Laplace's Rule of Succession in Information Geometry

Yann Ollivier

CNRS & Paris-Saclay University, France

Sequential prediction

Sequential prediction problem: given observations $x_1, \ldots, x_t$, build a probabilistic model p^{t+1} for x_{t+1}, iteratively.

Sequential prediction

Sequential prediction problem: given observations $x_1, \ldots, x_t$, build a probabilistic model p^{t+1} for x_{t+1}, iteratively.

Example: given that w women and m men entered this room, what is the probability that the next person who enters is a woman/man?

Sequential prediction

Sequential prediction problem: given observations $x_1, \ldots, x_t$, build a probabilistic model p^{t+1} for x_{t+1}, iteratively.

Example: given that w women and m men entered this room, what is the probability that the next person who enters is a woman/man?

Common performance criterion for prediction: cumulated log-loss

$$L_T := -\sum_{t=0}^{T-1} \log p^{t+1}(x_{t+1} \mid x_{1\ldots t})$$

to be minimized.

Sequential prediction

Sequential prediction problem: given observations $x_1, \ldots, x_t$, build a probabilistic model p^{t+1} for x_{t+1}, iteratively.

Example: given that w women and m men entered this room, what is the probability that the next person who enters is a woman/man?

Common performance criterion for prediction: cumulated log-loss

$$L_T := -\sum_{t=0}^{T-1} \log p^{t+1}(x_{t+1} \mid x_{1\ldots t})$$

to be minimized.
This corresponds to compression cost, and is also equal to square loss for Gaussian models.

Maximum likelihood estimator

Maximum likelihood strategy: Fix a parametric model $p_\theta(x)$. At each time, the best parameter based on past observations:

$$\theta_t^{\mathrm{ML}} := \arg\max_\theta \left\{ \prod_{s\leqslant t} p_\theta(x_s) \right\}$$

$$= \arg\min_\theta \left\{ -\sum_{s\leqslant t} \log p_\theta(x_s) \right\}$$

Maximum likelihood estimator

Maximum likelihood strategy: Fix a parametric model $p_\theta(x)$. At each time, the best parameter based on past observations:

$$\theta_t^{\mathrm{ML}} := \arg\max_\theta \left\{ \prod_{s \leqslant t} p_\theta(x_s) \right\}$$
$$= \arg\min_\theta \left\{ -\sum_{s \leqslant t} \log p_\theta(x_s) \right\}$$

Then, predict x_{t+1} using this θ_t^{ML}:

Maximum likelihood estimator

Maximum likelihood strategy: Fix a parametric model $p_\theta(x)$. At each time, the best parameter based on past observations:

$$\theta_t^{\mathrm{ML}} := \arg\max_\theta \left\{ \prod_{s \leqslant t} p_\theta(x_s) \right\} = \arg\min_\theta \left\{ - \sum_{s \leqslant t} \log p_\theta(x_s) \right\}$$

Then, predict x_{t+1} using this θ_t^{ML}:

$$p^{\mathrm{ML}}(x_{t+1}|x_{1\ldots t}) := p_{\theta_t^{\mathrm{ML}}}(x_{t+1})$$

This is the maximum likelihood or "plug-in" estimator.

Maximum likelihood estimator

Maximum likelihood strategy: Fix a parametric model $p_\theta(x)$. At each time, the best parameter based on past observations:

$$\theta_t^{\mathrm{ML}} := \arg\max_\theta \left\{ \prod_{s\leqslant t} p_\theta(x_s) \right\}$$
$$= \arg\min_\theta \left\{ -\sum_{s\leqslant t} \log p_\theta(x_s) \right\}$$

Then, predict x_{t+1} using this θ_t^{ML}:

$$p^{\mathrm{ML}}(x_{t+1}|x_{1\ldots t}) := p_{\theta_t^{\mathrm{ML}}}(x_{t+1})$$

This is the maximum likelihood or “plug-in” estimator.

Heavily used in machine learning. Argmax often computed via gradient descent.

Problems with the ML estimator

Example: Given that there are w women and m men in this room, the next person to enter will be a man with probability

Problems with the ML estimator

Example: Given that there are w women and m men in this room, the next person to enter will be a man with probability

$$\frac{m}{w+m}$$

according to the ML estimator.

Problems with the ML estimator

Example: Given that there are w women and m men in this room, the next person to enter will be a man with probability

$$\frac{m}{w + m}$$

according to the ML estimator.

$\Longrightarrow$ The ML estimator is not satisfying:

Problems with the ML estimator

Example: Given that there are w women and m men in this room, the next person to enter will be a man with probability

$$\frac{m}{w+m}$$

according to the ML estimator.

$\Longrightarrow$ The ML estimator is not satisfying:

- How do you predict the first observation?

Problems with the ML estimator

Example: Given that there are w women and m men in this room, the next person to enter will be a man with probability

$$\frac{m}{w+m}$$

according to the ML estimator.

$\Longrightarrow$ The ML estimator is not satisfying:

- How do you predict the first observation?
- Zero-frequency problem: If you have seen only women so far, the probability to see a man is estimated to 0.

Problems with the ML estimator

Example: Given that there are w women and m men in this room, the next person to enter will be a man with probability

$$\frac{m}{w+m}$$

according to the ML estimator.

$\Longrightarrow$ The ML estimator is not satisfying:

- How do you predict the first observation?
- Zero-frequency problem: If you have seen only women so far, the probability to see a man is estimated to 0.
- Often overfits in machine learning.

Laplace's rule of succession

Laplace suggested a quick fix for these problems: add one to the counts of each possibility. That is, predict according to

$$p(\text{woman}) = \frac{w + 1}{w + m + 2} \qquad p(\text{man}) = \frac{m + 1}{w + m + 2}$$

instead of $\frac{w}{w+m}$ and $\frac{m}{w+m}$. This is Laplace's rule of succession.

Laplace's rule of succession

Laplace suggested a quick fix for these problems: add one to the counts of each possibility. That is, predict according to

$$p(\text{woman}) = \frac{w + 1}{w + m + 2} \qquad p(\text{man}) = \frac{m + 1}{w + m + 2}$$

instead of $\frac{w}{w+m}$ and $\frac{m}{w+m}$. This is Laplace's rule of succession.

- Solves the zero-frequency problem: After having seen t women and no men, the probability to see a man is estimated to $1/(t+2)$.

Laplace's rule of succession

Laplace suggested a quick fix for these problems: add one to the counts of each possibility. That is, predict according to

$$p(\text{woman}) = \frac{w+1}{w+m+2} \qquad p(\text{man}) = \frac{m+1}{w+m+2}$$

instead of $\frac{w}{w+m}$ and $\frac{m}{w+m}$. This is Laplace's rule of succession.

- Solves the zero-frequency problem: After having seen t women and no men, the probability to see a man is estimated to $1/(t+2)$.
- Generalizes to other discrete data ("additive smoothing").

Laplace's rule of succession

Laplace suggested a quick fix for these problems: add one to the counts of each possibility. That is, predict according to

$$p(\text{woman}) = \frac{w+1}{w+m+2} \qquad p(\text{man}) = \frac{m+1}{w+m+2}$$

instead of $\frac{w}{w+m}$ and $\frac{m}{w+m}$. This is Laplace's rule of succession.

- Solves the zero-frequency problem: After having seen t women and no men, the probability to see a man is estimated to $1/(t+2)$.
- Generalizes to other discrete data ("additive smoothing").
- May seem arbitrary but has a beautiful Bayesian interpretation.

Bayesian predictors

Bayesian predictors start with a parametric model $p_\theta(x)$ together with a prior $\alpha(\theta)$ on θ.

At time t, the next symbol x_{t+1} is predicted by mixing all possible models p_θ with all values of θ,

$$p^{\alpha\text{-Bayes}}(x_{t+1}|x_{1\ldots t}) = \int_\theta p_\theta(x_{t+1})\, q_t(\theta)\; \mathrm{d}\theta$$

where $q_t(\theta)$ is the Bayesian posterior on θ given data $x_{1\ldots t}$,

$$q_t(\theta) \propto \alpha(\theta) \prod_{s \leqslant t} p_\theta(x_s)$$

Bayesian predictors

Bayesian predictors start with a parametric model $p_\theta(x)$ together with a prior $\alpha(\theta)$ on θ.

At time t, the next symbol x_{t+1} is predicted by mixing all possible models p_θ with all values of θ,

$$p^{\alpha\text{-Bayes}}(x_{t+1}|x_{1\dots t}) = \int_\theta p_\theta(x_{t+1})\, q_t(\theta)\; \mathrm{d}\theta$$

where $q_t(\theta)$ is the Bayesian posterior on θ given data $x_{1\dots t}$,

$$q_t(\theta) \propto \alpha(\theta) \prod_{s \leqslant t} p_\theta(x_s)$$

Proposition (folklore)

For Bernoulli distributions on a binary variable, e.g., $\{$woman, man$\}$, Laplace's rule coincides with the Bayesian predictor with a uniform prior on the Bernoulli parameter $\theta \in [0;1]$.

Bayesian predictors (2)

Bayesian predictors

- solve the zero-frequency problem
- have theoretical guarantees
- do not overfit

Bayesian predictors (2)

Bayesian predictors

- solve the zero-frequency problem
- have theoretical guarantees
- do not overfit

  but

- are difficult to compute: one must perform an integral over θ, and keep track of the Bayesian posterior which is an arbitrary function of θ.

Bayesian predictors (2)

Bayesian predictors

- solve the zero-frequency problem
- have theoretical guarantees
- do not overfit

but

- are difficult to compute: one must perform an integral over θ, and keep track of the Bayesian posterior which is an arbitrary function of θ.

Is there a simple way to approximate Bayesian predictors that would generalize Laplace's rule?

Theorem

Let (p_θ) be an exponential family of probability distributions.

Theorem

Let (p_θ) be an exponential family of probability distributions.

Under suitable regularity conditions, these two predictors coincide at first order in $1/t$:

- *The Bayesian predictor using the non-informative Jeffreys prior $\alpha(\theta) \propto \sqrt{\det I(\theta)}$ with $I(\theta)$ the Fisher information matrix.*

Theorem

Let (p_θ) be an exponential family of probability distributions.

Under suitable regularity conditions, these two predictors coincide at first order in $1/t$:

- *The Bayesian predictor using the non-informative Jeffreys prior $\alpha(\theta) \propto \sqrt{\det I(\theta)}$ with $I(\theta)$ the Fisher information matrix.*
- *The average*

$$\frac{1}{2}p^{\mathrm{ML}} + \frac{1}{2}p^{\mathrm{SNML}}$$

  of the maximum likelihood predictor and the "sequential normalized maximum likelihood" predictor [Shtarkov 1987, Roos, Rissanen...]

Theorem

Let (p_θ) be an exponential family of probability distributions.

Under suitable regularity conditions, these two predictors coincide at first order in $1/t$:

- *The Bayesian predictor using the non-informative Jeffreys prior $\alpha(\theta) \propto \sqrt{\det I(\theta)}$ with $I(\theta)$ the Fisher information matrix.*
- *The average*

$$\frac{1}{2}p^{\mathrm{ML}} + \frac{1}{2}p^{\mathrm{SNML}}$$

  of the maximum likelihood predictor and the "sequential normalized maximum likelihood" predictor [Shtarkov 1987, Roos, Rissanen...]

"The predictors p and p' coincide at first order" means that

$$p'(x_{t+1}|x_{1...t}) = p(x_{t+1}|x_{1...t})\,(1 + O(1/t^2))$$

for any sequence (x_t), assuming both are non-zero.

The sequential normalized maximum likelihood predictor

The SNML predictor p^{SNML} is defined as follows. For each possible value y of x_{t+1}, let

$$\theta^{\mathrm{ML}+y} := \arg\max_{\theta} \left\{ \log p_\theta(y) + \sum_{s \leqslant t} \log p_\theta(x_s) \right\}$$

be the value of the ML estimator if this value of x_{t+1} had already been observed.

The sequential normalized maximum likelihood predictor

The SNML predictor p^{SNML} is defined as follows. For each possible value y of x_{t+1}, let

$$\theta^{\mathrm{ML}+y} := \arg\max_{\theta} \left\{ \log p_\theta(y) + \sum_{s \leqslant t} \log p_\theta(x_s) \right\}$$

be the value of the ML estimator if this value of x_{t+1} had already been observed.

Define

$$q(y) := p_{\theta^{\mathrm{ML}+y}}(y)$$

Usually q is not a probability distribution, $\int_y q(y) > 1$.

The sequential normalized maximum likelihood predictor

The SNML predictor p^{SNML} is defined as follows. For each possible value y of x_{t+1}, let

$$\theta^{\mathrm{ML}+y} := \arg\max_\theta \left\{ \log p_\theta(y) + \sum_{s \leqslant t} \log p_\theta(x_s) \right\}$$

be the value of the ML estimator if this value of x_{t+1} had already been observed.

Define

$$q(y) := p_{\theta^{\mathrm{ML}+y}}(y)$$

Usually q is not a probability distribution, $\int_y q(y) > 1$.
$\Longrightarrow$ Rescale q:

$$p^{\mathrm{SNML}} := \frac{q}{\int q}$$

and use this for prediction of x_{t+1}.

Example: For a Bernoulli distribution, the SNML predictor is the same as Laplace's rule.

Example: For a Bernoulli distribution, the SNML predictor is the same as Laplace's rule.

The theorem states that the Bayesian predictor with canonical Jeffreys prior is approximately the average of the ML and SNML predictors.

Example: For a Bernoulli distribution, the SNML predictor is the same as Laplace's rule.

The theorem states that the Bayesian predictor with canonical Jeffreys prior is approximately the average of the ML and SNML predictors.

$\Longrightarrow$ For Bernoulli distributions, we recover the "add-one-half" rule for the Jeffreys prior (Krichevsky–Trofimov estimator).

Example: For a Bernoulli distribution, the SNML predictor is the same as Laplace's rule.

The theorem states that the Bayesian predictor with canonical Jeffreys prior is approximately the average of the ML and SNML predictors.

$\Longrightarrow$ For Bernoulli distributions, we recover the "add-one-half" rule for the Jeffreys prior (Krichevsky–Trofimov estimator).

- Relatively easy to compute

Example: For a Bernoulli distribution, the SNML predictor is the same as Laplace's rule.

The theorem states that the Bayesian predictor with canonical Jeffreys prior is approximately the average of the ML and SNML predictors.

$\Longrightarrow$ For Bernoulli distributions, we recover the "add-one-half" rule for the Jeffreys prior (Krichevsky–Trofimov estimator).

- Relatively easy to compute
- Different estimators usually differ at first order in $1/t$ (e.g., ML estimator or Bayesian estimators with different priors). The theorem is precise at first order in $1/t$ so recovers these differences.

Example: For a Bernoulli distribution, the SNML predictor is the same as Laplace's rule.

The theorem states that the Bayesian predictor with canonical Jeffreys prior is approximately the average of the ML and SNML predictors.

$\Longrightarrow$ For Bernoulli distributions, we recover the "add-one-half" rule for the Jeffreys prior (Krichevsky–Trofimov estimator).

- Relatively easy to compute
- Different estimators usually differ at first order in $1/t$ (e.g., ML estimator or Bayesian estimators with different priors). The theorem is precise at first order in $1/t$ so recovers these differences.
- Multiplicative error $(1 + O(1/t^2))$ in the theorem yields at most a bounded difference on cumulated log-loss.

Corollary: From ML to Bayes

For exponential families, there is an explicit approximate formula to compute the Bayesian predictor with Jeffreys prior if the ML predictor is known:

Corollary: From ML to Bayes

For exponential families, there is an explicit approximate formula to compute the Bayesian predictor with Jeffreys prior if the ML predictor is known:

$$p^{\mathrm{Jeffreys}}(x_{t+1}) \approx p^{\mathrm{ML}}(x_{t+1}) \left(1 + \frac{1}{2t} \left\|\partial_\theta \log p_\theta(x_{t+1})\right\|^2_{Fisher} - \frac{\dim(\theta)}{2t}\right)$$

up to $O(1/t^2)$.

Here $\|\partial_\theta \log p_\theta(x_{t+1})\|_{Fisher}$ is the norm of the gradient of $\log p_\theta(x_{t+1})$ in the Riemannian metric given by the Fisher information matrix. (Compare "flattened ML" [Kotłowski–Grünwald–de Rooij 2010].)

Corollary: From ML to Bayes

For exponential families, there is an explicit approximate formula to compute the Bayesian predictor with Jeffreys prior if the ML predictor is known:

$$p^{\mathrm{Jeffreys}}(x_{t+1}) \approx p^{\mathrm{ML}}(x_{t+1}) \left(1 + \frac{1}{2t} \left\|\partial_\theta \log p_\theta(x_{t+1})\right\|^2_{Fisher} - \frac{\dim(\theta)}{2t}\right)$$

up to $O(1/t^2)$.

Here $\|\partial_\theta \log p_\theta(x_{t+1})\|_{Fisher}$ is the norm of the gradient of $\log p_\theta(x_{t+1})$ in the Riemannian metric given by the Fisher information matrix. (Compare "flattened ML" [Kotłowski–Grünwald–de Rooij 2010].)

Note: valid only when these probabilities are non-zero, so does not solve the zero-frequency problem.

From ML to Bayes (2)

Proof of the corollary (idea):

$$\theta^{\mathrm{ML}+x_{t+1}} \approx \theta^{\mathrm{ML}} + \frac{1}{t}\,\tilde{\nabla}_\theta \log p_\theta(x_{t+1})$$

where $\tilde{\nabla}_\theta = I(\theta)^{-1}\frac{\partial}{\partial\theta}$ is Amari's natural gradient given by the Fisher matrix.

"When adding a data point, the ML estimator moves by $1/t$ times the natural gradient of the new point's log-likelihood."

From ML to Bayes (2)

Proof of the corollary (idea):

$$\theta^{\mathrm{ML}+x_{t+1}} \approx \theta^{\mathrm{ML}} + \frac{1}{t}\,\tilde{\nabla}_\theta \log p_\theta(x_{t+1})$$

where $\tilde{\nabla}_\theta = I(\theta)^{-1}\frac{\partial}{\partial\theta}$ is Amari's natural gradient given by the Fisher matrix.

"When adding a data point, the ML estimator moves by $1/t$ times the natural gradient of the new point's log-likelihood."

What about other priors?

Arbitrary Bayesian priors

Consider a Bayesian prior with density $\beta(\theta)$ with respect to the Jeffreys prior

$$\alpha(\theta) = \beta(\theta)\sqrt{\det I(\theta)}$$

Arbitrary Bayesian priors

Consider a Bayesian prior with density $\beta(\theta)$ with respect to the Jeffreys prior

$$\alpha(\theta) = \beta(\theta)\sqrt{\det I(\theta)}$$

Then a similar theorem holds if the definition of the SNML predictor p^{SNML} is modified as

$$q(y) := \beta\left(\theta^{\mathrm{ML}+y}\right)^2 \, p_{\theta^{\mathrm{ML}+y}}(y)\,, \qquad p^{\mathrm{SNML}} := \frac{q}{\int q}$$

for each possible value y of x_{t+1}.

Posterior means: a systematic shift between ML and Bayes

Another way to compare ML and Bayesian predictors is to use test functions.

Posterior means: a systematic shift between ML and Bayes

Another way to compare ML and Bayesian predictors is to use test functions.

"Is the Bayesian posterior approximately centered around the ML estimator?"

Posterior means: a systematic shift between ML and Bayes

Another way to compare ML and Bayesian predictors is to use test functions.

"Is the Bayesian posterior approximately centered around the ML estimator?"

$\Longrightarrow$ No!

Posterior means: a systematic shift between ML and Bayes

Let $f(\theta)$ be a smooth test function of θ. Then there is a systematic direction of the difference between $f(\theta^{\mathrm{ML}})$ and the Bayesian posterior mean of f.

Posterior means: a systematic shift between ML and Bayes

Let $f(\theta)$ be a smooth test function of θ. Then there is a systematic direction of the difference between $f(\theta^{\mathrm{ML}})$ and the Bayesian posterior mean of f. For exponential families and the Jeffreys prior, this difference is approximately

$$\frac{1}{t}\,\partial_\theta f \cdot V(\theta^{\mathrm{ML}})$$

where $V(\theta)$ is an intrinsic vector field on Θ, independent from f:

Posterior means: a systematic shift between ML and Bayes

Let $f(\theta)$ be a smooth test function of θ. Then there is a systematic direction of the difference between $f(\theta^{\mathrm{ML}})$ and the Bayesian posterior mean of f. For exponential families and the Jeffreys prior, this difference is approximately

$$\frac{1}{t}\,\partial_\theta f \cdot V(\theta^{\mathrm{ML}})$$

where $V(\theta)$ is an intrinsic vector field on Θ, independent from f:

$$V(\theta) = \frac{1}{4}\, I(\theta)^{-1} \cdot T(\theta) \cdot I(\theta)^{-1}$$

with I the Fisher matrix and T the skewness tensor [Amari–Nagaoka]

$$T(\theta)_{ijk} := \mathbb{E}_{x\sim p_\theta} \frac{\partial \ln p_\theta(x)}{\partial \theta^i} \frac{\partial \ln p_\theta(x)}{\partial \theta^j} \frac{\partial \ln p_\theta(x)}{\partial \theta^k}$$

Conclusions

- For exponential families, Bayesian predictors can be approximated using modified ML predictors.
- The difference between Bayesian and ML predictors can be computed from the Fisher metric.
- There is a systematic direction of the shift from ML to Bayesian posterior means.
- Extends to non-i.i.d. models if $p_\theta(x_{t+1}|x_{1\ldots t})$ is an exponential family.

Conclusions

- For exponential families, Bayesian predictors can be approximated using modified ML predictors.
- The difference between Bayesian and ML predictors can be computed from the Fisher metric.
- There is a systematic direction of the shift from ML to Bayesian posterior means.
- Extends to non-i.i.d. models if $p_\theta(x_{t+1}|x_{1\ldots t})$ is an exponential family.

Thank you!