $z_i=\xi(x_t|x_{<t})$, $m=y_t^{\Lambda_\mu}$, $s=y_t^{\Lambda_\xi}$, $\ell_{xy}=\mathsf{Loss}(x,y)$.

This and definition of $y_t^{\Lambda_\mu}$ and $y_t^{\Lambda_\xi}$ and $\sum_i z_i\ell_{is}\leq\sum_i z_i\ell_{ij}\,\forall j$ implies

$$\mathsf{Loss}_t(\Lambda_\xi)-\mathsf{Loss}_t(\Lambda_\mu)\equiv\sum_i y_i\ell_{is}-\sum_i y_i\ell_{im}\overset{(a)}{\leq}\sum_i(y_i-z_i)(\ell_{is}-\ell_{im})$$
$$\leq\sum_i|y_i-z_i|\cdot|\ell_{is}-\ell_{im}|\overset{(b)}{\leq}\sum_i|y_i-z_i|\overset{(c)}{\leq}\sqrt{\sum_i y_i\ln\frac{y_i}{z_i}}\equiv\sqrt{2d_t(x_{<t})}$$

(a) We added $\sum_i z_i(\ell_{im}-\ell_{is})\geq 0$.

(b) $|\ell_{is}-\ell_{im}|\leq 1$ since $\ell\in[0,1]$.

(c) Pinsker's inequality (elementary, but not trivial)

Generalization: Continuous Probability Classes $\mathcal{M}$

In statistical parameter estimation one often has a continuous hypothesis class (e.g. a Bernoulli(θ) process with unknown $\theta \in [0,1]$).

$$\mathcal{M} := \{\mu_\theta : \theta \in \mathbb{R}^d\}, \quad \xi(x_{1:n}) := \int_{\mathbb{R}^d} d\theta\, w(\theta)\, \mu_\theta(x_{1:n}), \quad \int_{\mathbb{R}^d} d\theta\, w(\theta) = 1$$

We only used $\xi(x_{1:n}) \geq w_\mu \cdot \mu(x_{1:n})$ which was obtained by dropping the sum over μ. Here, restrict integral over $\mathbb{R}^d$ to a small vicinity N_δ of θ. For sufficiently smooth μ_θ and $w(\theta)$ we expect

$$\xi(x_{1:n}) \gtrsim |N_{\delta_n}| \cdot w(\theta) \cdot \mu_\theta(x_{1:n}) \quad \Longrightarrow \quad D_n \lesssim \ln w_\mu^{-1} + \ln |N_{\delta_n}|^{-1}$$

Average Fisher information $\bar{\jmath}_n$ measures curvature (parametric complexity) of $\ln \mu_\theta$. Weak regularity conditions on $\bar{\jmath}_n \Rightarrow$

Theorem: $D_n \leq \ln w_\mu^{-1} + \frac{d}{2} \ln \frac{n}{2\pi} + \frac{1}{2} \ln \det \bar{\jmath}_n + o(1)$

i.e. D_n grows only logarithmically with n.

Bayesian Sequence Prediction: Summary

- The aim of probability theory is to describe uncertainty.
- Various sources and interpretations of uncertainty: frequency, objective, and subjective probabilities.
- They all respect the same rules.
- General sequence prediction: Use known (subj.) Bayes mixture $\xi = \sum_{\nu\in\mathcal{M}} w_\nu \nu$ in place of unknown (obj.) true distribution μ.
- Bound on the relative entropy between ξ and μ.

$\Rightarrow$ posterior of ξ converges rapidly to the true posterior μ.

- ξ is also optimal in a decision-theoretic sense w.r.t. any bounded loss function.
- No structural assumptions on $\mathcal{M}$ and $\nu \in \mathcal{M}$.

Minimum Description Length: Contents

- Questions left open by Bayes
- Indifference=Symmetry and Maximum Entropy Principles
- Occam's Razor – The Simplicity Principle
- Priors from Prefix Sets/Codes – Kraft Inequality
- A Universal Choice of ξ and $\mathcal{M}$
- Optimality of the Universal Predictor
- The Minimum Description Length Principle
- Application: Sequence Prediction
- Application: Regression / Polynomial Fitting
- Summary

Minimum Description Length: Abstract

The Minimum Description/Message Length principle is one of the most important concepts in Machine Learning, and serves as a scientific guide, in general. The motivation is as follows: To make predictions involves finding regularities in past data, regularities in data allows for compression, hence short descriptions of data should help in making predictions. In this lecture series we approach MDL from a Bayesian perspective and relate it to a MAP (maximum a posteriori) model choice. The Bayesian prior is chosen in accordance with Occam and Epicurus and the posterior is approximated by the MAP solution. We reconsider (un)fair coin flips and compare the M(D)L to Bayes-Laplace's solution, and similarly for general sequence prediction tasks. Finally I present an application to regression / polynomial fitting.

When is a Sequence Random?

a) Is 01100101001011011010011111011 generated by a fair coin flip?
b) Is 11111111111111111111111111111 generated by a fair coin flip?
c) Is 11001001000011111110110101010 generated by a fair coin flip?
d) Is 01010101010101010101010101010 generated by a fair coin flip?

- Intuitively: (a) and (c) look random, but (b) and (d) look unlikely.
- Problem: Formally (a-d) have equal probability $(\frac{1}{2})^{length}$.
- Classical solution: Consider hypothesis class $H := \{\text{Bernoulli}(p) : p \in \Theta \subseteq [0,1]\}$ and determine p for which sequence has maximum likelihood $\Longrightarrow$ (a,c,d) are fair Bernoulli($\frac{1}{2}$) coins, (b) not.
- Problem: (d) is non-random, also (c) is binary expansion of π.
- Solution: Choose H larger, but how large? Overfitting? MDL?
- AIT Solution: A sequence is **random** *iff* it is **incompressible**.

What does Probability Mean?

Naive frequency interpretation is circular:

- Probability of event E is $p := \lim_{n\to\infty} \frac{k_n(E)}{n}$,
  $n = \#$ i.i.d. trials, $k_n(E) = \#$ occurrences of event E in n trials.
- Problem: Limit may be anything (or nothing):
  e.g. a fair coin can give: Head, Head, Head, Head, ... $\Rightarrow p = 1$.
- Of course, for a fair coin this sequence is "unlikely".
  For fair coin, $p = 1/2$ with "high probability".
- But to make this statement rigorous we need to formally know what "high probability" means. Circularity!

Also: In complex domains typical for AI, sample size is often 1.
(e.g. a single non-iid historic weather data sequences is given).
We want to know whether certain properties hold for this *particular* seq.

How to Choose the Prior?

The probability axioms allow relating probabilities and plausibilities of different events, but they do not uniquely fix a numerical value for each event, except for the sure event Ω and the empty event $\{\}$.

We need new principles for determining values for at least some basis events from which others can then be computed.

There seem to be only 3 general principles:

- The principle of indifference — the symmetry principle
- The maximum entropy principle
- Occam's razor — the simplicity principle

Concrete: How shall we choose the hypothesis space $\{H_i\}$ and their prior $p(H_i)$ –or– $\mathcal{M} = \{\nu\}$ and their weight w_ν.

Indifference or Symmetry Principle

Assign same probability to all hypotheses:

$p(H_i) = \frac{1}{|I|}$ for finite I

$p(H_\theta) = [\text{Vol}(\Theta)]^{-1}$ for compact and measurable Θ.

$\Rightarrow p(H_i|D) \propto p(D|H_i) \stackrel{\wedge}{=}$ classical Hypothesis testing (Max.Likelihood).

Prev. Example: H_θ =Bernoulli(θ) with $p(\theta) = 1$ for $\theta \in \Theta := [0,1]$.

Problems: Does not work for "large" hypothesis spaces:

(a) Uniform distr. on infinite $I = \mathbb{N}$ or noncompact Θ not possible!

(b) Reparametrization: $\theta \rightsquigarrow f(\theta)$. Uniform in θ is not uniform in $f(\theta)$.

Example: "Uniform" distr. on space of all (binary) sequences $\{0,1\}^\infty$:

$p(x_1...x_n) = (\frac{1}{2})^n \; \forall n \forall x_1...x_n \Rightarrow p(x_{n+1} = 1|x_1...x_n) = \frac{1}{2}$ always!

Inference so not possible (No-Free-Lunch myth).

Predictive setting: All we need is $p(x)$.

The Maximum Entropy Principle ...

is based on the foundations of statistical physics.

The symmetry principle is a special case of the maximum entropy principle.

Occam's Razor — The Simplicity Principle

- Only Occam's razor (in combination with Epicurus' principle) is general enough to assign prior probabilities in *every* situation.
- The idea is to assign high (subjective) probability to simple events, and low probability to complex events.
- Simple events (strings) are more plausible a priori than complex ones.
- This gives (approximately) justice to both Occam's razor and Epicurus' principle.

Prefix Sets/Codes

String x is (proper) prefix of y $:\Longleftrightarrow$ $\exists\, z(\neq \epsilon)$ such that $xz = y$.

Set $\mathcal{P}$ is prefix-free or a prefix code $:\Longleftrightarrow$ no element is a proper prefix of another.

Example: A self-delimiting code is prefix-free.

Kraft Inequality

For a prefix code $\mathcal{P}$ we have $\sum_{x\in\mathcal{P}} 2^{-\ell(x)} \leq 1$.

Conversely, let l_1, l_2, ... be a countable sequence of natural numbers such that Kraft's inequality $\sum_k 2^{-l_k} \leq 1$ is satisfied. Then there exists a prefix code $\mathcal{P}$ with these lengths of its binary code.

Proof of the Kraft-Inequality

Proof $\Rightarrow$: Assign to each $x \in \mathcal{P}$ the interval $\Gamma_x := [0.x, 0.x + 2^{-\ell(x)})$.

Length of interval Γ_x is $2^{-\ell(x)}$.

Intervals are disjoint, since $\mathcal{P}$ is prefix free, hence

$$\sum_{x \in \mathcal{P}} 2^{-\ell(x)} = \sum_{x \in \mathcal{P}} \text{Length}(\Gamma_x) \leq \text{Length}([0,1]) = 1$$

$\Leftarrow$: Idea: Choose l_1, l_2, ... in increasing order. Successively chop off intervals of lengths 2^{-l_1}, 2^{-l_2}, ... from left to right from $[0,1)$ and define left interval boundary as code.

Priors from Prefix Codes

- Let Code(H_ν) be a prefix code of hypothesis H_ν.
- Define complexity $Kw(\nu)$:=Length(Code(H_ν))
- Choose prior $w_\nu = p(H_\nu) = 2^{-Kw(\nu)}$
  $\Rightarrow \sum_{\nu\in\mathcal{M}} w_\nu \leq 1$ is semi-probability (by Kraft).
- How to choose a Code and hypothesis space $\mathcal{M}$?
- Praxis: Choose a code which is reasonable for your problem and $\mathcal{M}$ large enough to contain the true model.
- Theory: Choose a universal code and consider "all" hypotheses ...

A Universal Choice of ξ and $\mathcal{M}$

- We have to assume the existence of some structure on the environment to avoid the No-Free-Lunch Theorems [Wolpert 96].
- We can only unravel effective structures which are describable by (semi)computable probability distributions.
- So we may include *all* (semi)computable (semi)distributions in $\mathcal{M}$.
- Occam's razor and Epicurus' principle of multiple explanations tell us to assign high prior belief to simple environments.
- Using Kolmogorov's universal complexity measure $K(\nu)$ for environments ν one should set $w_\nu = 2^{-K(\nu)}$, where $K(\nu)$ is the length of the shortest program on a universal TM computing ν.
- The resulting mixture ξ is Solomonoff's (1964) universal prior.
- In the following we consider generic $\mathcal{M}$ and w_ν.

Optimality of the Universal Predictor

- There are $\mathcal{M}$ and $\mu \in \mathcal{M}$ and weights w_μ for which the loss bounds are tight.
- The universal prior ξ is pareto-optimal, in the sense that there is no ρ with $\mathcal{F}(\nu,\rho) \leq \mathcal{F}(\nu,\xi)$ for all $\nu \in \mathcal{M}$ and strict inequality for at least one ν, where $\mathcal{F}$ is the instantaneous or total squared distance s_t, S_n, or entropy distance d_t, D_n, or general Loss_t, $\text{Loss}_{1:n}$.
- ξ is balanced pareto-optimal in the sense that by accepting a slight performance decrease in some environments one can only achieve a slight performance increase in other environments.
- Within the set of enumerable weight functions with short program, the universal weights $w_\nu = 2^{-K(\nu)}$ lead to the smallest performance bounds within an additive (to $\ln w_\mu^{-1}$) constant in all enumerable environments.

The Minimum Description Length Principle

Identification of probabilistic model "best" describing data:

Probabilistic model(=hypothesis) H_ν with $\nu \in \mathcal{M}$ and data D.

Most probable model is $\nu^{\text{MDL}} = \arg\max_{\nu\in\mathcal{M}} p(H_\nu|D)$.

Bayes' rule: $p(H_\nu|D) = p(D|H_\nu)\cdot p(H_\nu)/p(D)$.

Occam's razor: $p(H_\nu) = 2^{-Kw(\nu)}$.

By definition: $p(D|H_\nu) = \nu(x)$, $D = x$ =data-seq., $p(D)$ =const.

Take logarithm $\quad\Longrightarrow\quad$ $\boxed{\nu^{\text{MDL}} = \arg\min_{\nu\in\mathcal{M}}\{K\nu(x) + Kw(\nu)\}}$

$K\nu(x) := -\log\nu(x)$ = length of Shannon-Fano code of x given H_ν.

$Kw(\nu)$ = length of model H_ν.

Names: Two-part MDL or MAP or MML ($\exists$ "slight" differences)

Predict with Best Model

- Use best model from class of models $\mathcal{M}$ for prediction:
- Predict y with probability $\nu^{\text{MDL}}(y|x) = \frac{\nu^{\text{MDL}}(xy)}{\nu^{\text{MDL}}(x)}$ (3 variants)
- $y^{\text{MDL}} = \arg\max_y\{\nu^{\text{MDL}}(y|x)\}$ is most likely continuation of x
- Special case: $Kw(\nu) =$const.
  $\Longrightarrow$ MDL $\rightsquigarrow$ ML:=Maximum likelihood principle.
- Example: $H_\theta =$Bernoulli(θ) with $\theta \in [0,1]$ and $Kw(\theta) :=$const. and $\nu(x_{1:n}) = \theta^{n_1}(1-\theta)^{n_0}$ with $n_1 = x_1 + ... + x_n = n - n_0$.
  $\Rightarrow \theta^{\text{MDL}} = \arg\min_\theta\{-\log\theta^{n_1}(1-\theta)^{n_0} + K(\theta)\} = \frac{n_1}{n} = \nu^{\text{MDL}}(1|x)$
  $=$ ML frequency estimate. (overconfident, e.g. $n_1 = 0$)
- Compare with Laplace' rule based on Bayes' rule: $\theta^{Laplace} = \frac{n_1+1}{n+2}$.

Application: Sequence Prediction

- Instead of Bayes mixture $\xi(x) = \sum_\nu w_\nu \nu(x)$, consider MAP/MDL
- $\nu^{\text{MDL}}(x) = \max\{w_\nu \nu(x) : \nu \in \mathcal{M}\} = \arg\min_{\nu\in\mathcal{M}} \{K\nu(x) + Kw(\nu)\}$.
- $$\boxed{\sum_{t=1}^{\infty} \mathbf{E}\Big[\sum_{x_t} (\mu(x_t|x_{<t}) - \nu^{\text{MDL}}(x_t|x_{<t}))^2\Big] \leq 8 w_\mu^{-1}} \Leftarrow$$ no log as for ξ!

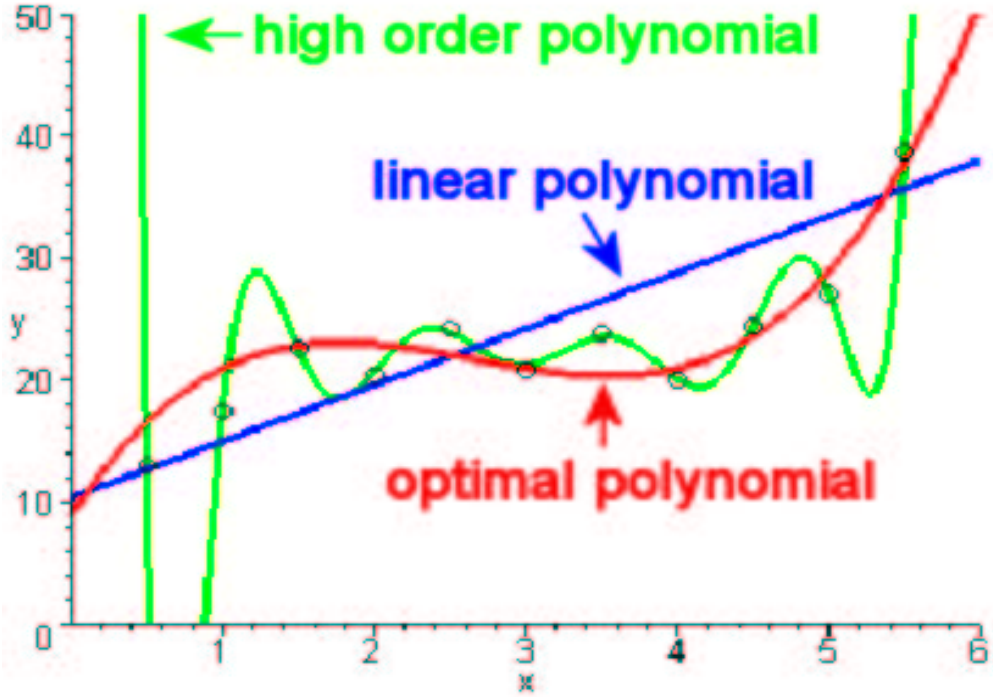
$\Rightarrow$ MDL converges, but speed can be exponentially worse than Bayes

$\Rightarrow$ be careful (bound is tight).

- For continuous smooth model class $\mathcal{M}$ and prior w_ν, MDL is as good as Bayes.

Application: Regression / Polynomial Fitting

- Data $D = \{(x_1, y_1), ..., (x_n, y_n)\}$
- Fit polynomial $f_d(x) := a_0 + a_1 x + a_2 x^2 + ... + a_d x^d$ of degree d through points D
- Measure of error: $SQ(a_0...a_d) = \sum_{i=1}^{n} (y_i - f_d(x_i))^2$
- Given d, minimize $SQ(a_{0:d})$ w.r.t. parameters $a_0...d_d$.
- This classical approach does not tell us how to choose d? ($d \geq n-1$ gives perfect fit)



MDL Solution to Polynomial Fitting

Assume y is Gaussian with variance σ^2 and mean $f_d(x)$, i.e.

$$P((x,y)|f_d) := P(y|x,f_d) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(y-f_d(x))^2}{2\sigma^2})$$

$$\Longrightarrow \quad P(D|f_d) = \prod_{i=1}^{d} P((x_i,y_i)|f_d) = \frac{e^{-SQ(a_{0:d})/2\sigma^2}}{(2\pi\sigma^2)^{n/2}}$$

The larger the error SQ, the less likely the data.

Occam: $P(f_d) = 2^{-Kw(f_d)}$. Simple coding: $Kw(f_d) \approx (d+1)\cdot C$, where C is the description length=accuracy of each coefficient a_k in bits $\Longrightarrow$

$$f^{\text{MDL}} = \operatorname{argmin}_f\{-\log P(D|f)+Kw(f)\} = \operatorname{argmin}_{d,a_{0:d}}\Big\{\frac{SQ(a_{0:d})}{2\sigma^2\ln 2}+(d+1)C\Big\}$$

Fixed d $\quad\Rightarrow\quad$ $a_{0:d}^{\text{ML}} = \arg\min_{a_{0:d}} SQ(a_{0:d})$ = classical solution

(by linear invariance of $\arg\min$)

MDL Polynomial Fitting: Determine Degree d

Determine d $(\min_f = \min_d \min_{f_d})$:

$$d = \arg\min_d \{ \frac{1}{2\sigma^2 \ln 2} \underbrace{SQ(a^{\text{ML}}_{0:d})}_{\text{least square fit}} + \underbrace{\frac{n}{2}\log(2\pi\sigma^2)}_{\text{“constant”}} + \underbrace{(d+1)C}_{\text{complexity penalty}} \}$$

Interpretation: Tradeoff between SQuare error and compleity penalty

σ and C may also be determined by minimizing this expression w.r.t. σ and C, but some subtleties have to be paid attention to.

Minimum Description Length: Summary

- Probability axioms give no guidance of how to choose the prior.
- Occam's razor is the only general (always applicable) principle for determining priors, especially in complex domains typical for AI.
- Prior $= 2^{-\text{descr.length}}$ — Universal prior $= 2^{-\text{Kolmogorov complexity}}$.
- Prediction $\widehat{=}$ finding regularities $\widehat{=}$ compression $\widehat{=}$ MDL.
- MDL principle: from a model class, a model is chosen that: minimizes the joint description length of the model and the data observed so far given the model.
- Similar to (Bayesian) Maximum a Posteriori (MAP) principle.
- MDL often as good as Bayes but not always.

The Similarity Metric: Contents

- Kolmogorov Complexity
- The Universal Similarity Metric
- Tree-Based Clustering
- Genomics & Phylogeny: Mammals, SARS Virus & Others
- Classification of Different File Types
- Language Tree (Re)construction
- Classify Music w.r.t. Composer
- Further Applications
- Summary

The Similarity Metric: Abstract

The MDL method has been studied from very concrete and highly tuned practical applications to general theoretical assertions. Sequence prediction is just one application of MDL. The MDL idea has also been used to define the so called information distance or universal similarity metric, measuring the similarity between two individual objects. I will present some very impressive recent clustering applications based on standard Lempel-Ziv or bzip2 compression, including a completely automatic reconstruction (a) of the evolutionary tree of 24 mammals based on complete mtDNA, and (b) of the classification tree of 52 languages based on the declaration of human rights and (c) others.

Based on [Cilibrasi&Vitanyi'03]

Kolmogorov Complexity

Question: When is object=string x similar to object=string y?

Universal solution: x similar y $\Leftrightarrow$ x can be easily (re)constructed from y $\Leftrightarrow$ Kolmogorov complexity $K(x|y) := \min\{\ell(p) : U(p,y) = x\}$ is small

Examples:

1) x is very similar to itself ($K(x|x) \stackrel{+}{=} 0$)
2) A processed x is similar to x ($K(f(x)|x) \stackrel{+}{=} 0$ if $K(f) = O(1)$).
   e.g. doubling, reverting, inverting, encrypting, partially deleting x.
3) A random string is with high probability not similar to any other string ($K(\text{random}|y) = \text{length}(\text{random})$).

The problem with $K(x|y)$ as similarity=distance measure is that it is neither symmetric nor normalized nor computable.

The Universal Similarity Metric

- Symmetrization and normalization leads to a/the universal metric d:

$$0 \;\leq\; d(x,y) := \frac{\max\{K(x|y),K(y|x)\}}{\max\{K(x),K(y)\}} \;\leq\; 1$$

- Every effective similarity between x and y is detected by d
- Use $K(x|y) \approx K(xy) - K(y)$ (coding T) and $K(x) \equiv K_U(x) \approx K_T(x)$ $\Longrightarrow$ computable approximation: Normalized compression distance:

$$d(x,y) \;\approx\; \frac{K_T(xy) - \min\{K_T(x),K_T(y)\}}{\max\{K_T(x),K_T(y)\}} \lesssim 1$$

- For T choose Lempel-Ziv or gzip or bzip(2) (de)compressor in the applications below.
- Theory: Lempel-Ziv compresses asymptotically better than any probabilistic finite state automaton predictor/compressor.

Tree-Based Clustering

- If many objects $x_1, ..., x_n$ need to be compared, determine the

  similarity matrix $M_{ij} = d(x_i, x_j)$ for $1 \leq i, j \leq n$

- Now cluster similar objects.

- There are various clustering techniques.

- Tree-based clustering: Create a tree connecting similar objects,

- e.g. quartet method (for clustering)

Genomics & Phylogeny: Mammals

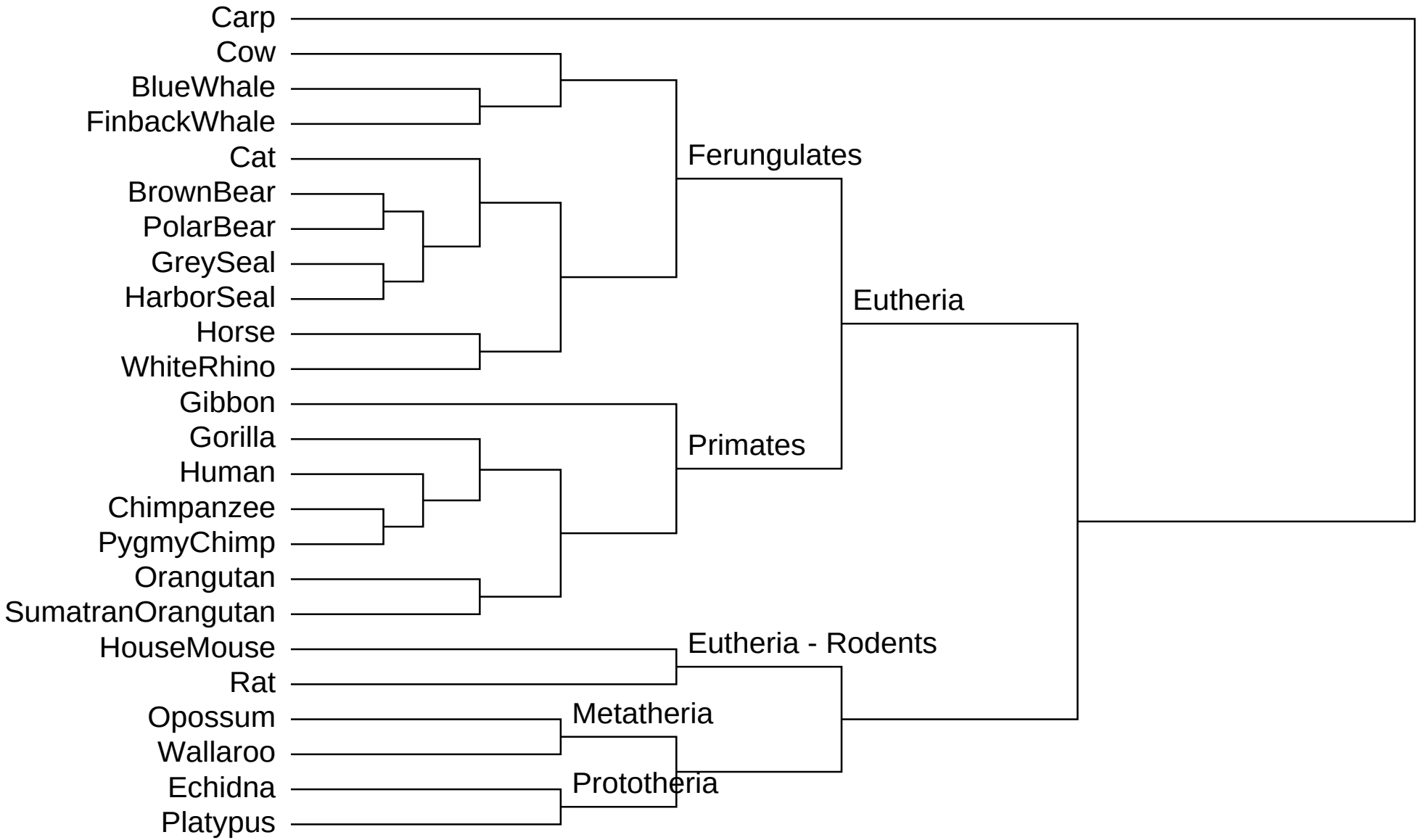
Let $x_1, ..., x_n$ be mitochondrial genome sequences of different mammals:

Partial distance matrix M_{ij} using bzip2(?)

			Cat		Echidna		Gorilla			...
	BrownBear		Chimpanzee		FinWhale		HouseMouse			...
		Carp		Cow		Gibbon			Human	...
BrownBear	0.002	0.943	0.887	0.935	0.906	0.944	0.915	0.939	0.940	0.934	0.930	...
Carp	0.943	0.006	0.946	0.954	0.947	0.955	0.952	0.951	0.957	0.956	0.946	...
Cat	0.887	0.946	0.003	0.926	0.897	0.942	0.905	0.928	0.931	0.919	0.922	...
Chimpanzee	0.935	0.954	0.926	0.006	0.926	0.948	0.926	0.849	0.731	0.943	0.667	...
Cow	0.906	0.947	0.897	0.926	0.006	0.936	0.885	0.931	0.927	0.925	0.920	...
Echidna	0.944	0.955	0.942	0.948	0.936	0.005	0.936	0.947	0.947	0.941	0.939	...
FinbackWhale	0.915	0.952	0.905	0.926	0.885	0.936	0.005	0.930	0.931	0.933	0.922	...
Gibbon	0.939	0.951	0.928	0.849	0.931	0.947	0.930	0.005	0.859	0.948	0.844	...
Gorilla	0.940	0.957	0.931	0.731	0.927	0.947	0.931	0.859	0.006	0.944	0.737	...
HouseMouse	0.934	0.956	0.919	0.943	0.925	0.941	0.933	0.948	0.944	0.006	0.932	...
Human	0.930	0.946	0.922	0.667	0.920	0.939	0.922	0.844	0.737	0.932	0.005	...
	...	...	...	...	...	...	...	...	...	...	...	...

Genomics & Phylogeny: Mammals

Evolutionary tree built from complete mammalian mtDNA of 24 species:

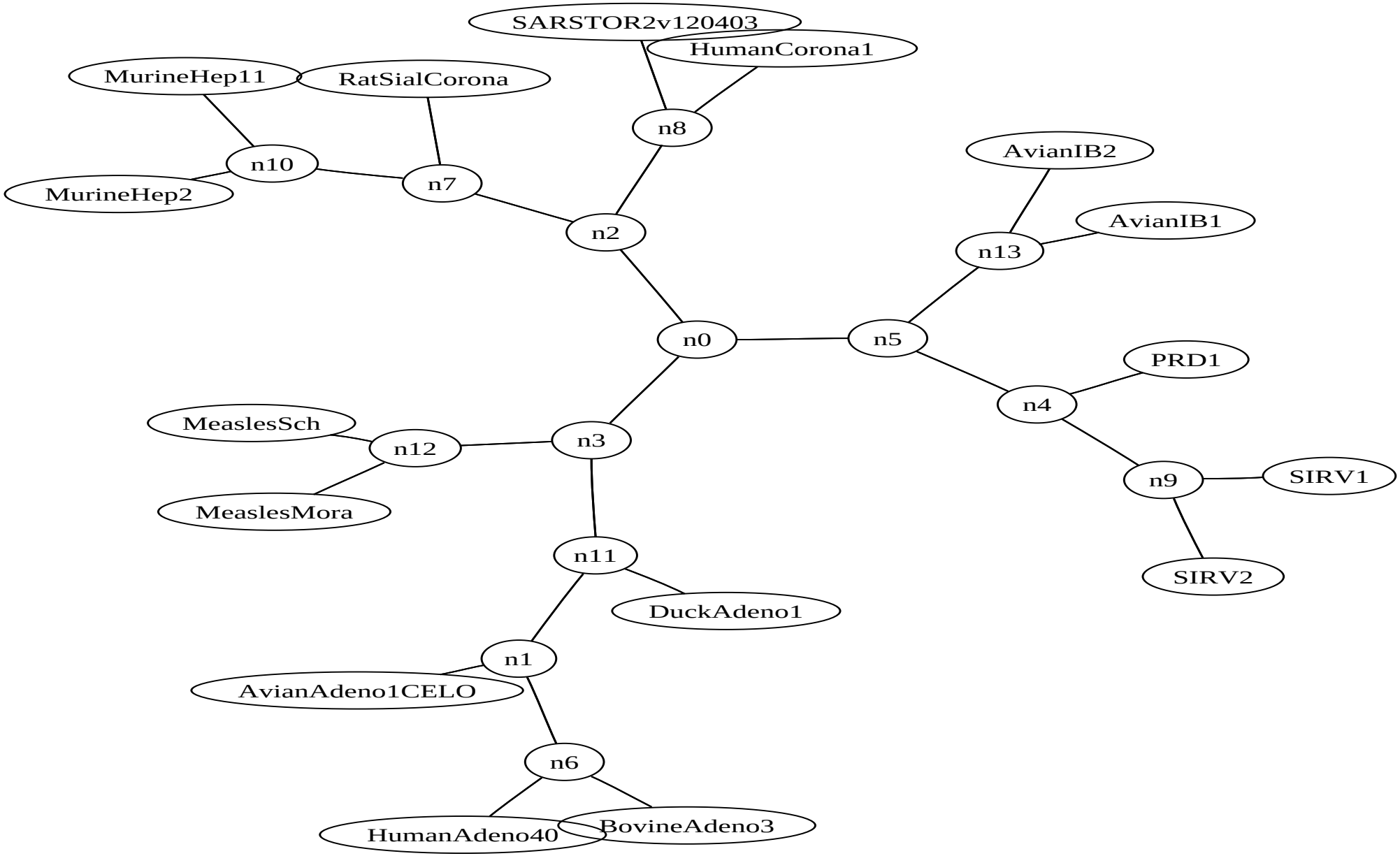


Genomics & Phylogeny: SARS Virus and Others

- Clustering of SARS virus in relation to potential similar virii based on complete sequenced genome(s) using bzip2:

- The relations are very similar to the definitive tree based on medical-macrobio-genomics analysis from biologists.

Genomics & Phylogeny: SARS Virus and Others



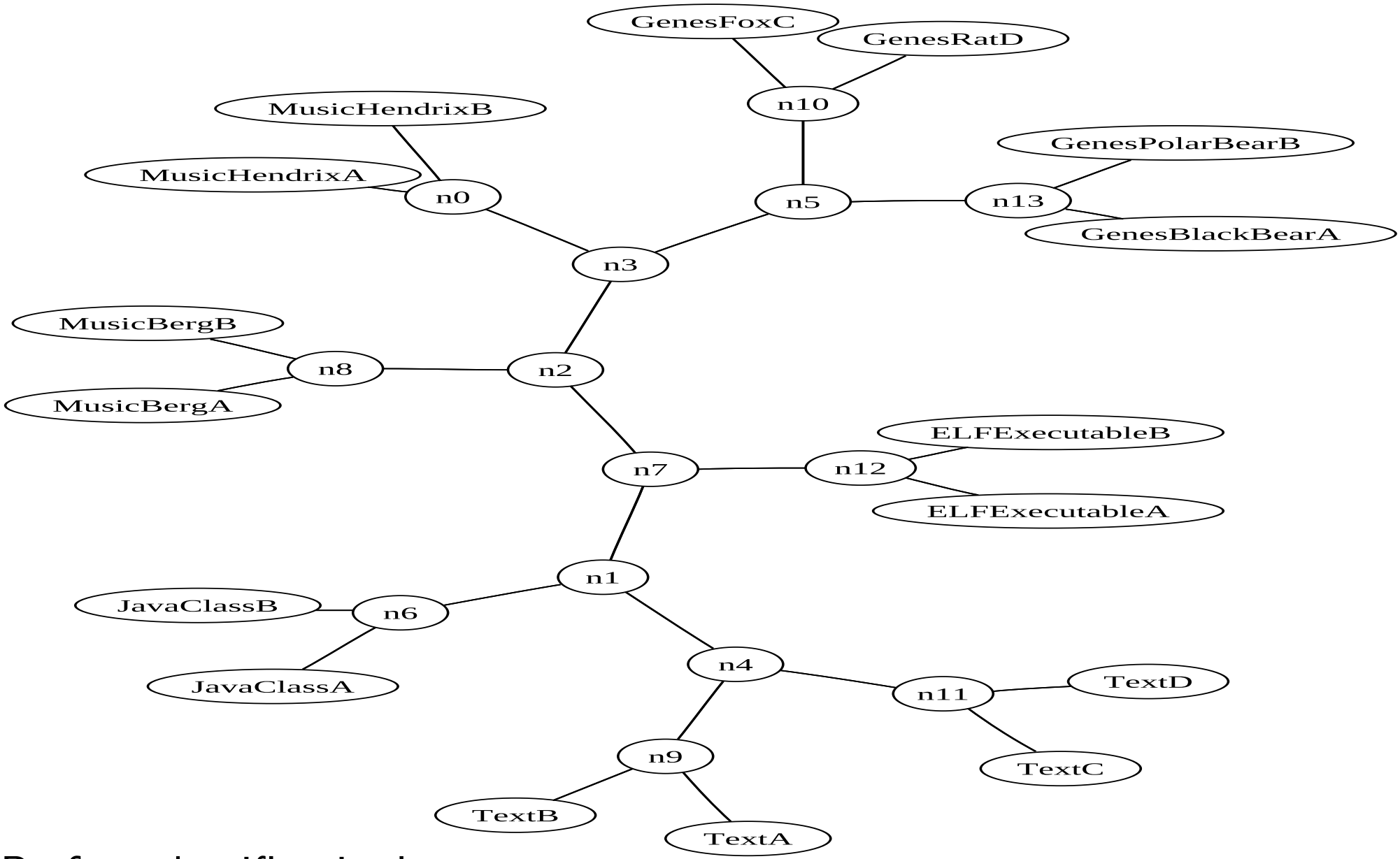
Classification of Different File Types

Classification of files based on markedly different file types using bzip2

- Four mitochondrial gene sequences
- Four excerpts from the novel "The Zeppelin's Passenger"
- Four MIDI files without further processing
- Two Linux x86 ELF executables (the cp and rm commands)
- Two compiled Java class files

No features of any specific domain of application are used!

Classification of Different File Types



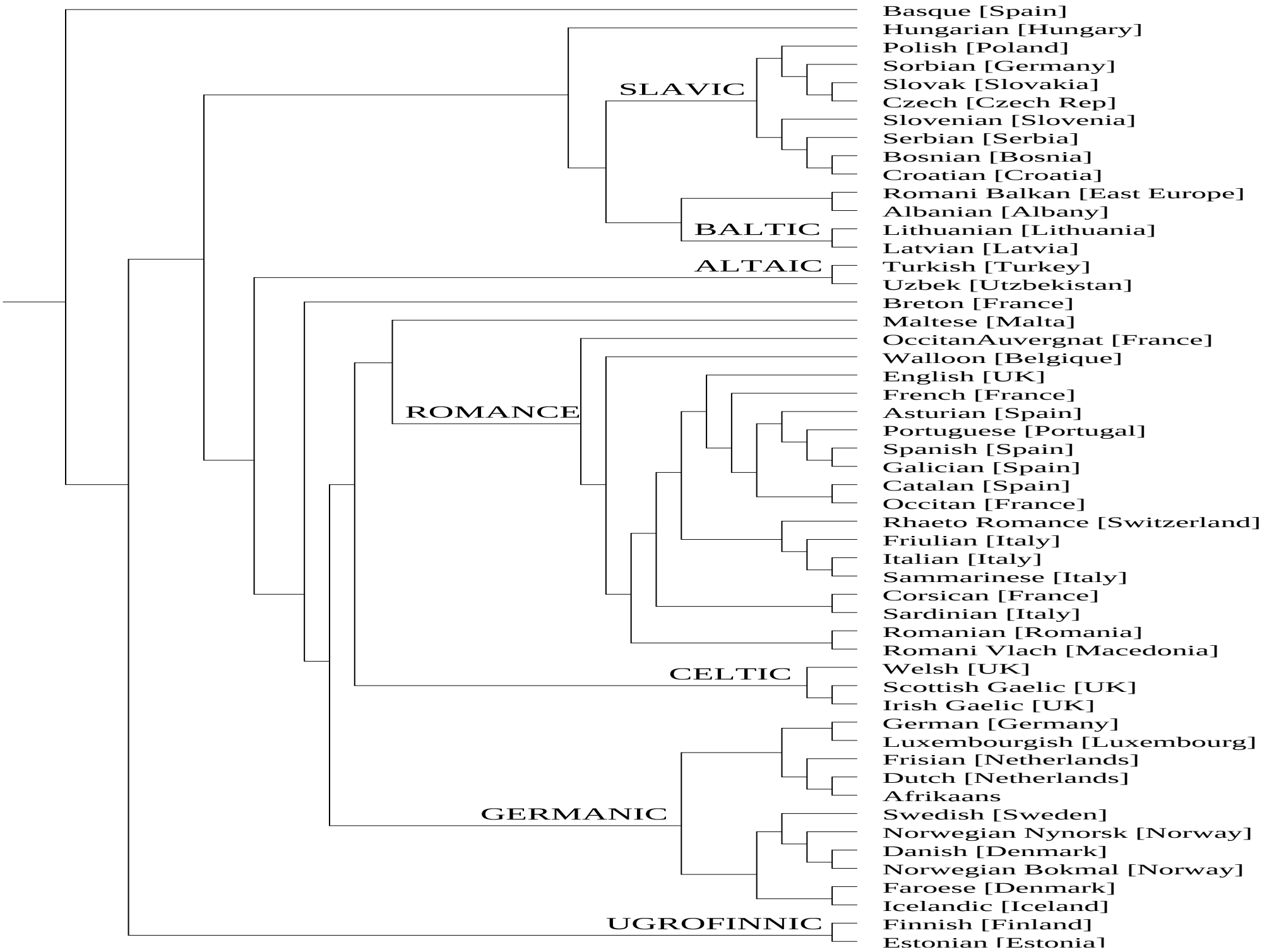
Perfect classification!

Language Tree (Re)construction

- Let $x_1, ..., x_n$ be the "The Universal Declaration of Human Rights" in various languages $1, ..., n$.

- Distance matrix M_{ij} based on gzip. Language tree constructed from M_{ij} by the Fitch-Margoliash method [Li&al'03]

- All main linguistic groups can be recognized (next slide)


Basque [Spain]
Hungarian [Hungary]
Polish [Poland]
Sorbian [Germany]
Slovak [Slovakia]
Czech [Czech Rep]
Slovenian [Slovenia]
Serbian [Serbia]
Bosnian [Bosnia]
Croatian [Croatia]
Romani Balkan [East Europe]
Albanian [Albany]
Lithuanian [Lithuania]
Latvian [Latvia]
Turkish [Turkey]
Uzbek [Utzbekistan]
Breton [France]
Maltese [Malta]
OccitanAuvergnat [France]
Walloon [Belgique]
English [UK]
French [France]
Asturian [Spain]
Portuguese [Portugal]
Spanish [Spain]
Galician [Spain]
Catalan [Spain]
Occitan [France]
Rhaeto Romance [Switzerland]
Friulian [Italy]
Italian [Italy]
Sammarinese [Italy]
Corsican [France]
Sardinian [Italy]
Romanian [Romania]
Romani Vlach [Macedonia]
Welsh [UK]
Scottish Gaelic [UK]
Irish Gaelic [UK]
German [Germany]
Luxembourgish [Luxembourg]
Frisian [Netherlands]
Dutch [Netherlands]
Afrikaans
Swedish [Sweden]
Norwegian Nynorsk [Norway]
Danish [Denmark]
Norwegian Bokmal [Norway]
Faroese [Denmark]
Icelandic [Iceland]
Finnish [Finland]
Estonian [Estonia]
SLAVIC
BALTIC
ALTAIC
ROMANCE
CELTIC
GERMANIC
UGROFINNIC

Classify Music w.r.t. Composer

Let $m_1, ..., m_n$ be pieces of music in MIDI format.

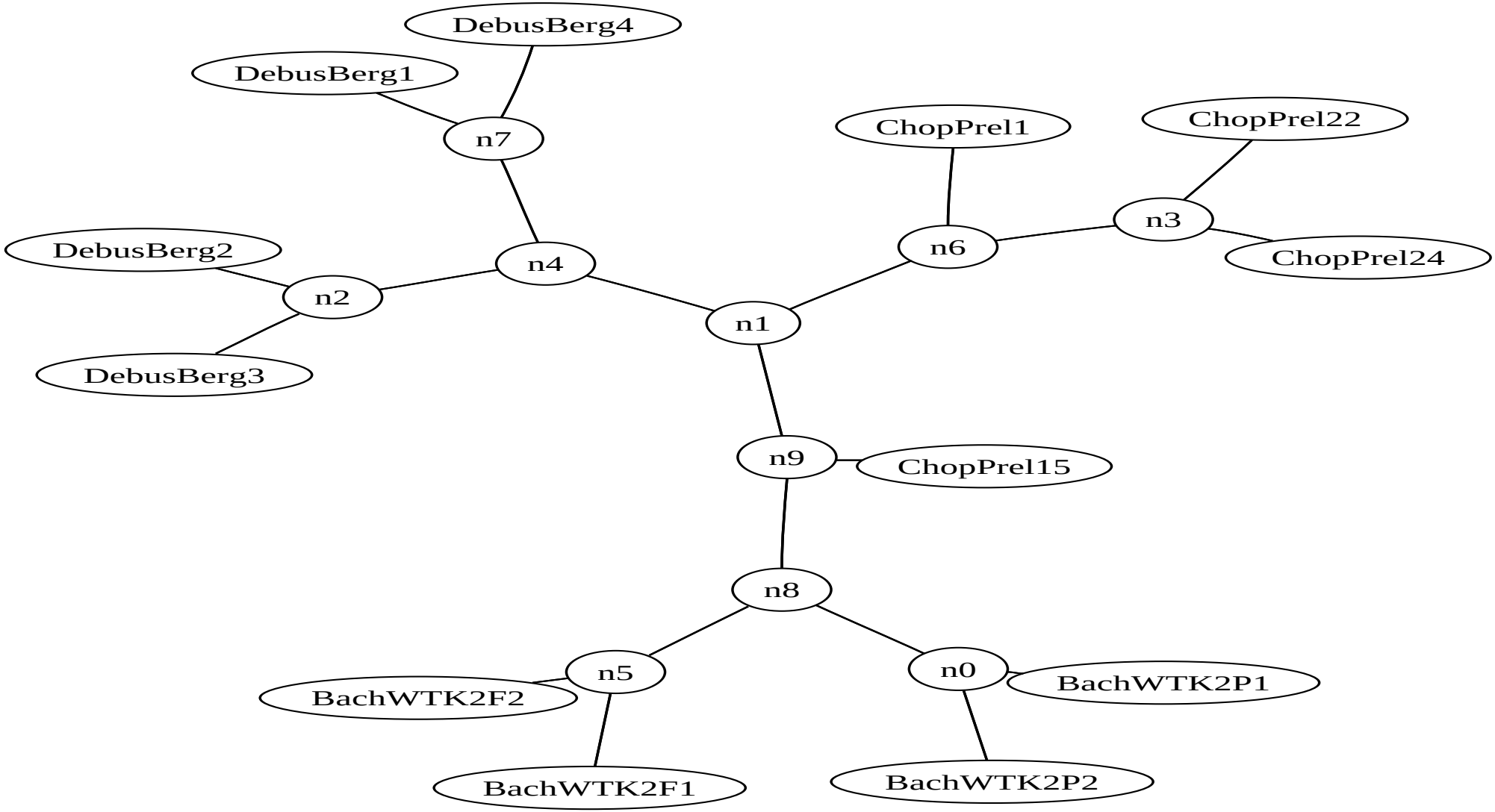
Preprocessing the MIDI files:

- Delete identifying information (composer, title, ...), instrument indicators, MIDI control signals, tempo variations, ...
- Keep only note-on and note-off information.
- A note, $k \in \mathbb{Z}$ half-tones above the average note is coded as a signed byte with value k.
- The whole piece is quantized in 0.05 second intervals.
- Tracks are sorted according to decreasing average volume, and then output in succession.

Processed files $x_1, ..., x_n$ still sounded like the original.

Classify Music w.r.t. Composer

12 pieces of music: 4×Bach + 4×Chopin + 4×Debussy. Class. by bzip2



Perfect grouping of processed MIDI files w.r.t. composers.

Further Applications

- Classification of Fungi
- Optical character recognition
- Classification of Galaxies
- Clustering of novels w.r.t. authors
- Larger data sets

See [Cilibrasi&Vitanyi'03]

The Clustering Method: Summary

- based on the universal similarity metric,
- based on Kolmogorov complexity,
- approximated by bzip2,
- with the similarity matrix represented by tree,
- approximated by the quartet method

- **leads to excellent classification in many domains.**

Prediction with Expert Advice: Contents

Prediction with Expert Advice: Abstract

Prediction with expert advice is a very active field of research that studies prediction in a worst-case setting, i.e. the actual sequences are generated by an adversary. Instead of trying to predict absolutely well, the goal is a good relative performance with respect to a pool of experts. One can prove that the regret is essentially bounded by the square root of a complexity term and the loss of the best expert. This method and corresponding performance bounds are related or dual to the Bayesian and MDL approaches to sequence prediction. This lecture gives an introduction and presents the two major variants, "Weighted Majority" and "Follow the Perturbed Leader". I present loss bounds for adaptive learning rate and both finite expert classes with uniform weights and countable expert classes with arbitrary weights. Finally I compare the bounds to corresponding Bayes-bounds.

Prediction with Expert Advice (PEA) - Informal

Given a class of n experts $\{\text{Expert}_1, ..., \text{Expert}_n\}$, each Expert_i at times $t = 1, 2, ...$ makes a prediction y_t^i.

The goal is to construct a master algorithm, which exploits the experts, and predicts asymptotically as well as the best expert in hindsight.

	Expert_1	Expert_2	...	Expert_n	PEA	true	Loss
day_1	0	0	...	0	0	1	1
day_2	0	1	...	1	1	1	0
day_3	1	0	...	1	1	0	1
...	...	...	...	...	...	...	...
day_t	y_t^1	y_t^2	...	y_t^n	y_t^{PEA}	x_t	$\lvert y_t^{\text{PEA}} - x_t \rvert$

Prediction with Expert Advice (PEA) - Setup

More formally, a PEA-Master is defined as:

For $t = 1, 2, ..., T$
- Predict $y_t^{\text{PEA}} := \text{PEA}(x_{<t}, \mathbf{y}_t, \text{Loss})$
- Observe $x_t := \text{Env}(\mathbf{y}_{<t}, x_{<t}, y_{<t}^{\text{PEA}})$
- Receive $\text{Loss}_t(\text{Expert}_i) := \text{Loss}(x_t, y_t^i)$ for each Expert $(i = 1, ..., n)$
- Suffer $\text{Loss}_t(\text{PEA}) := \text{Loss}_t(x_t, y_t^{\text{PEA}})$

Notation: $x_{<t} := (x_1, ..., x_{t-1})$ and $\mathbf{y}_t = (y_t^1, ..., y_t^n)$.

Goals

BEH := Best Expert in Hindsight = Expert of minimal total Loss.
$\text{Loss}_{1:T}(\text{BEH}) = \min_{e\in\mathcal{E}} \text{Loss}_{1:T}(\text{Expert}_e)$.

0) Regret := $\text{Loss}_{1:T}(\text{PEA}) - \text{Loss}_{1:T}(\text{BEH})$
shall be small ($O(\sqrt{\text{Loss}_{1:T}(\text{BEH})})$).

1) Any bounded Loss function (w.l.g. $0 \leq \text{Loss}_t \leq 1$).
Literature: Mostly specific Loss (absolute, 0/1, log, square)

2) Neither (non-trivial) upper bound on total Loss,
nor sequence length T is known. Solution: Adaptive learning rate.

3) Infinite number of Experts. Motivation:
- Expert_e = polynomial of degree $e = 1, 2, 3, ...$ through data -or-
- $\mathcal{E}$ = class of all computable (or finite state or ...) Experts.

Best Expert in Hindsight (BEH)

$$\begin{aligned}
\text{BEH} &:= \text{Expert of minimal total Loss}, \quad \text{i.e.} \\
i^{\text{BEH}} &:= \arg\min_i\{\text{Loss}_{1:T}(\text{Expert}_i)\}, \quad \text{where} \\
\text{Loss}_{1:T} &:= \text{Loss}_1 + ... + \text{Loss}_T \\
\text{Total Loss} &:= \text{sum of instantaneous losses}
\end{aligned}$$

Goal

Total Loss of PEA shall not be much more
than Loss of BEH, i.e. of any Expert.

$$\text{Loss}_{1:T}(\text{PEA}) \overset{?}{\lesssim} \text{Loss}_{1:T}(\text{BEH}) \overset{\surd}{\leq} \text{Loss}_{1:T}(\text{Expert}_i) \quad \forall i$$

Naive Ansatz: Follow the Leader (FL)

FL exploits prediction of expert which performed best in past, i.e.

$$i_t^{\text{FL}} := \arg\min_i \{\text{Loss}_{<t}(\text{Expert}_i)\} \quad \text{(known at time } t\text{)}$$

At time t, FL predicts $y_t^{\text{FL}} := y_t^{i_t^{\text{FL}}}$.

Problem: The predictor which performed best in the past my oscillate.

$\Longrightarrow$ FL often selects suboptimal expert.

Example (2 Experts): $\text{Loss}_{t=1,2,\ldots,T}(\text{Expert}_2^1) = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1/2 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$

$\Longrightarrow \quad \text{Loss}_{1:T}(\text{Expert}_2^1) \approx T/2 \qquad \longleftarrow$ twice as large $\searrow$

$\Longrightarrow \quad i_t^{\text{FL}} = \begin{cases} 1 \text{ if } t \text{ is even} \\ 2 \text{ if } t \text{ is odd} \end{cases}$, but $\text{Loss}_t(\text{FL}) = 1 \quad \Rightarrow \quad \text{Loss}_{1:T}(\text{FL}) = T$

Solution: Smooth decision by randomization

Weighted Majority (WM)

Take expert which performed best in past with high probability and others with smaller probability.

At time t, select Expert I_t^{WM} with probability

$$P[I_t^{\text{WM}} = i] \;\propto\; \exp[-\eta_t \cdot \text{Loss}_{<t}(\text{Expert}_i)]$$

η_t = learning rate.

[Littlestone&Warmuth'90 (Classical)]: 0/1 loss and η_t=const.

[Freund&Shapire'97 (Hedge)] and others: General Loss, but η_t=const.

[Cesa-Bianchi et al.'97]: Piecewise constant η_t.

[Auer&Gentile'00, Yaroshinsky et al.'04]: Smooth $\eta_t \searrow 0$, but only 0/1 Loss.

Follow the Perturbed Leader (FPL)

Select expert of minimal perturbed Loss.

Let Q_t^i be i.i.d. random variables.

Select expert $I_t^{\text{FPL}} := \arg\min_i\{\text{Loss}_{<t}(\text{Expert}_i) - Q_t^i/\eta\}$.

[Hannan'57]: $Q_t^i \overset{d.}{\sim} -\text{Uniform}[0,1]$,

[Kalai&Vempala'03]: $P[Q_t^i = u] = \frac{1}{2}e^{-|u|}$,

[Hutter&Poland'04]: $P[Q_t^i = u] = e^{-u} \quad (u \geq 0)$.

For all PEA variants (WM & FPL & others) it holds:

$P[I_t = i] = \{^{large}_{small}\}$ if Expert_i has $\{^{small}_{large}\}$ Loss.

$I_t \overset{\eta\to\infty}{\longrightarrow}$ Best Expert in Past $= i_t^{\text{FL}}$ $\qquad (\eta =$ learning rate$)$

$I_t \overset{\eta\to 0}{\longrightarrow}$ Uniform distribution among Experts.

Infinite number of Experts

Example 1) Expert_i = polynomial of degree $i = 1, 2, 3, ...$ through data

Example 2) $\{\text{Expert}_i : i \in \mathbb{N}\}$ = class of all computable Experts.

Solution: Penalize "complex" Experts (Occam's razor).

Assign complexity k^i to Expert_i -or- a-priori probability $w^i = e^{-k^i}$.

Assume Kraft inequality $\sum_i w^i \leq 1$.

$\Rightarrow k^i$ = prefix code length -and- w^i=(semi)probability.

Examples: Finite number n of Experts: $k^i = \ln n$.

Infinite #Experts: $k^i = \frac{1}{2} + 2\ln i$ increases slowly with i.

WM: $\boxed{P[I_t^{\text{WM}} = i] \;\propto\; w^i \cdot \exp[-\eta_t \cdot \text{Loss}_{<t}(\text{Expert}_i)]}$

FPL: $\boxed{I_t^{\text{FPL}} \;:=\; \arg\min_i\{\text{Loss}_{<t}(\text{Expert}_i) + (k^i - Q_t^i)/\eta_t\}}$

The FPL Algorithm

For $t = 1, ..., T$

- Choose i.i.d. random vector $Q_t \stackrel{d.}{\sim} \exp$, i.e. $P[Q_t^i] = e^{-Q_t^i}$ $(Q_t^i \geq 0)$.
- Choose learning rate η_t.
- Output prediction of expert i which minimizes $\text{Loss}_{<t}(\text{Expert}_i) + (k^i - Q_t^i)/\eta_t$.
- Receive $\text{Loss}_t(\text{Expert}_i)$ for each expert i.
- Suffer $\text{Loss}_t(\text{FPL})$.

Key Analysis Tool: Implicit or Infeasible FPL

$I_t^{\text{IFPL}} := \arg\min_i\{\text{Loss}_{1:t}(\text{Expert}_i) + (k^i - Q_t^i)/\eta_i\}$

IFPL is infeasible, since it depends on $\text{Loss}_t(x_t, {y_t}^i)$, unknown at time t.

One can show: $\underline{\text{Loss}}_{1:T}(\text{FPL}) \lesssim \underline{\text{Loss}}_{1:T}(\text{IFPL}) \lesssim \text{Loss}_{1:T}(\text{BEH})$

Since FPL is randomized, we need to consider expected-Loss $=: \underline{\text{Loss}}$

$$\underline{\text{Loss}}_{1:T}(\text{IFPL}) \quad \leq \quad \begin{cases} \text{Loss}_{1:T}(\text{Expert}_i) + k^i/\eta_T & \forall i, \\ \text{Loss}_{1:T}(\text{BEH}) + \frac{\ln n}{\eta_T} & \text{if} \quad k^i = \ln n. \end{cases}$$

$$\underline{\text{Loss}}_t(\text{FPL}) \quad \leq \quad e^{\eta_t} \cdot \underline{\text{Loss}}_t(\text{IFPL})$$

Choose η_t, and sum latter bound over $t = 1, ..., T$, and chain with first bound to get final bounds ...

Regret Bounds for $n < \infty$ and $k^i = \ln n$

Regret $:= \underline{\text{L}}\text{oss}_{1:T}(\text{FPL}) - \text{Loss}_{1:T}(\text{BEH})$

Static $\quad \eta_t = \sqrt{\frac{\ln n}{T}} \quad \Longrightarrow \quad \text{Regret} \leq 2\sqrt{T \cdot \ln n}$
Dynamic $\eta_t = \sqrt{\frac{\ln n}{2t}} \quad \Longrightarrow \quad \text{Regret} \leq 2\sqrt{2T \cdot \ln n}$
Self-confident $\eta_t = \sqrt{\frac{\ln n}{2(\underline{\text{L}}\text{oss}_{<t}(\text{FPL})+1)}} \quad \Longrightarrow$ $\text{Regret} \leq 2\sqrt{2(\text{Loss}_{1:T}(\text{BEH}) + 1) \cdot \ln n} + 8 \ln n$
Adaptive $\eta_t = \sqrt{\frac{1}{2}} \min\left\{1, \sqrt{\frac{\ln n}{\text{Loss}_{<t}(\text{“BEH”})}}\right\} \quad \Longrightarrow$ $\text{Regret} \leq 2\sqrt{2\text{Loss}_{1:T}(\text{BEH}) \cdot \ln n} + 5 \ln n \cdot \ln \text{Loss}_{1:T}(\text{BEH}) + 3\ln n + 6$

No hidden $O()$ terms!

Regret Bounds for $n = \infty$ and general k^i

We expect $\ln n \rightsquigarrow k^i$, i.e. Regret $= O(\sqrt{k^i \cdot (\text{Loss or } T)})$.

Problem: Choice of $\eta_t = \sqrt{k^i/...}$ depends on i. Proofs break down.

Choose: $\eta_t = \sqrt{1/...} \quad \Rightarrow \quad$ Regret $\leq \; k^i\sqrt{\cdots}$, i.e. k^i not under $\sqrt{\ }$.

Solution: Two-Level **Hierarchy of Experts**:

Group all experts of (roughly) equal complexity.

- FPL^K over subclass of experts with complexity $k^i \in (K-1, K]$. Choose $\eta_t^K = \sqrt{K/2\text{Loss}_{<t}} =$ constant within subclass.
- Regard each FPL^K as a (meta)expert. Construct from them (meta) $\widetilde{\text{FPL}}$. Choose $\tilde{\eta}_t = \sqrt{1/\text{Loss}_{<t}}$.

$$\Longrightarrow \qquad \boxed{\text{Regret} \leq \; 2\sqrt{2\,k^i \cdot \text{Loss}_{1:T}(\text{Expert}_i)} \cdot (1 + O(\tfrac{\ln k^i}{\sqrt{k^i}})) + O(k^i)}$$

Some more FPL Results

Lower bound: $\underline{\text{Loss}}_{1:T}(\text{IFPL}) \geq \text{Loss}_{1:T}(\text{BEH}) + \frac{\ln n}{\eta_T}$ if $k^i = \ln n$.

Bounds with high probability (Chernoff-Hoeffding):
$P[|\text{Loss}_{1:T} - \underline{\text{Loss}}_{1:T}| \geq \sqrt{3c\underline{\text{Loss}}_{1:T}}] \leq 2e^{-c}$ is tiny for e.g. $c = 5$.

Computational aspects: It is trivial to generate the randomized decision of FPL. If we want to *explicitly* compute the probability we need to compute a 1D integral.

Deterministic prediction: FPL can be derandomized if prediction space $\mathcal{Y}$ and loss-function $\text{Loss}(x, y)$ are convex.

PEA versus Bayes Bounds – Formal

Formal similarity and duality between Bayes and PEA bounds is striking:

$$\bar{\text{Loss}}_{1:T}(\text{Bayes}_\xi) \;\leq\; \bar{\text{Loss}}_{1:T}(\text{Any } \Lambda) + 2\sqrt{\bar{\text{Loss}}_{1:T}(\text{Any } \Lambda)\cdot k^\mu} + 2k^\mu$$

$$\underline{\text{Loss}}_{1:T}(\text{PEA}) \leq \text{Loss}_{1:T}(\text{Expert}_e) + c\cdot\sqrt{\text{Loss}_{1:T}(\text{Expert}_e)\cdot k^e} + b\cdot k^e$$

$$c = 2\sqrt{2} \text{ and } b = 8 \text{ for PEA} = \text{FPL.}$$

	beats predictors	in environment	expectation w.r.t.	function of
Bayes	all Λ	$\mu \in \mathcal{M}$	environment μ	$\mathcal{M}$
PEA	$\text{Expert}_e \in \mathcal{E}$	any $x_1...x_T$	prob. prediction	$\mathcal{E}$

Apart from this formal duality, there is a real connection between both bounds ...

PEA Bound reduced to Bayes Bound

Regard class of Bayes-predictors $\{\text{Bayes}_\nu : \nu \in \mathcal{M}\}$ as class of experts $\mathcal{E}$.

The corresponding FPL algorithm then satisfies PEA bound

$$\underline{\text{L}}\text{oss}_{1:T}(\text{PEA}) \leq \text{Loss}_{1:T}(\text{Bayes}_\mu) + c\cdot\sqrt{\text{Loss}_{1:T}(\text{Bayes}_\mu)k^\mu} + b\cdot k^\mu.$$

Take the μ-expectation, and use $\bar{\text{L}}\text{oss}_{1:T}(\text{Bayes}_\mu) \leq \bar{\text{L}}\text{oss}_{1:T}(\text{Any } \Lambda)$ and Jensen's inequality, to get a Bayes-like bound for PEA

$$\boxed{\underline{\bar{\text{L}}}\text{oss}(\text{PEA}) \leq \bar{\text{L}}\text{oss}_{1:T}(\text{Any } \Lambda) + c\cdot\sqrt{\bar{\text{L}}\text{oss}_{1:T}(\text{Any } \Lambda)\cdot k^\mu} + b\cdot k^\mu \quad \forall \mu \in \mathcal{M}}$$

Ignoring details, instead of using Bayes_ξ, one may use
PEA with same/similar performance guarantees as Bayes_ξ.

Additionally, PEA has worst-case guarantees, which Bayes lacks.

So why use Bayes at all?

Prediction with Expert Advice: Summary

- PEA considers *any* environmental sequence (worst case approach).
- PEA predicts nearly as well as the best expert of some class.
- Major PEA variants:
  "Weighted Majority" and "Follow the Perturbed Leader".
- Loss bounds for adaptive learning rate and both
  finite expert classes with uniform weights and
  countable expert classes with arbitrary weights.
- Approach and corresponding performance bounds are related or dual to the Bayesian and MDL approaches to sequence prediction.

Wrap Up

- Setup: Given (non)iid data $D = (x_1, ..., x_n)$, predict x_{n+1}
- Ultimate goal is to maximize profit or minimize loss
- Consider Models/Hypothesis $H_i \in \mathcal{M}$
- Max.Likelihood: $H_{best} = \arg\max_i p(D|H_i)$ (overfits if $\mathcal{M}$ large)
- Bayes: Posterior probability of H_i is $p(H_i|D)$
- MDL: $H_{best} = \arg\min_{H_i}\{\text{CodeLength}(D|H_i) + \text{CodeLength}(H_i)\}$ (Complexity penalization)
- Bayes needs prior(H_i), MDL needs CodeLength(H_i)
- Occam+Epicurus: High prior for simple models with short codes.
- Kolmogorov/Solomonoff: Quantification of simplicity/complexity
- MDL & Bayes work if D is sampled from $H_{true} \in \mathcal{M}$
- Prediction with Expert Advice works w/o assumption on D.

Literature

- Y. Freund and R. E. Schapire. *A Decision-Theoretic Generalization of On-line Learning and an Application to Boosting.* Journal of Computer and System Sciences, 55(1):119–139, 1997. http://www.research.att.com/~schapire/papers/FreundSc95.ps.Z
- M. Hutter and J. Poland. *Prediction with Expert Advice by Following the Perturbed Leader for General Weights.* Proc. 15th International Conf. on Algorithmic Learning Theory (ALT). LNAI 3244, 279–293, 2004. http://arxiv.org/abs/cs.LG/0405043
- P. D. Grünwald, *Tutorial on Minimum Description Length.* Chapters 1 and 2 of Minimum Description Length: recent advances in theory and practice MIT Press, 2004, to appear. http://www.cwi.nl/~pdg/ftp/mdlintro.pdf
- R. Cilibrasi and P.M.B. Vitanyi, *Clustering by Compression.* CWI manuscript 2003. http://arXiv.org/abs/cs/0312044
- M. Hutter, *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability.* Springer, Berlin, 300 pages, 2004. http://www.idsia.ch/~marcus/ai/uaibook.htm