Bayesian machine learning: a tutorial

Rémi Bardenet

CNRS & CRIStAL, Univ. Lille, France

Outline

Outline

Typical jobs for statisticians

Estimation

- You have data $x_1, \ldots, x_n$ that you assume drawn from $p(x_1, \ldots, x_n|\theta^\star)$, with $\theta^\star \in \mathbb{R}^d$.
- You want an estimate $\hat{\theta}(x_1, \ldots, x_n)$ of $\theta^\star \in \mathbb{R}^d$.

Confidence regions

- You have data $x_1, \ldots, x_n$ that you assume drawn from $p(x_1, \ldots, x_n \cdot |\theta^\star)$, with $\theta^\star \in \mathbb{R}^d$.
- You want a region $A(x_1, \ldots, x_n) \subset \mathbb{R}^d$ and make a statement that $\theta \in A(x_1, \ldots, x_n)$ with some certainty.

Typical jobs for statisticians

Estimation

- You have data $x_1, \ldots, x_n$ that you assume drawn from $p(x_1, \ldots, x_n|\theta^\star)$, with $\theta^\star \in \mathbb{R}^d$.
- You want an estimate $\hat{\theta}(x_1, \ldots, x_n)$ of $\theta^\star \in \mathbb{R}^d$.

Confidence regions

- You have data $x_1, \ldots, x_n$ that you assume drawn from $p(x_1, \ldots, x_n \cdot |\theta^\star)$, with $\theta^\star \in \mathbb{R}^d$.
- You want a region $A(x_1, \ldots, x_n) \subset \mathbb{R}^d$ and make a statement that $\theta \in A(x_1, \ldots, x_n)$ with some certainty.

Statistical decision theory[1]

Figure: Abraham Wald (1902–1950)

[1] A. Wald. *Statistical decision functions*. Wiley, 1950.

Statistical decision theory

- Let Θ be the "states of the world", typically the space of parameters of interest.
- Decisions are functions $d(x_1, \ldots, x_n) \in \mathcal{D}$.
- Let $L(d, \theta)$ denote the loss of making decision d when the state of the world is θ.
- Wald defines the risk of a decision as

$$R(d, \theta) = \int L(d, \theta) p(x_{1:n}|\theta) dx_{1:n}.$$

- Wald says d_1 is a better decision than d_2 if

$$\forall \theta \in \Theta, \quad L(d_1, \theta) \leqslant L(d_2, \theta). \tag{1}$$

- d is called admissible if there is no better decision than d.

Statistical decision theory

- Let Θ be the "states of the world", typically the space of parameters of interest.
- Decisions are functions $d(x_1, \ldots, x_n) \in \mathcal{D}$.
- Let $L(d, \theta)$ denote the loss of making decision d when the state of the world is θ.
- Wald defines the risk of a decision as

$$R(d, \theta) = \int L(d, \theta) p(x_{1:n}|\theta) dx_{1:n}.$$

- Wald says d_1 is a better decision than d_2 if

$$\forall \theta \in \Theta, \quad L(d_1, \theta) \leqslant L(d_2, \theta). \tag{1}$$

- d is called admissible if there is no better decision than d.

Statistical decision theory

- Let Θ be the "states of the world", typically the space of parameters of interest.
- Decisions are functions $d(x_1, \ldots, x_n) \in \mathcal{D}$.
- Let $L(d, \theta)$ denote the loss of making decision d when the state of the world is θ.
- Wald defines the risk of a decision as

$$R(d, \theta) = \int L(d, \theta) p(x_{1:n}|\theta) dx_{1:n}.$$

- Wald says d_1 is a better decision than d_2 if

$$\forall \theta \in \Theta, \quad L(d_1, \theta) \leqslant L(d_2, \theta). \tag{1}$$

- d is called admissible if there is no better decision than d.

Statistical decision theory

- Let Θ be the "states of the world", typically the space of parameters of interest.
- Decisions are functions $d(x_1, \ldots, x_n) \in \mathcal{D}$.
- Let $L(d, \theta)$ denote the loss of making decision d when the state of the world is θ.
- Wald defines the risk of a decision as

$$R(d, \theta) = \int L(d, \theta) p(x_{1:n}|\theta) dx_{1:n}.$$

- Wald says d_1 is a better decision than d_2 if

$$\forall \theta \in \Theta, \quad L(d_1, \theta) \leqslant L(d_2, \theta). \tag{1}$$

- d is called admissible if there is no better decision than d.

Statistical decision theory

- Let Θ be the "states of the world", typically the space of parameters of interest.
- Decisions are functions $d(x_1, \ldots, x_n) \in \mathcal{D}$.
- Let $L(d, \theta)$ denote the loss of making decision d when the state of the world is θ.
- Wald defines the risk of a decision as

$$R(d, \theta) = \int L(d, \theta) p(x_{1:n}|\theta) dx_{1:n}.$$

- Wald says d_1 is a better decision than d_2 if

$$\forall \theta \in \Theta, \quad L(d_1, \theta) \leqslant L(d_2, \theta). \tag{1}$$

- d is called admissible if there is no better decision than d.

Statistical decision theory

- Let Θ be the "states of the world", typically the space of parameters of interest.
- Decisions are functions $d(x_1, \ldots, x_n) \in \mathcal{D}$.
- Let $L(d, \theta)$ denote the loss of making decision d when the state of the world is θ.
- Wald defines the risk of a decision as

$$R(d, \theta) = \int L(d, \theta) p(x_{1:n}|\theta) dx_{1:n}.$$

- Wald says d_1 is a better decision than d_2 if

$$\forall \theta \in \Theta, \quad L(d_1, \theta) \leqslant L(d_2, \theta). \tag{1}$$

- d is called admissible if there is no better decision than d.

Illustration with a simple estimation problem

- You have data $x_1, \ldots, x_n$ that you assume drawn from

$$p(x_1, \ldots, x_n | \theta^\star) = \prod_{i=1}^{n} \mathcal{N}(x_i | \theta^\star, \sigma^2),$$

  and you know σ^2.
- You choose a loss function, say $L(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|^2$.
- You restrict your decision space to unbiased estimators.
- The sample mean $\tilde{\theta} := n^{-1} \sum_{i=1}^{n} x_i$ is unbiased, and has minimum variance among unbiased estimators.
- Since

$$R(\tilde{\theta}, \theta) = \mathrm{Var}\tilde{\theta},$$

  $\tilde{\theta}$ is the best decision you can make in Wald's framework.

Wald's view of frequentist estimation

Estimation

- You have data $x_1, \ldots, x_n$ that you assume drawn from $p(x_1, \ldots, x_n|\theta^\star)$, with $\theta^\star \in \mathbb{R}^d$.
- You want an estimate $\hat{\theta}(x_1, \ldots, x_n)$ of $\theta^\star \in \mathbb{R}^d$.

A Waldian answer

- Our decisions are estimates $d(x_1, \ldots, x_n) = \hat{\theta}(x_1, \ldots, x_n)$.
- We pick a loss, say $L(d, \theta) = L(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|^2$.
- If you have an unbiased estimator with minimum variance, then this is the best decision among unbiased estimators.

Wald's view of frequentist estimation

Estimation

- You have data $x_1, \ldots, x_n$ that you assume drawn from $p(x_1, \ldots, x_n|\theta^\star)$, with $\theta^\star \in \mathbb{R}^d$.
- You want an estimate $\hat{\theta}(x_1, \ldots, x_n)$ of $\theta^\star \in \mathbb{R}^d$.

A Waldian answer

- Our decisions are estimates $d(x_1, \ldots, x_n) = \hat{\theta}(x_1, \ldots, x_n)$.
- In general, the loss can be more complex and unbiased estimors unknown/irrelevant.
- In these cases, you may settle for a minimax estimator

$$\hat{\theta}(x_1, \ldots, x_n) = \arg\min_d \sup_\theta R(d, \theta).$$

Wald's is only one view of frequentist statistics...

- On estimation, some would argue in favour of the maximum likelihood[2].

Figure: Ronald Fisher (1890–1962)

[2] S. M. Stigler. "The epic story of maximum likelihood". In: *Statistical Science* (2007), pp. 598–620.

... but bear with me, since it is predominant in machine learning

For instance, supervised learning is usually formalized as

$$g^\star = \arg\min_g \mathbb{E} L(y, g(x)). \tag{2}$$

which you approximate by

$$\hat{g} = \arg\min_g \sum_{i=1}^{n} L(y_i, g(x_i)) + \text{penalty}(g),$$

while trying to control the excess risk

$$\mathbb{E} L(y, \hat{g}(x)) - \mathbb{E} L(y, g^\star(x)).$$

Wald's view of frequentist confidence regions

Confidence regions

- You have data $x_1, \ldots, x_n$ that you assume drawn from $p(x_1, \ldots, x_n \cdot |\theta^\star)$, with $\theta^\star \in \mathbb{R}^d$.
- You want a region $A(x_1, \ldots, x_n) \subset \mathbb{R}^d$ and make a statement that $\theta \in A(x_1, \ldots, x_n)$ with some certainty.

A Waldian answer

- Our decisions are subsets of $\mathbb{R}^d$: $d(x_{1:n}) = A(x_{1:n})$.
- A common loss is $L(d, \theta) = L(A, \theta) = 1_{\theta \notin A} + \gamma |A|$.
- So you want to find $A(x_{1:n})$ that minimizes

$$L(A, \theta) = \int [1_{\theta^\star \notin A} p(x_{1:n}|\theta^\star) + \gamma |A|] \, dx_{1:n}.$$

Illustration with a simple confidence interval problem

- You have data $x_1, \ldots, x_n$ that you assume drawn from

$$p(x_1, \ldots, x_n | \theta^\star) = \prod_{i=1}^{n} \mathcal{N}(x_i | \theta^\star, \sigma^2).$$

- You choose a loss function, say $L(A, \theta) = 1_{\theta \notin A} + \gamma |A|$.
- You restrict your decisions to intervals centered around the sample mean $\tilde{\theta}$.
- Since $\frac{\theta - \tilde{\theta}}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$, we know (exercise) that for

$$\tilde{A} := [\tilde{\theta} - k\sigma/\sqrt{n}, \tilde{\theta} + k\sigma/\sqrt{n}],$$

  it comes

$$R(A, \theta) = \mathbb{P}(|\mathcal{N}(0,1)| \geqslant k) + \frac{2\gamma k \sigma}{\sqrt{n}}.$$

- All is left to do is choose k.
- Textbook examples bypass the need for γ: they fix $\alpha > 0$ and find the smallest k such that $\mathbb{P}(|\mathcal{N}(0,1)| \geqslant k) \leqslant \alpha$.

Summary so far

- Waldian frequentists measure risks as expectations w.r.t. the data-generating process.

$$R(d, \theta) = \int L(d(x_{1:n}), \theta) p(x_{1:n}|\theta) dx_{1:n}$$

- One major difficulty is that the risk remains a function of θ.
- Without additional structure (unbiasedness, Gaussianity, etc.), it is difficult to go beyond minimax rules.

Idea

What if we introduced a distribution on Θ, and tried to minimize

$$\begin{aligned} r(d) &= \int R(d, \theta) p(\theta) d\theta \\ &= \int \left[\int L(d(x_{1:n}), \theta) p(x_{1:n}|\theta) dx_{1:n} \right] p(\theta) d\theta \end{aligned}$$

From expected frequentist loss to posterior expected loss

$$
\begin{aligned}
r(d) &= \int R(d,\theta)p(\theta)d\theta \\
&= \int \left[\int L(d(x_{1:n}),\theta)p(x_{1:n}|\theta)dx_{1:n} \right] p(\theta)d\theta \\
&= \int \left[\int L(d(x_{1:n}),\theta)p(x_{1:n}|\theta)p(\theta)d\theta \right] dx_{1:n} \\
&= \int \left[\int L(d(x_{1:n}),\theta)\frac{p(x_{1:n}|\theta)p(\theta)}{p(x_{1:n})}d\theta \right] p(x_{1:n})dx_{1:n} \\
&= \int \left[\int L(d(x_{1:n}),\theta)p(\theta|x_{1:n})d\theta \right] p(x_{1:n})dx_{1:n}
\end{aligned}
$$

Bayesians minimize posterior expected utility

The posterior expected utility paradigm: Bayes rules

Pick d to solve

$$\arg\min_d \int L(d(x_{1:n}), \theta) p(\theta|x_{1:n}) d\theta.$$

Bayes rules have good frequentist properties[3]

- Under general conditions, Bayes decision rules are admissible, all admissible rules are limits of Bayes rules.
- Bayes rules with "least favourable priors" are minimax.

[3] G. Parmigiani and L. Inoue. *Decision theory: principles and approaches*. Vol. 812. John Wiley & Sons, 2009.

Illustration with a simple estimation problem

- You have data $x_1, \ldots, x_n$ that you assume drawn from

$$p(x_1, \ldots, x_n | \theta^\star) = \prod_{i=1}^{n} \mathcal{N}(x_i | \theta^\star, \sigma^2),$$

  and you know σ^2.
- You choose a loss function, say $L(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|^2$.
- You choose a prior p over θ.
- Your Bayes decision minimizes

$$\int \|\hat{\theta} - \theta\|^2 p(\theta | x_{1:n}) d\theta,$$

  so you pick

$$\hat{\theta} = \int \theta p(\theta | x_{1:n}) d\theta.$$

- Conceptually, it is simpler. In practice, you need to compute an integral.

Illustration with a simple confidence interval problem

- You have data $x_1, \ldots, x_n$ that you assume drawn from

$$p(x_1, \ldots, x_n | \theta^\star) = \prod_{i=1}^{n} \mathcal{N}(x_i | \theta^\star, \sigma^2).$$

- You choose a loss function, say $L(A, \theta) = 1_{\theta \notin A} + \gamma |A|$.
- You choose a prior p over θ.
- Your Bayes decision minimizes

$$\int 1_{\theta \notin A} p(\theta | x_{1:n}) d\theta + \gamma |A|,$$

- Conceptually, it is simpler. In practice, you need to carefully pick your prior and/or restrict the decision space and/or compute many integrals.

Summary so far

- Bayes rules fit into Wald's framework.
- For a fixed prior, the Bayesian risk completely orders decision rules.
- The key idea is posterior expected utility.
- You can answer most basic statistical questions using this principle: [more examples].

A recent motivating success

PRL **119,** 141101 (2017) PHYSICAL REVIEW LETTERS week ending 6 OCTOBER 2017

GW170814: A Three-Detector Observation of Gravitational Waves from a Binary Black Hole Coalescence

B. P. Abbott *et al.*[*]
(LIGO Scientific Collaboration and Virgo Collaboration)
(Received 23 September 2017; published 6 October 2017)

On August 14, 2017 at 10:30:43 UTC, the Advanced Virgo detector and the two Advanced LIGO detectors coherently observed a transient gravitational-wave signal produced by the coalescence of two stellar mass black holes, with a false-alarm rate of $\lesssim 1$ in 27 000 years. The signal was observed with a three-detector network matched-filter signal-to-noise ratio of 18. The inferred masses of the initial black holes are $30.5^{+5.7}_{-3.0} M_\odot$ and $25.3^{+2.8}_{-4.2} M_\odot$ (at the 90% credible level). The luminosity distance of the source is 540^{+130}_{-210} Mpc, corresponding to a redshift of $z = 0.11^{+0.03}_{-0.04}$. A network of three detectors improves the sky localization of the source, reducing the area of the 90% credible region from 1160 deg^2 using only the two LIGO detectors to 60 deg^2 using all three detectors. For the first time, we can test the nature of gravitational-wave polarizations from the antenna response of the LIGO-Virgo network, thus enabling a new class of phenomenological tests of gravity.

DOI: 10.1103/PhysRevLett.119.141101

I. INTRODUCTION

The era of gravitational-wave (GW) astronomy began with the detection of binary black hole (BBH) mergers, by the Advanced Laser Interferometer Gravitational-Wave Observatory (LIGO) detectors [1] during the first of the waveform obtained from analysis of the LIGO detectors' data alone, we find that the probability, in 5000 s of data around the event, of a peak in SNR from Virgo data due to noise and as large as the one observed, within a time window determined by the maximum possible time of

Outline

The subjectivistic viewpoint

- Top requirement is internal coherence of decisions.
- Favourizes interpreting probability distributions as personal beliefs.

Figure: Bruno de Finetti (1906–1985) and L. Jimmie Savage (1917–1971)

The logical justification

- Top requirement is to find a version of propositional logic that allows taking into account uncertainty.
- Also favourizes interpreting probability distributions as beliefs, but aims for objective priors.

Figure: Richart T. Cox (1898–1991), Edwin T. Jaynes (1917–1971), and Harold Jeffreys (1891–1989)

The hybrid view[4]

- The starting point is posterior expected utility, loosely justified by Wald's theory.
- It is simple, widely applicable, has good frequentist properties.
- It satisfies the likelihood principle.
- It is easy to interpret: beliefs are
  - represented by probabilities,
  - updated using Bayes' rule,
  - integrated when making decisions.
- It is easy to communicate your uncertainty
  - Simply give your posterior.
  - When making a decision, make sure that the priors of everyone involved would yield the same decision.

[4] C. P. Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

Practical advantages of posterior expected utility

- Conceptually answers all ML problems.
- Suits all applications where quantifying uncertainty is vital vs computational complexity: all basic sciences, health, even one-shot commercial decisions.
- We never invoked any large-sample argument, so suits all sizes of datasets.

Outline

Conjugacy

- Say we have a linear regression problem

$$y_i = f(x_i) + \varepsilon_i,$$

$f(x) = \theta^T x$, ε_i i.i.d. Gaussians $\mathcal{N}(0, \sigma^2)$.

- If we choose $p(\theta) \sim \mathcal{N}(0, \Sigma)$, then (exercise)

$$\begin{aligned} p(\theta|(x,y)_{1:n}) &\propto p((x,y)_{1:n}|\theta)p(\theta) \\ &= \mathcal{N}(\sigma^{-2}A^{-1}X\mathbf{y}, A^{-1}), \end{aligned}$$

where $A = \sigma^{-2}X^T X + \Sigma^{-1}$.

- If the loss is not too complicated, then integrals are easy. For instance, prediction with L^2 loss is simple:

Conjugacy

- Say we have a linear regression problem

$$y_i = f(x_i) + \varepsilon_i,$$

  $f(x) = \theta^T x$, ε_i i.i.d. Gaussians $\mathcal{N}(0, \sigma^2)$.
- If we choose $p(\theta) \sim \mathcal{N}(0, \Sigma)$, then (exercise)

$$\begin{aligned} p(\theta|(x,y)_{1:n}) &\propto p((x,y)_{1:n}|\theta)p(\theta) \\ &= \mathcal{N}(\sigma^{-2}A^{-1}X\mathbf{y}, A^{-1}), \end{aligned}$$

  where $A = \sigma^{-2}X^T X + \Sigma^{-1}$.
- If the loss is not too complicated, then integrals are easy. For instance, prediction with L^2 loss is simple:

Conjugacy

- Say we have a linear regression problem

$$y_i = f(x_i) + \varepsilon_i,$$

  $f(x) = \theta^T x$, ε_i i.i.d. Gaussians $\mathcal{N}(0, \sigma^2)$.
- If we choose $p(\theta) \sim \mathcal{N}(0, \Sigma)$, then (exercise)

$$\begin{aligned} p(\theta|(x,y)_{1:n}) &\propto p((x,y)_{1:n}|\theta)p(\theta) \\ &= \mathcal{N}(\sigma^{-2}A^{-1}X\mathbf{y}, A^{-1}), \end{aligned}$$

  where $A = \sigma^{-2}X^T X + \Sigma^{-1}$.
- If the loss is not too complicated, then integrals are easy. For instance, prediction with L^2 loss is simple:

$$\arg\min_{y_\star} \int \|y_\star - f(x_\star)\|^2 p(\theta|(x,y)_{1:n}) d\theta$$

Conjugacy

- Say we have a linear regression problem

$$y_i = f(x_i) + \varepsilon_i,$$

  $f(x) = \theta^T x$, ε_i i.i.d. Gaussians $\mathcal{N}(0, \sigma^2)$.
- If we choose $p(\theta) \sim \mathcal{N}(0, \Sigma)$, then (exercise)

$$\begin{aligned} p(\theta|(x,y)_{1:n}) &\propto p((x,y)_{1:n}|\theta)p(\theta) \\ &= \mathcal{N}(\sigma^{-2}A^{-1}X\mathbf{y}, A^{-1}), \end{aligned}$$

  where $A = \sigma^{-2}X^TX + \Sigma^{-1}$.
- If the loss is not too complicated, then integrals are easy. For instance, prediction with L^2 loss is simple:

$$\arg\min_{y_\star} \int \|y_\star - \theta^T x_\star\|^2 p(\theta|(x,y)_{1:n}) d\theta$$

Conjugacy

- Say we have a linear regression problem

$$y_i = f(x_i) + \varepsilon_i,$$

  $f(x) = \theta^T x$, ε_i i.i.d. Gaussians $\mathcal{N}(0, \sigma^2)$.
- If we choose $p(\theta) \sim \mathcal{N}(0, \Sigma)$, then (exercise)

$$\begin{aligned} p(\theta|(x,y)_{1:n}) &\propto p((x,y)_{1:n}|\theta)p(\theta) \\ &= \mathcal{N}(\sigma^{-2}A^{-1}X\mathbf{y}, A^{-1}), \end{aligned}$$

  where $A = \sigma^{-2}X^TX + \Sigma^{-1}$.
- If the loss is not too complicated, then integrals are easy. For instance, prediction with L^2 loss is simple:

$$\hat{y}_\star := \sigma^{-2}x_\star^T A^{-1} X\mathbf{y}.$$

Monte Carlo methods

- Sometimes, you're less lucky. Say we're doing logistic regression.

$$y_i = \text{Bernoulli}\left[\sigma(f(x_i))\right],$$

with $f(x) = \theta^T x$, $\sigma(x) = 1/(1+e^{-x})$.

- Even if we choose $p(\theta) \sim \mathcal{N}(0, \Sigma)$,

$$p(\theta|(x,y)_{1:n}) \propto p((x,y)_{1:n}|\theta)p(\theta)$$

does not have a simple closed form.

- We need powerful numerical integration methods, that is, constructions of nodes (θ_i) and weights w_i such that

$$\int h d\pi \approx \sum_{i=1}^{N} w_i h(\theta_i).$$

Monte Carlo methods

- Sometimes, you're less lucky. Say we're doing logistic regression.

$$y_i = \text{Bernoulli}\left[\sigma(f(x_i))\right],$$

with $f(x) = \theta^T x$, $\sigma(x) = 1/(1+e^{-x})$.

- Even if we choose $p(\theta) \sim \mathcal{N}(0, \Sigma)$,

$$p(\theta|(x,y)_{1:n}) \propto p((x,y)_{1:n}|\theta)p(\theta)$$

does not have a simple closed form.

- We need powerful numerical integration methods, that is, constructions of nodes (θ_i) and weights w_i such that

$$\int h d\pi \approx \sum_{i=1}^{N} w_i h(\theta_i).$$

Monte Carlo methods

- Sometimes, you're less lucky. Say we're doing logistic regression.

$$y_i = \text{Bernoulli}\left[\sigma(f(x_i))\right],$$

with $f(x) = \theta^T x$, $\sigma(x) = 1/(1 + e^{-x})$.

- Even if we choose $p(\theta) \sim \mathcal{N}(0, \Sigma)$,

$$p(\theta|(x, y)_{1:n}) \propto p((x, y)_{1:n}|\theta)p(\theta)$$

does not have a simple closed form.

- We need powerful numerical integration methods, that is, constructions of nodes (θ_i) and weights w_i such that

$$\int h d\pi \approx \sum_{i=1}^{N} w_i h(\theta_i).$$

Metropolis-Hastings

$\mathrm{MH}\big(\pi(\theta),\, q(\theta'|\theta),\, \theta_0,\, N_{\text{iter}}\big)$

1 **for** $k \leftarrow 1$ **to** N_{iter}

2 $\quad \theta \leftarrow \theta_{k-1}$

3 $\quad \theta' \sim q(.|\theta),\, u \sim \mathcal{U}_{(0,1)},$

4 $\quad \alpha = \frac{\pi(\theta')}{\pi(\theta)} \frac{q(\theta|\theta')}{q(\theta'|\theta)}$

5 $\quad$ **if** $u < \alpha$

6 $\quad\quad \theta_k \leftarrow \theta'$ $\quad \triangleright$ *Accept*

7 $\quad$ **else** $\theta_k \leftarrow \theta$ $\quad \triangleright$ *Reject*

8 **return** $(\theta_k)_{k=1,\ldots,N_{\text{iter}}}$

Metropolis-Hastings

$\mathrm{MH}\big(\pi(\theta),\, q(\theta'|\theta),\, \theta_0,\, N_{\text{iter}}\big)$

1 **for** $k \leftarrow 1$ **to** N_{iter}

2 $\quad \theta \leftarrow \theta_{k-1}$

3 $\quad \theta' \sim q(.|\theta),\ u \sim \mathcal{U}_{(0,1)},$

4 $\quad \alpha = \frac{\pi(\theta')}{\pi(\theta)} \frac{q(\theta|\theta')}{q(\theta'|\theta)}$

5 $\quad$ **if** $u < \alpha$

6 $\quad\quad \theta_k \leftarrow \theta' \qquad \triangleright\ Accept$

7 $\quad$ **else** $\theta_k \leftarrow \theta \qquad \triangleright\ Reject$

8 **return** $(\theta_k)_{k=1,\dots,N_{\text{iter}}}$

Metropolis-Hastings

$\mathrm{MH}\big(\pi(\theta),\, q(\theta'|\theta),\, \theta_0,\, N_{\text{iter}}\big)$

1 **for** $k \leftarrow 1$ **to** N_{iter}

2 $\quad \theta \leftarrow \theta_{k-1}$

3 $\quad \theta' \sim q(.|\theta),\ u \sim \mathcal{U}_{(0,1)},$

4 $\quad \alpha = \frac{\pi(\theta')}{\pi(\theta)} \frac{q(\theta|\theta')}{q(\theta'|\theta)}$

5 $\quad$ **if** $u < \alpha$

6 $\quad\quad \theta_k \leftarrow \theta' \qquad \triangleright$ *Accept*

7 $\quad$ **else** $\theta_k \leftarrow \theta \qquad \triangleright$ *Reject*

8 **return** $(\theta_k)_{k=1,\ldots,N_{\text{iter}}}$

Metropolis-Hastings

$\mathrm{MH}\big(\pi(\theta),\, q(\theta'|\theta),\, \theta_0,\, N_{\text{iter}}\big)$

1 **for** $k \leftarrow 1$ **to** N_{iter}

2 $\quad \theta \leftarrow \theta_{k-1}$

3 $\quad \theta' \sim q(.|\theta),\; u \sim \mathcal{U}_{(0,1)},$

4 $\quad \alpha = \frac{\pi(\theta')}{\pi(\theta)} \frac{q(\theta|\theta')}{q(\theta'|\theta)}$

5 $\quad$ **if** $u < \alpha$

6 $\quad\quad \theta_k \leftarrow \theta'$ ▷ *Accept*

7 $\quad$ **else** $\theta_k \leftarrow \theta$ ▷ *Reject*

8 **return** $(\theta_k)_{k=1,\ldots,N_{\text{iter}}}$

Metropolis-Hastings

$\mathrm{MH}\big(\pi(\theta),\, q(\theta'|\theta),\, \theta_0,\, N_{\text{iter}}\big)$

1 **for** $k \leftarrow 1$ **to** N_{iter}

2 $\quad \theta \leftarrow \theta_{k-1}$

3 $\quad \theta' \sim q(.|\theta),\; u \sim \mathcal{U}_{(0,1)},$

4 $\quad \alpha = \frac{\pi(\theta')}{\pi(\theta)} \frac{q(\theta|\theta')}{q(\theta'|\theta)}$

5 $\quad$ **if** $u < \alpha$

6 $\quad\quad \theta_k \leftarrow \theta'$ ▷ *Accept*

7 $\quad$ **else** $\theta_k \leftarrow \theta$ ▷ *Reject*

8 **return** $(\theta_k)_{k=1,\ldots,N_{\text{iter}}}$

The MCMC magic

- Under assumptions[5],

$$\sqrt{N_{\text{iter}}}\left[\frac{1}{N_{\text{iter}}}\sum_{k=0}^{N_{\text{iter}}} h(\theta_k) - \int h(\theta)\,\pi(\theta)d\theta\right] \to \mathcal{N}(0, \sigma^2_{\text{lim}}(h)),$$

- If you choose q carefully, you can hope for a polynomial increase of the mixing time and $\sigma^2_{\text{lim}}(h)$ with d.
- Most MCMC algorithms are instances of Metropolis-Hastings with clever choices of proposal[6], even the NUTS HMC of Stan and PyMC3.
- For nice illustrations, check out `https://chi-feng.github.io/mcmc-demo/`

[5] R. Douc et al. *Nonlinear time series*. Chapman-Hall, 2014.

[6] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. New York: Springer-Verlag, 2004.

Variational approximations

- When in a hurry, you can settle for a good approximation to your posterior

$$\pi(\theta) = p(\theta|\mathbf{x}) \propto p(\mathbf{x}|\theta)p(\theta),$$

say minimize in q

$$\begin{aligned}\mathrm{KL}(q, \pi) &= \mathbb{E}_q \log q - \mathbb{E}_q \log p(\theta|\mathbf{x}) \\ &= -[-\mathbb{E}_q \log q + \mathbb{E}_q \log p(\mathbf{x}, \theta)] + \log p(\mathbf{x}).\end{aligned}$$

- Equivalently, we can maximize the evidence lower bound (ELBO)[7].
- Ideally, I would rather cast the choice of q into a Wald-like problem.

[7] D. M. Blei et al. "Variational inference: A review for statisticians". In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877.

Outline

Linear regression

- $y_i = f(x_i) + \varepsilon_i$, $f(x) = \theta^T x$, ε_i i.i.d. Gaussians $\mathcal{N}(0, \sigma^2)$.
- If we choose $p(\theta) \sim \mathcal{N}(0, \Sigma^2)$, then

$$\begin{aligned} p(\theta|(x,y)_{1:n}) &\propto p((x,y)_{1:n}|\theta)p(\theta) \\ &= \text{[Exercise]} \end{aligned}$$

where $A = \sigma^{-2} X^T X + \Sigma^{-1}$.

Linear regression

- $y_i = f(x_i) + \varepsilon_i$, $f(x) = \theta^T x$, ε_i i.i.d. Gaussians $\mathcal{N}(0, \sigma^2)$.
- If we choose $p(\theta) \sim \mathcal{N}(0, \Sigma^2)$, then

$$\begin{aligned} p(\theta|(x,y)_{1:n}) &\propto p((x,y)_{1:n}|\theta)p(\theta) \\ &\propto \exp\left[-\frac{\|\mathbf{y} - X\theta\|^2}{2\sigma^2} - \frac{\theta^T \Sigma^{-1} \theta}{2}\right] \end{aligned}$$

where $A = \sigma^{-2} X^T X + \Sigma^{-1}$.

Linear regression

- $y_i = f(x_i) + \varepsilon_i$, $f(x) = \theta^T x$, ε_i i.i.d. Gaussians $\mathcal{N}(0, \sigma^2)$.
- If we choose $p(\theta) \sim \mathcal{N}(0, \Sigma^2)$, then

$$\begin{aligned} p(\theta|(x,y)_{1:n}) &\propto p((x,y)_{1:n}|\theta)p(\theta) \\ &\propto \exp\left[-\frac{\|\mathbf{y} - X\theta\|^2}{2\sigma^2} - \frac{\theta^T\Sigma^{-1}\theta}{2}\right] \\ &\propto \exp\left[\frac{\mathbf{y}^T X\theta}{\sigma^2} - \frac{\theta^T X^T X\theta}{2\sigma^2} - \frac{\theta^T\Sigma^{-1}\theta}{2}\right] \end{aligned}$$

where $A = \sigma^{-2}X^T X + \Sigma^{-1}$.

Linear regression

- $y_i = f(x_i) + \varepsilon_i$, $f(x) = \theta^T x$, ε_i i.i.d. Gaussians $\mathcal{N}(0, \sigma^2)$.
- If we choose $p(\theta) \sim \mathcal{N}(0, \Sigma^2)$, then

$$
\begin{aligned}
p(\theta|(x,y)_{1:n}) &\propto p((x,y)_{1:n}|\theta)p(\theta) \\
&\propto \exp\left[-\frac{\|\mathbf{y} - X\theta\|^2}{2\sigma^2} - \frac{\theta^T\Sigma^{-1}\theta}{2}\right] \\
&\propto \exp\left[\frac{\mathbf{y}^T X\theta}{\sigma^2} - \frac{\theta^T X^T X\theta}{2\sigma^2} - \frac{\theta^T\Sigma^{-1}\theta}{2}\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\theta - \sigma^{-2}A^{-1}X\mathbf{y}\right)^T A\left(\theta - \sigma^{-2}A^{-1}X\mathbf{y}\right)\right]
\end{aligned}
$$

where $A = \sigma^{-2}X^T X + \Sigma^{-1}$.

Linear regression

- $y_i = f(x_i) + \varepsilon_i$, $f(x) = \theta^T x$, ε_i i.i.d. Gaussians $\mathcal{N}(0, \sigma^2)$.
- If we choose $p(\theta) \sim \mathcal{N}(0, \Sigma^2)$, then

$$
\begin{aligned}
p(\theta|(x,y)_{1:n}) &\propto p((x,y)_{1:n}|\theta)p(\theta) \\
&\propto \exp\left[-\frac{\|\mathbf{y} - X\theta\|^2}{2\sigma^2} - \frac{\theta^T\Sigma^{-1}\theta}{2}\right] \\
&\propto \exp\left[\frac{\mathbf{y}^T X\theta}{\sigma^2} - \frac{\theta^T X^T X\theta}{2\sigma^2} - \frac{\theta^T\Sigma^{-1}\theta}{2}\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\theta - \sigma^{-2}A^{-1}X\mathbf{y}\right)^T A\left(\theta - \sigma^{-2}A^{-1}X\mathbf{y}\right)\right] \\
&= \mathcal{N}(\sigma^{-2}A^{-1}X\mathbf{y}, A^{-1}),
\end{aligned}
$$

where $A = \sigma^{-2}X^T X + \Sigma^{-1}$.

Linear regression

- $y_i = f(x_i) + \varepsilon_i$, $f(x) = \theta^T x$, ε_i i.i.d. Gaussians $\mathcal{N}(0, \sigma^2)$.
- If we choose $p(\theta) \sim \mathcal{N}(0, \Sigma^2)$, then

$$
\begin{aligned}
p(\theta|(x,y)_{1:n}) &\propto p((x,y)_{1:n}|\theta)p(\theta) \\
&\propto \exp\left[-\frac{\|\mathbf{y} - X\theta\|^2}{2\sigma^2} - \frac{\theta^T\Sigma^{-1}\theta}{2}\right] \\
&\propto \exp\left[\frac{\mathbf{y}^T X\theta}{\sigma^2} - \frac{\theta^T X^T X\theta}{2\sigma^2} - \frac{\theta^T\Sigma^{-1}\theta}{2}\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\theta - \sigma^{-2}A^{-1}X\mathbf{y}\right)^T A\left(\theta - \sigma^{-2}A^{-1}X\mathbf{y}\right)\right] \\
&= \mathcal{N}(\sigma^{-2}A^{-1}X\mathbf{y}, A^{-1}),
\end{aligned}
$$

  where $A = \sigma^{-2}X^T X + \Sigma^{-1}$.

- Remember prediction with L^2 loss is simple:

$$
\arg\min_{y_\star} \int \|y_\star - f(x_\star)\|^2 p(\theta|(x,y)_{1:n}) d\theta = \sigma^{-2} x_\star^T A^{-1} X\mathbf{y}.
$$

Linear regression

- $y_i = f(x_i) + \varepsilon_i$, $f(x) = \theta^T x$, ε_i i.i.d. Gaussians $\mathcal{N}(0, \sigma^2)$.
- If we choose $p(\theta) \sim \mathcal{N}(0, \Sigma^2)$, then

$$
\begin{aligned}
p(\theta|(x,y)_{1:n}) &\propto p((x,y)_{1:n}|\theta)p(\theta) \\
&\propto \exp\left[-\frac{\|\mathbf{y} - X\theta\|^2}{2\sigma^2} - \frac{\theta^T\Sigma^{-1}\theta}{2}\right] \\
&\propto \exp\left[\frac{\mathbf{y}^T X\theta}{\sigma^2} - \frac{\theta^T X^T X\theta}{2\sigma^2} - \frac{\theta^T\Sigma^{-1}\theta}{2}\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\theta - \sigma^{-2}A^{-1}X\mathbf{y}\right)^T A\left(\theta - \sigma^{-2}A^{-1}X\mathbf{y}\right)\right] \\
&= \mathcal{N}(\sigma^{-2}A^{-1}X\mathbf{y}, A^{-1}),
\end{aligned}
$$

where $A = \sigma^{-2}X^T X + \Sigma^{-1}$.

- Actually, we can even check that

$$p(f(x_\star)|x_\star, (x,y)_{1:n}) = \mathcal{N}(\sigma^{-2}x_\star^T A^{-1}X\mathbf{y}, x^{\star T}A^{-1}x^\star).$$

Linear regression with nonlinear features

- Replace each x by a vector of features $\varphi(x) \in \mathbb{R}^p$:

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

$f(x) = \theta^T \varphi(x)$, ε_i i.i.d. Gaussians $\mathcal{N}(0, \sigma^2)$, $\theta \sim \mathcal{N}(0, \Sigma)$.

- Think $\varphi(x) = (1, x^1, x^2, x^1 x^2, \dots)$
- Recall

$$p(f(x_\star)|x_\star, (x, y)_{1:n}) = \mathcal{N}(\sigma^{-2}\Phi_\star^T A^{-1}\mathbf{\Phi}\mathbf{y}, \Phi^\star A^{-1}\Phi^\star)$$

where $A = \sigma^{-2}\mathbf{\Phi}^T\mathbf{\Phi} + \Sigma^{-1}$.

- Requires $p \times p$ inversion.
- But let $\mathbf{K} = \mathbf{\Phi}\Sigma\mathbf{\Phi}^T$, then can rewrite (Exercise)

$$p(f(x_\star)|(x, y)_{1:n}) = \mathcal{N}(\mu_\star, \sigma_\star^2),$$

where

$$\mu_\star = \Phi_\star^T \Sigma \mathbf{\Phi}^T (\mathbf{K} + \sigma^2 I)^{-1}\mathbf{y},$$
$$\sigma_\star^2 = \varphi_\star \Sigma \varphi_\star - \varphi_\star^T \Sigma \mathbf{\Phi}^T (\mathbf{K} + \sigma^2 I)^{-1}\mathbf{\Phi}\Sigma\varphi_\star.$$

Gaussian processes

- A distribution over a space of functions $f : \mathbb{R}^d \to \mathbb{R}$.

Gaussian processes

If $\forall p \in \mathbb{N}, \forall x_1, \ldots, x_p \in \mathbb{R}^d$

$$[f(x_1), \ldots, f(x_p)]^T \sim \mathcal{N}(\mathbf{m}, \mathbf{K}),$$

where $\mathbf{m} = [\mu(x_1), \ldots, \mu(x_p)]$ and

$$\mathbf{K} = ((K(x_i, x_j))),$$

then we say $f \sim GP(\mu, K)$.

- Unicity is usually easy, existence is tricky.
- Necessary condition is that all matrices $\mathbf{K}$ are positive definite.

Sampling, conditioning and predicting

See notebook 01 on
`https://github.com/rbardenet/bnp-course`

Commonly-used kernels

covariance function	expression	S	ND
constant	σ_0^2	$\surd$	
linear	$\sum_{d=1}^{D} \sigma_d^2 x_d x'_d$		
polynomial	$(\mathbf{x} \cdot \mathbf{x}' + \sigma_0^2)^p$		
squared exponential	$\exp(-\frac{r^2}{2\ell^2})$	$\surd$	$\surd$
Matérn	$\frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{\ell} r\right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}}{\ell} r\right)$	$\surd$	$\surd$
exponential	$\exp(-\frac{r}{\ell})$	$\surd$	$\surd$
γ-exponential	$\exp\left(-(\frac{r}{\ell})^{\gamma}\right)$	$\surd$	$\surd$
rational quadratic	$(1 + \frac{r^2}{2\alpha\ell^2})^{-\alpha}$	$\surd$	$\surd$
neural network	$\sin^{-1}\left(\frac{2\tilde{\mathbf{x}}^\top \Sigma \tilde{\mathbf{x}}'}{\sqrt{(1+2\tilde{\mathbf{x}}^\top \Sigma \tilde{\mathbf{x}})(1+2\tilde{\mathbf{x}}'^\top \Sigma \tilde{\mathbf{x}}')}}\right)$		$\surd$

Table 4.1: Summary of several commonly-used covariance functions. The covariances are written either as a function of $\mathbf{x}$ and $\mathbf{x}'$, or as a function of $r = |\mathbf{x} - \mathbf{x}'|$. Two columns marked 'S' and 'ND' indicate whether the covariance functions are stationary and nondegenerate respectively. Degenerate covariance functions have finite rank, see section 4.3 for more discussion of this issue.

Learning

- In regression,

$$p(\mathbf{y}|\mathbf{x}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{x})d\mathbf{f}$$
$$= \mathcal{N}(\mathbf{y}|0, \mathbf{K} + \sigma^2 I_n).$$

- So simply put a prior over $\eta = (\sigma, \theta)$ and integrate.
- Prediction becomes

$$f_\star \sim \int p(f_\star|\mathbf{x}, \eta)p(\eta)d\eta.$$

- Alternately, lots of people maximize the marginal likelihood.

Learning

- In regression,

$$p(\mathbf{y}|\mathbf{x}, \theta) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{x})d\mathbf{f}$$
$$= \mathcal{N}(\mathbf{y}|0, \mathbf{K}_\theta + \sigma^2 I_n).$$

- So simply put a prior over $\eta = (\sigma, \theta)$ and integrate.
- Prediction becomes

$$f_\star \sim \int p(f_\star|\mathbf{x}, \eta)p(\eta)d\eta.$$

- Alternately, lots of people maximize the marginal likelihood.

Beyond regression: classification[8]

- (Exercise) Find a simple classification model with GPs.
- Take for instance

$$p(y = +1|x, f) = \text{Bernoulli}(\sigma(f(x))), \quad f \sim \text{GP}(0, K).$$

- Problem: prediction is not easy anymore

$$p(f_\star|X, \mathbf{y}, x_\star) = \int p(f_\star|X, \mathbf{f}, x_\star) p(\mathbf{f}|X, \mathbf{y}) d\mathbf{f}$$

[8] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Beyond regression: classification[8]

- (Exercise) Find a simple classification model with GPs.
- Take for instance

$$p(y = +1|x, f) = \text{Bernoulli}(\sigma(f(x))), \quad f \sim \text{GP}(0, K).$$

- Problem: prediction is not easy anymore

$$p(f_\star|X, \mathbf{y}, x_\star) = \int p(f_\star|X, \mathbf{f}, x_\star) p(\mathbf{f}|X, \mathbf{y}) d\mathbf{f}$$

[8] Rasmussen and Williams, *Gaussian Processes for Machine Learning*.

Beyond regression: classification[8]

- (Exercise) Find a simple classification model with GPs.
- Take for instance

$$p(y = +1|x, f) = \text{Bernoulli}(\sigma(f(x))), \quad f \sim \text{GP}(0, K).$$

- Problem: prediction is not easy anymore

$$p(f_\star|X, \mathbf{y}, x_\star) = \int p(f_\star|X, \mathbf{f}, x_\star) p(\mathbf{f}|X, \mathbf{y}) d\mathbf{f}$$

[8] Rasmussen and Williams, *Gaussian Processes for Machine Learning*.

Beyond regression: ranking[9]

- (Exercise) Find a simple ranking model with GPs: your data is $(u, v)_{1:n}$ where $\forall i, u_i \prec v_i$. Your user wants to know whether a new $u_\star \prec v_\star$.
- Take for instance

$$p(u \prec v | u, v, f) = \varphi(f(v) - f(u)), \quad f \sim GP(0, K), \varphi \text{ increasing.}$$

- Same difficulties with learning.

[9] W. Chu and Z. Ghahramani. "Preference learning with Gaussian processes". In: *Proceedings of the 22nd International Conference on Machine Learning*. 2005, pp. 137–144.

Beyond regression: ranking[9]

- (Exercise) Find a simple ranking model with GPs: your data is $(u, v)_{1:n}$ where $\forall i, u_i \prec v_i$. Your user wants to know whether a new $u_\star \prec v_\star$.
- Take for instance

$$p(u \prec v | u, v, f) = \varphi(f(v) - f(u)), \quad f \sim GP(0, K), \varphi \text{ increasing.}$$

- Same difficulties with learning.

[9] Chu and Ghahramani, "Preference learning with Gaussian processes".

Beyond regression: ranking[9]

- (Exercise) Find a simple ranking model with GPs: your data is $(u, v)_{1:n}$ where $\forall i, u_i \prec v_i$. Your user wants to know whether a new $u_\star \prec v_\star$.
- Take for instance

$$p(u \prec v | u, v, f) = \varphi(f(v) - f(u)), \quad f \sim GP(0, K), \varphi \text{ increasing.}$$

- Same difficulties with learning.

[9] Chu and Ghahramani, "Preference learning with Gaussian processes".

Emulators of expensive models

PLOS ONE

RESEARCH ARTICLE

Bayesian Sensitivity Analysis of a Cardiac Cell Model Using a Gaussian Process Emulator

Eugene TY Chang[1,2], **Mark Strong**[3], **Richard H Clayton**[1,2]*

1 Insigneo Institute for *in-silico* Medicine, University of Sheffield, Sheffield, United Kingdom, **2** Department of Computer Science University of Sheffield, Sheffield, United Kingdom, **3** School of Health and Related Research, University of Sheffield, Sheffield, United Kingdom

* r.h.clayton@sheffield.ac.uk

CrossMark
click for updates

OPEN ACCESS

Citation: Chang ETY, Strong M, Clayton RH (2015)

Abstract

Models of electrical activity in cardiac cells have become important research tools as they can provide a quantitative description of detailed and integrative physiology. However, cardiac cell models have many parameters, and how uncertainties in these parameters affect the model output is difficult to assess without undertaking large numbers of model runs. In this study we show that a surrogate statistical model of a cardiac cell model (the Luo-Rudy 1991 model) can be built using Gaussian process (GP) emulators. Using this approach we

Nonparametric fits

Gaussian Process Cosmography

Arman Shafieloo[1], Alex G. Kim[2], Eric V. Linder[1,2,3]
[1] *Institute for the Early Universe WCU, Ewha Womans University, Seoul, Korea*
[2] *Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA and*
[3] *University of California, Berkeley, CA 94720, USA*
(Dated: July 11, 2012)

Gaussian processes provide a method for extracting cosmological information from observations without assuming a cosmological model. We carry out cosmography – mapping the time evolution of the cosmic expansion – in a model-independent manner using kinematic variables and a geometric probe of cosmology. Using the state of the art supernova distance data from the Union2.1 compilation, we constrain, without any assumptions about dark energy parametrization or matter density, the Hubble parameter and deceleration parameter as a function of redshift. Extraction of these relations is tested successfully against models with features on various coherence scales, subject to certain statistical cautions.

I. INTRODUCTION

Cosmic acceleration is a fundamental mystery of great interest and importance to understanding cosmology, gravitation, and high energy physics. The cosmic expansion rate is slowed down by gravitationally attractive matter and sped up by some other, unknown contribution to the dynamical equations. While great effort is being put into identifying the source of this extra dark energy contribution, the overall expansion behavior also holds important clues to origin, evolution, and present

ing procedures have been suggested, e.g. [6], but tend to induce bias in the function reconstruction due to parametric restriction of the behavior or to have poor error control. Using a general orthonormal basis or principal component analysis is another approach, to describe the distance-redshift relation (e.g. [7]) or the deceleration parameter [8], or using a correlated prior for smoothness on the dark energy equation of state [9], but in practice a finite (and small) number of modes is significant beyond the prior, essentially reducing to a parametric approach. Gaussian processes [10] offer an interesting possibility for

Natural language processing

Using Gaussian Processes for Rumour Stance Classification in Social Media

MICHAL LUKASIK, University of Sheffield
KALINA BONTCHEVA, University of Sheffield
TREVOR COHN, University of Melbourne
ARKAITZ ZUBIAGA, University of Warwick
MARIA LIAKATA, University of Warwick
ROB PROCTER, University of Warwick

Social media tend to be rife with rumours while new reports are released piecemeal during breaking news. Interestingly, one can mine multiple reactions expressed by social media users in those situations, exploring their stance towards rumours, ultimately enabling the flagging of highly disputed rumours as being potentially false. In this work, we set out to develop an automated, supervised classifier that uses multi-task learning to classify the stance expressed in each individual tweet in a rumourous conversation as either supporting, denying or questioning the rumour. Using a classifier based on Gaussian Processes, and exploring its effectiveness on two datasets with very different characteristics and varying distributions of stances, we show that our approach consistently outperforms competitive baseline classifiers. Our classifier is especially effective in estimating the distribution of different types of stance associated with a given rumour, which we set forth as a desired characteristic for a rumour-tracking system that will warn both ordinary users of Twitter and professional news practitioners when a rumour is being rebutted.

1. INTRODUCTION

There is an increasing need to interpret and act upon rumours spreading quickly through social media during breaking news, where new reports are released piecemeal and often have an unverified

Bayesian optimization for hyperparameter tuning

Algorithms for Hyper-Parameter Optimization

James Bergstra
The Rowland Institute
Harvard University
bergstra@rowland.harvard.edu

Rémi Bardenet
Laboratoire de Recherche en Informatique
Université Paris-Sud
bardenet@lri.fr

Yoshua Bengio
Dépt. d'Informatique et Recherche Opérationelle
Université de Montréal
yoshua.bengio@umontreal.ca

Balázs Kégl
Linear Accelerator Laboratory
Université Paris-Sud, CNRS
balazs.kegl@gmail.com

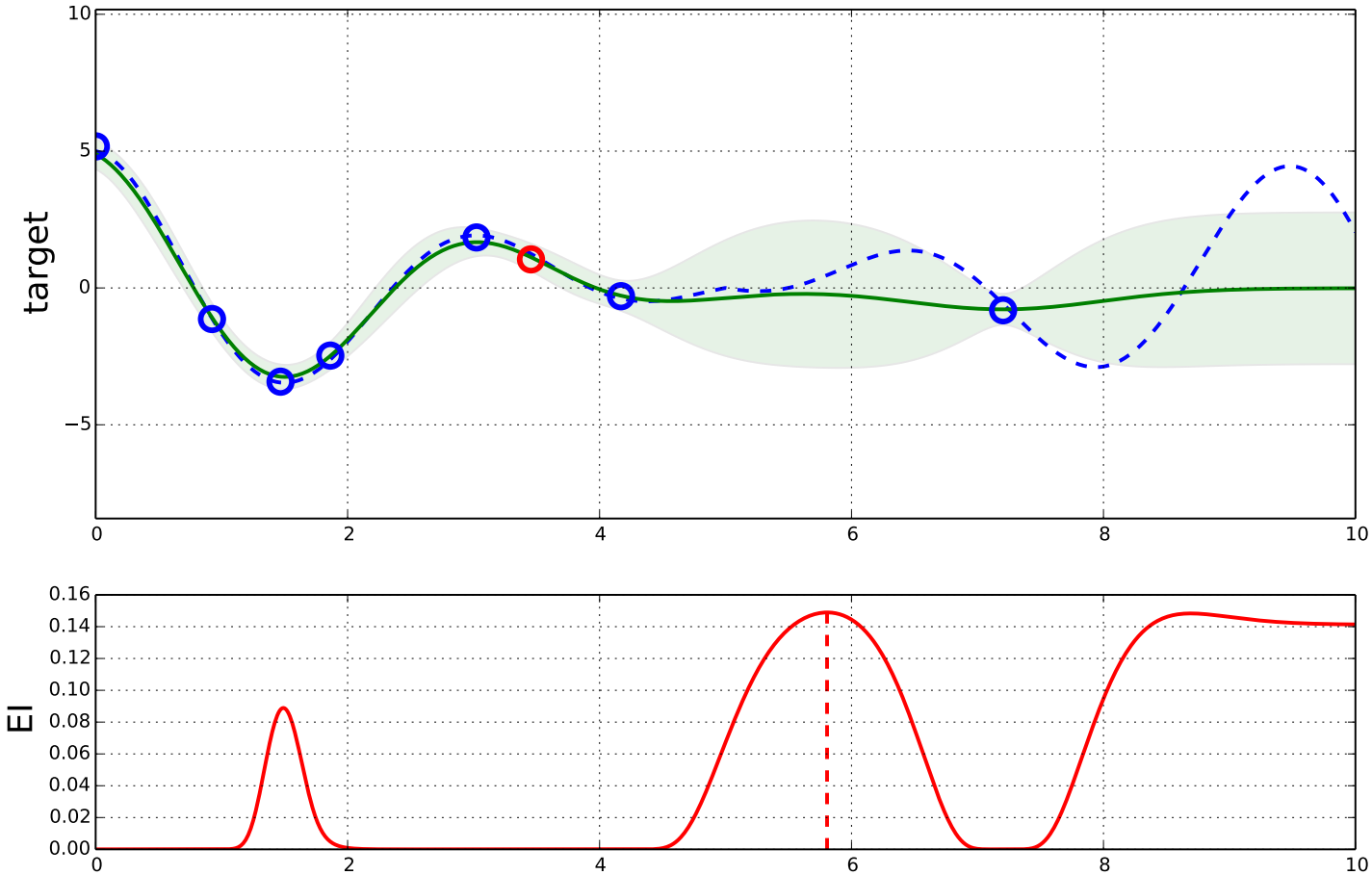
Abstract

Several recent advances to the state of the art in image classification benchmarks have come from better configurations of existing techniques rather than novel approaches to feature learning. Traditionally, hyper-parameter optimization has been the job of humans because they can be very efficient in regimes where only a few

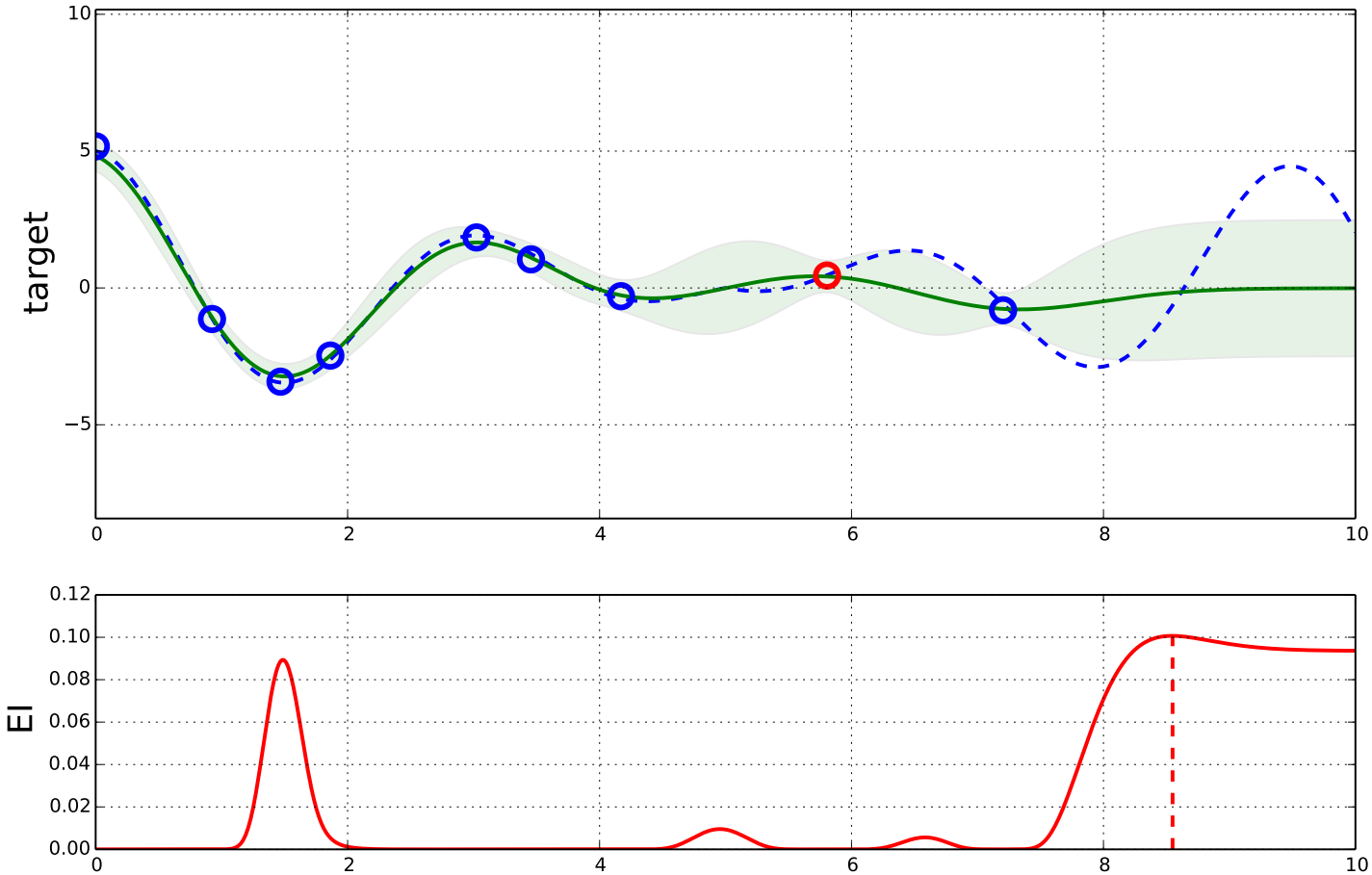
Bayesian optimization

- Goal is to minimize a noisy f with N iterations, N small.
- Key application: find the hyperparameters of your ML algorithm that minimize the validation error.
- Idea is to sequentially
  - update your model on f,
  - optimize an aquisition criterion.
- Check out notebook 03 on `https://github.com/rbardenet/bnp-course`.

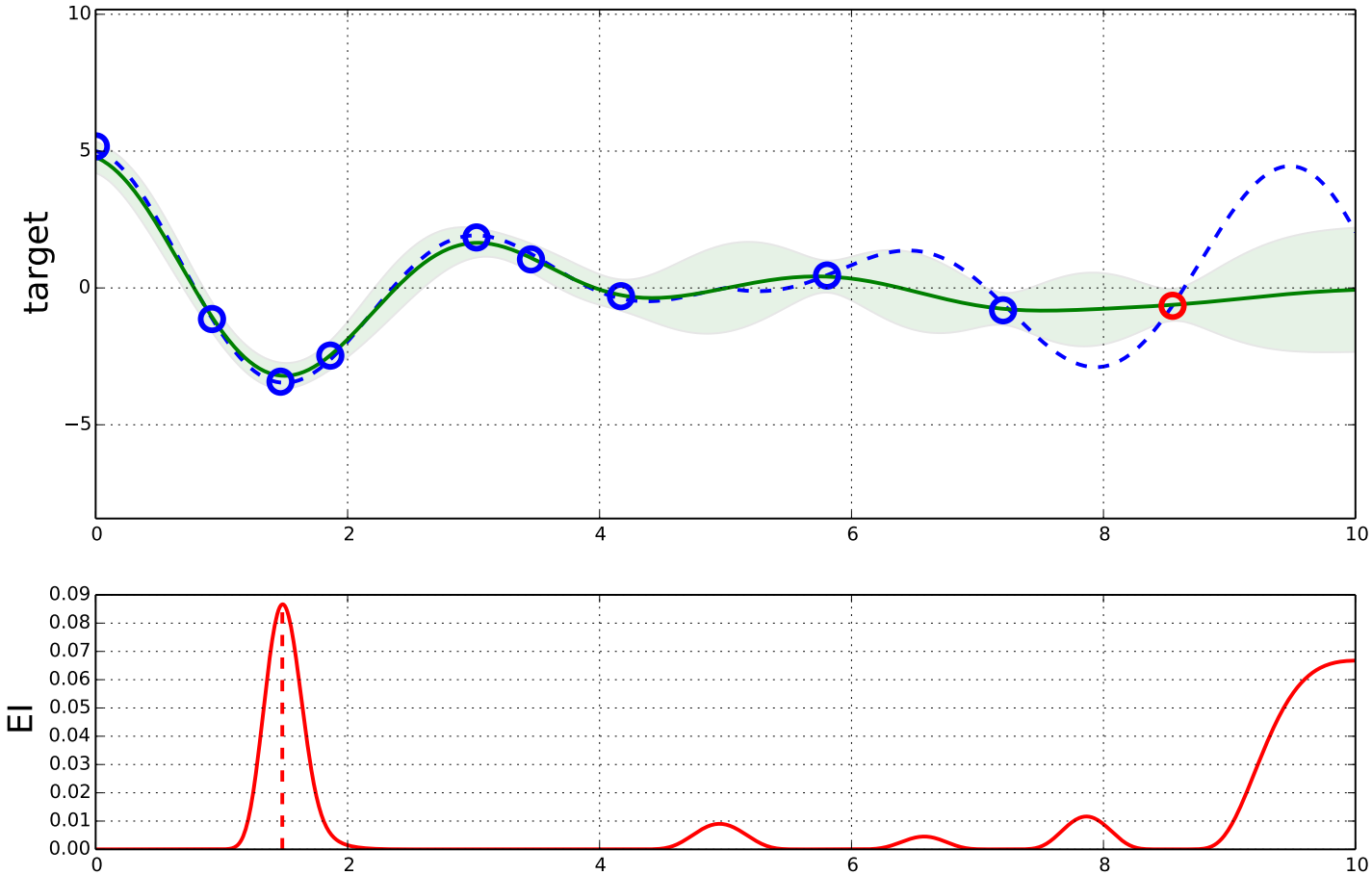
An example in 1D



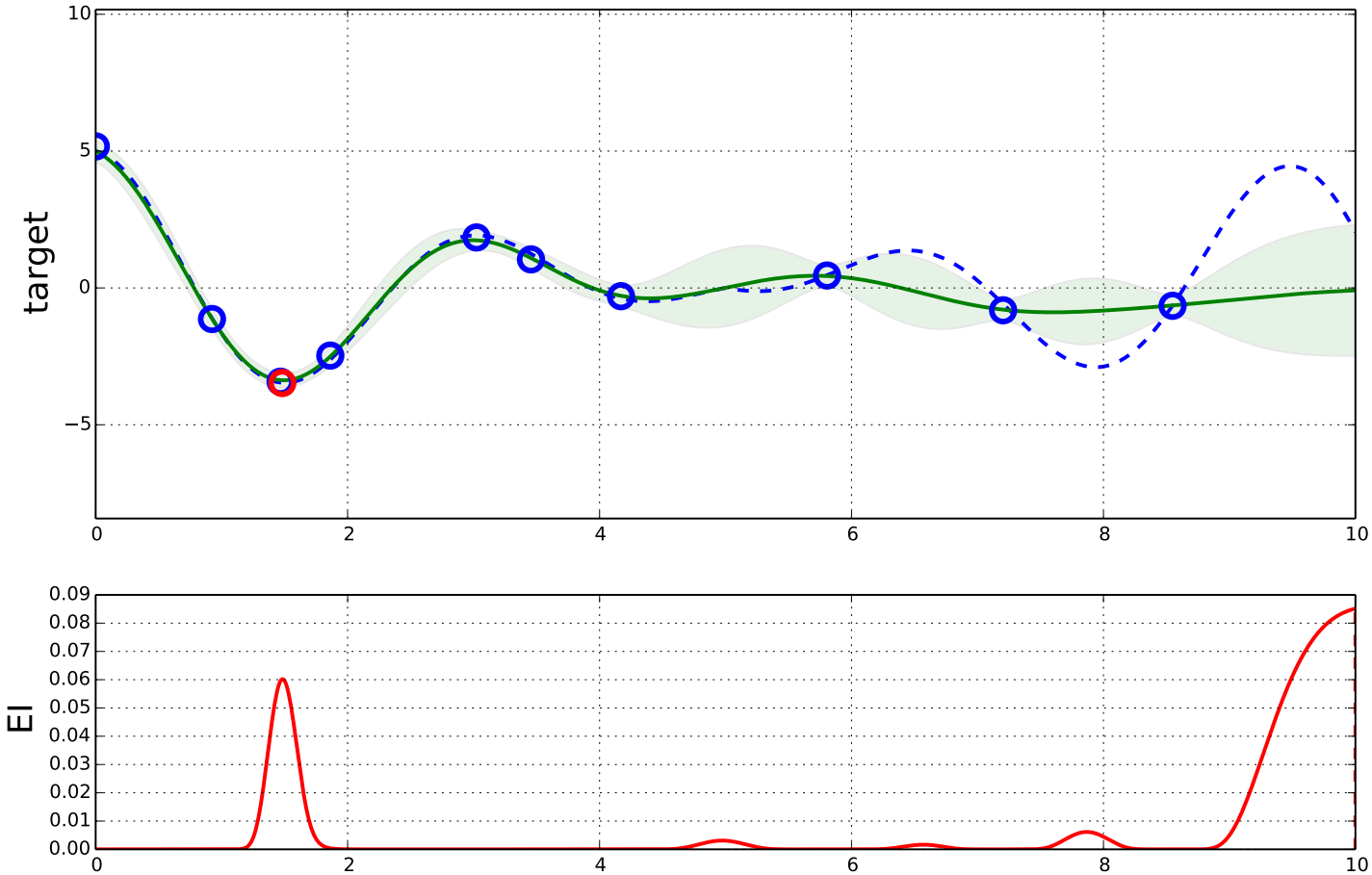
An example in 1D



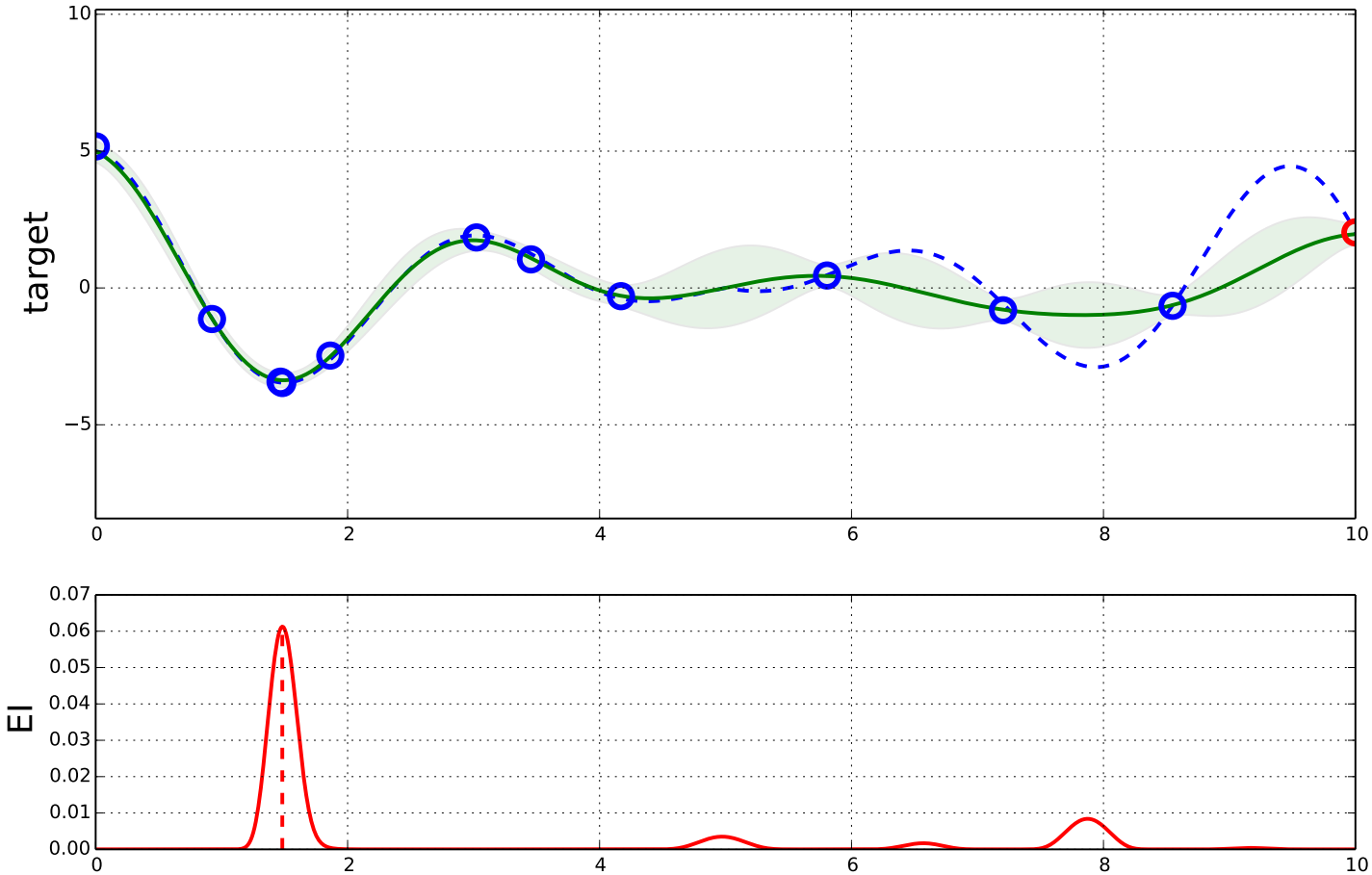
An example in 1D



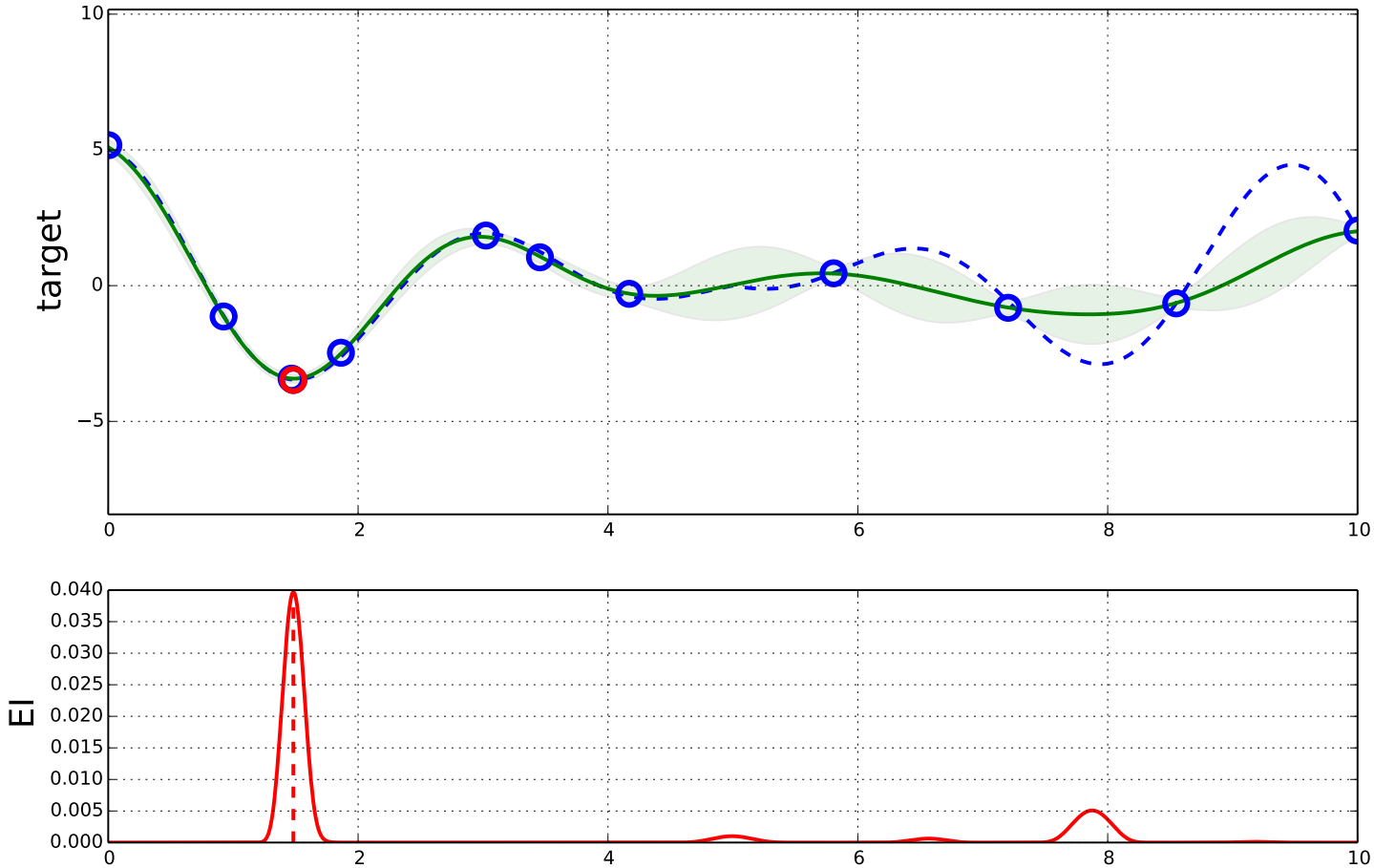
An example in 1D



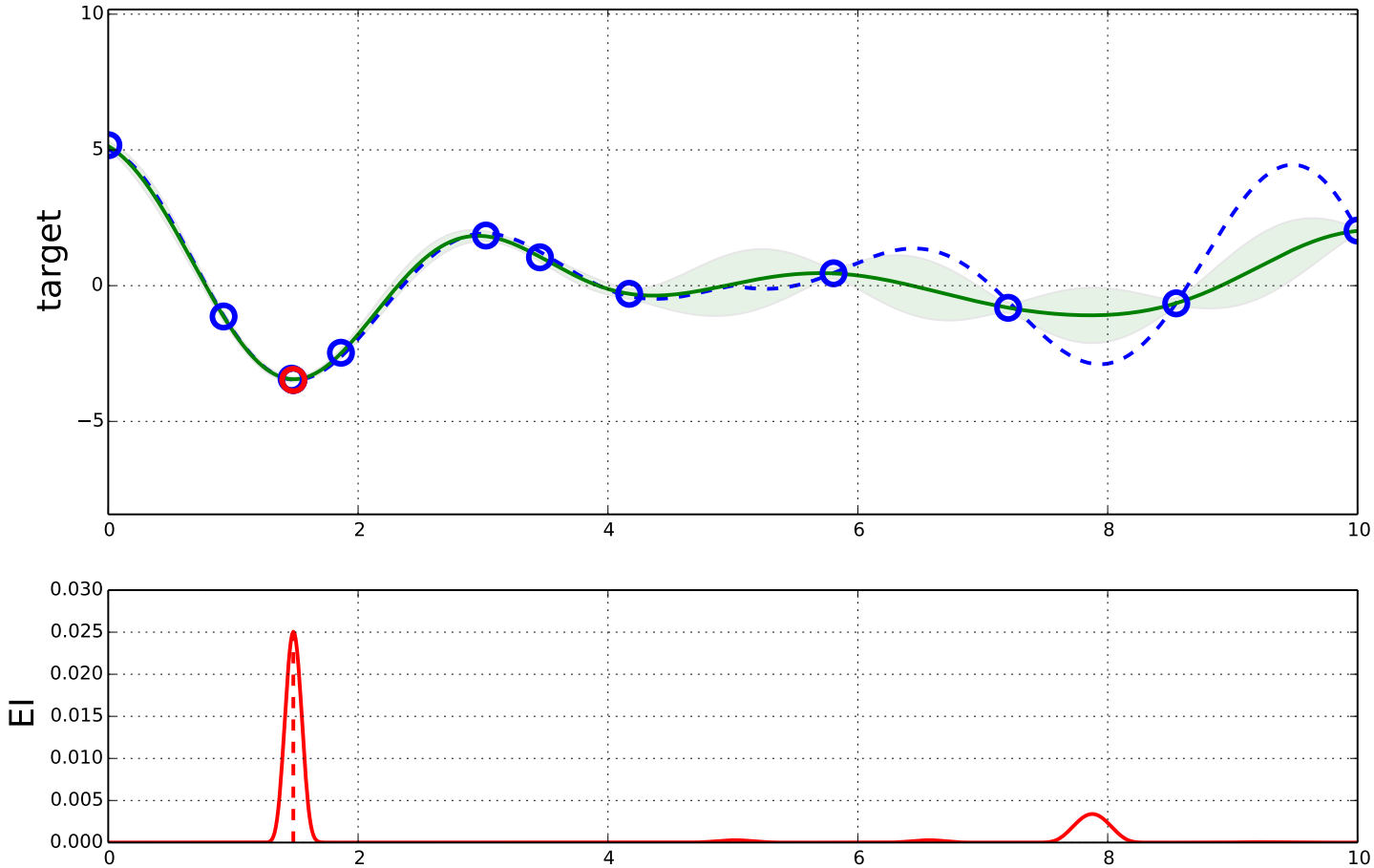
An example in 1D



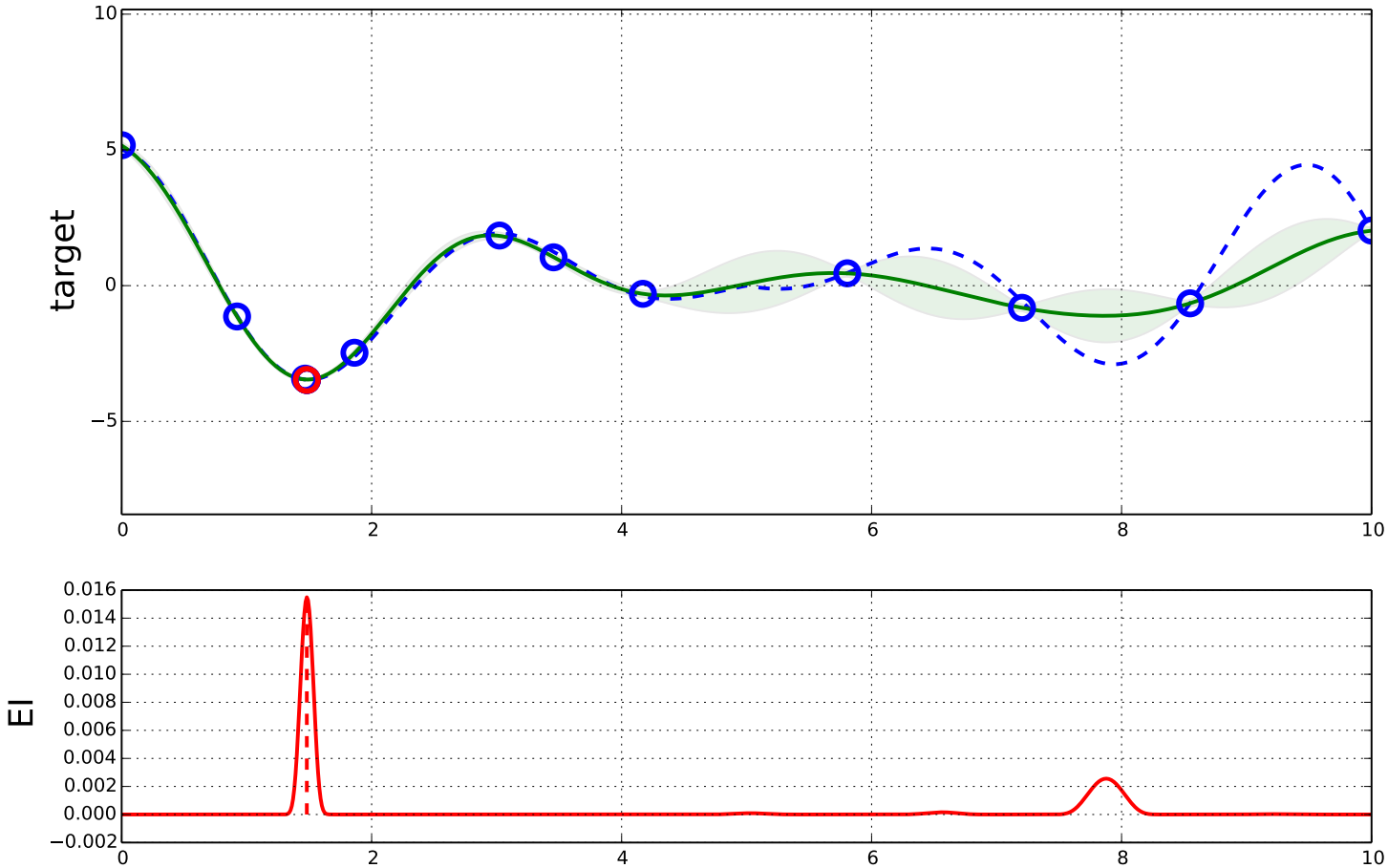
An example in 1D



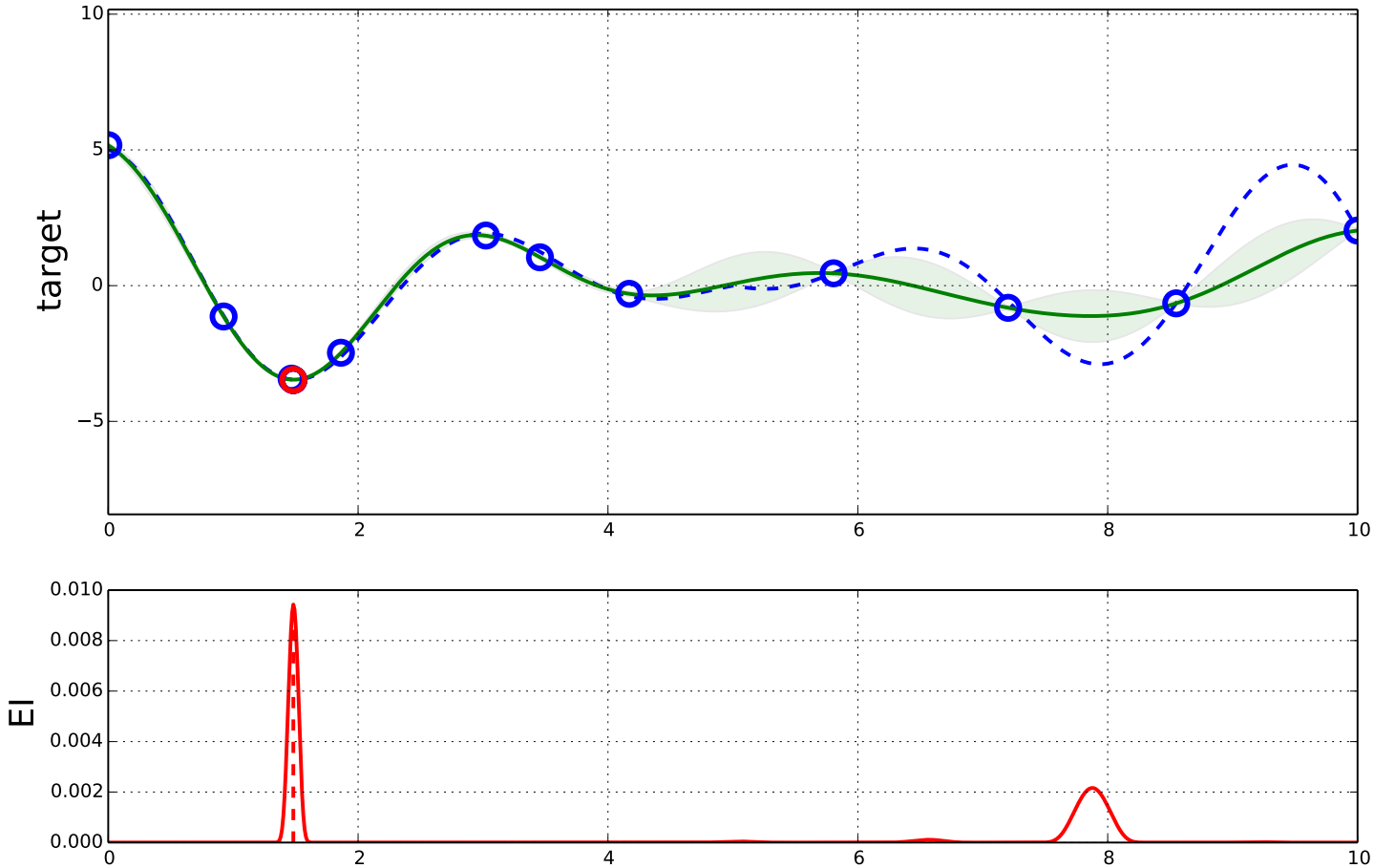
An example in 1D



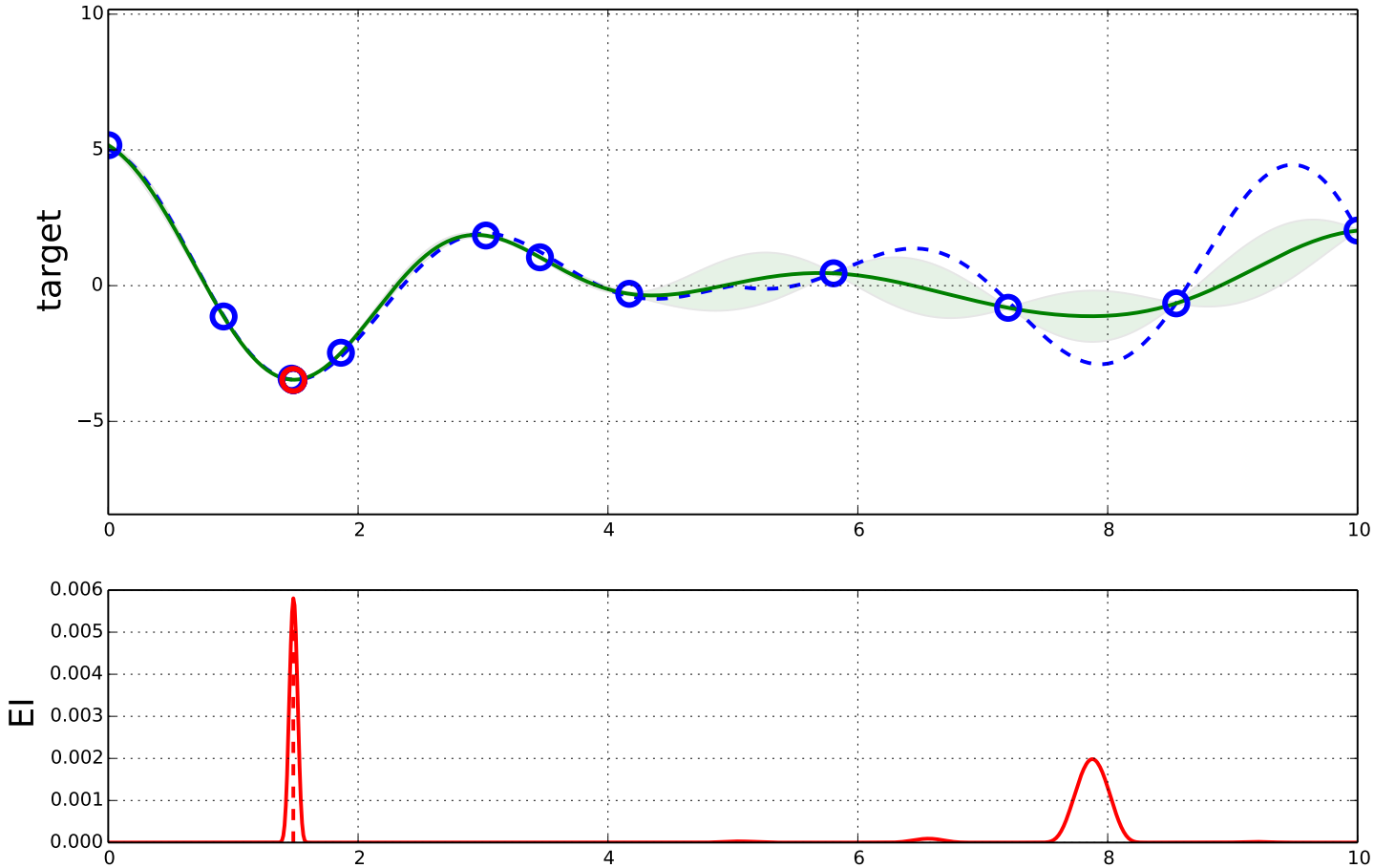
An example in 1D



An example in 1D



An example in 1D



Popular aquisition criteria

Expected improvement[10]

$$\mathrm{EI}(z) = \mathbb{E}\big(\max(m_N - f(z), 0) | (x, y)_{1:n} \big),$$

where $m_N = \min_{1 \leqslant i \leqslant N} f(x_i)$.

An easy computation yields

$$\mathrm{EI}(x) = \widetilde{\sigma}(x)\big(u\Phi(u) + \varphi(u)\big), \tag{3}$$

where

$$u = (m_n - \widetilde{m}(x))/\widetilde{\sigma}(x),$$

and Φ and φ denote the cdf and pdf of the $\mathcal{N}(0, 1)$ distribution.

[10] D. R. Jones. "A Taxonomy of Global Optimization Methods Based on Response Surfaces". In: *Journal of Global Optimization* 21 (2001), pp. 345–383.

GP-UCB (Srinivas et al., 2010)

GP-UCB

$$\text{GP-UCB}(z) = \widetilde{m}(z) + \beta\widetilde{\sigma}(z).$$

- If β properly tuned, can use bandit results.
- First criterion to give application theoretical results.

Bayesian optimization for hyperparameter tuning

Algorithms for Hyper-Parameter Optimization

James Bergstra
The Rowland Institute
Harvard University
bergstra@rowland.harvard.edu

Rémi Bardenet
Laboratoire de Recherche en Informatique
Université Paris-Sud
bardenet@lri.fr

Yoshua Bengio
Dépt. d'Informatique et Recherche Opérationelle
Université de Montréal
yoshua.bengio@umontreal.ca

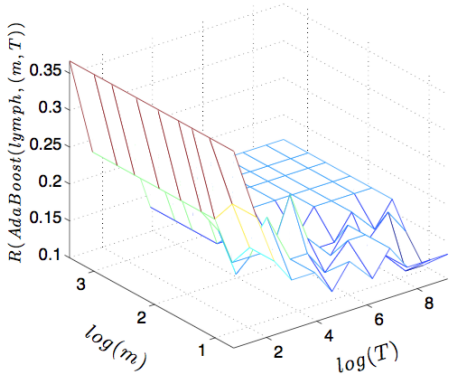
Balázs Kégl
Linear Accelerator Laboratory
Université Paris-Sud, CNRS
balazs.kegl@gmail.com

Abstract

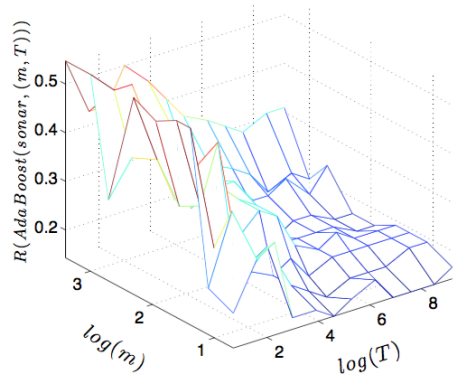
Several recent advances to the state of the art in image classification benchmarks have come from better configurations of existing techniques rather than novel approaches to feature learning. Traditionally, hyper-parameter optimization has been the job of humans because they can be very efficient in regimes where only a few

- Checkout hyperopt and spearmint.

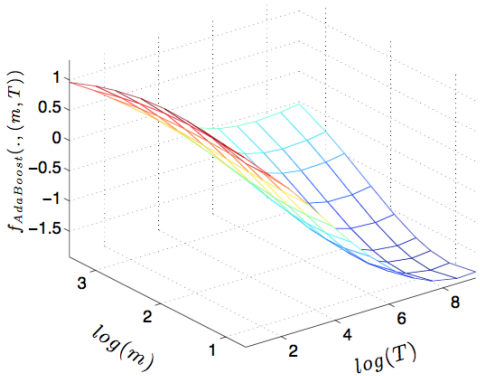
Going further: Hyperopt across datasets[10]



(a) Error surface of AdaBoost on lymph



(b) Error surface of AdaBoost on sonar



(c) The common latent ranker

[10] R. Bardenet et al. "Collaborative hyperparameter tuning". In: *International Conference on Machine Learning (ICML)*. Atlanta, Georgia, 2013.

Some useful hyperlinks

- Textbook by C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006,
  - great for understanding, methods, pointers to ML and stats.
- Videolecture by C. Rasmussen.
- lecture notes by P. Orbanz. *Lecture notes on Bayesian nonparametrics*. 2014.
  - mathematically clean, without losing the focus on ML.

Some open issues

- Fully Bayesian scalable approaches!
- Natural approaches to constrained GPs.
- Links with other models based on Gaussians and geometry.

Back to the roots

- Formulate HT across datasets and algorithms as a posterior expected loss problem, including computational constraints.
- Solve resulting dynamic programming problem.

References I

Bardenet, R. et al. "Collaborative hyperparameter tuning". In: *International Conference on Machine Learning (ICML)*. Atlanta, Georgia, 2013.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe. "Variational inference: A review for statisticians". In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877.

Chu, W. and Z. Ghahramani. "Preference learning with Gaussian processes". In: *Proceedings of the 22nd International Conference on Machine Learning*. 2005, pp. 137–144.

Douc, R., É. Moulines, and D. Stoffer. *Nonlinear time series*. Chapman-Hall, 2014.

Jones, D. R. "A Taxonomy of Global Optimization Methods Based on Response Surfaces". In: *Journal of Global Optimization* 21 (2001), pp. 345–383.

Orbanz, P. *Lecture notes on Bayesian nonparametrics*. 2014.

References II

Parmigiani, G. and L. Inoue. *Decision theory: principles and approaches*. Vol. 812. John Wiley & Sons, 2009.

Rasmussen, C. E. and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Robert, C. P. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

Robert, C. P. and G. Casella. *Monte Carlo Statistical Methods*. New York: Springer-Verlag, 2004.

Stigler, S. M. "The epic story of maximum likelihood". In: *Statistical Science* (2007), pp. 598–620.

Wald, A. *Statistical decision functions*. Wiley, 1950.