STAT G8325
Gaussian Processes and Kernel Methods
Lecture Notes §01: Basics

John P. Cunningham

Department of Statistics
Columbia University

Outline

Outline

Stochastic processes: reminder and notational conventions

Definition (stochastic process). A stochastic process is an S-valued function $f : T \times \Omega \to S$ where $(\Omega, \mathcal{B}, P)$ is a probability space, and for each $t \in T$, $f(t, \cdot)$ is measurable on Ω. More informally, we obscure the sample space and the σ-algebra and say $f : T \to S$ is a stochastic process indexed over T such that each $f(t)$ is an S-valued random variable.

Notes:

- X is often used; here we use f to be consistent with the gp literature.
- $t \in T$ is often used (time); we also use $x \in \mathcal{X}$, as in $f(x) \sim P$.
- We will focus on the "machine learning" of gp: how to *use* gp, but not its theory (in the measure theory sense), nor its applications (in the non-stats, non-CS sense). **Takeaway: the informal definition is just fine for us.**
- Note: our department has numerous great people in the theory of stochastic processes (e.g., Jingchen, Richard, Peter, the visitor Gennady Samorodnitsky,...).

Gaussian Process

Definition (gaussian process). Let:

i. T *be any set,*

ii. $m : T \to \mathbb{R}$ *be any function, and*

iii. $k : T \times T \to \mathbb{R}$ *be any function that is symmetric (i.e., $k(t, t') = k(t', t) \forall t, t' \in T$) and positive definite (i.e., for any finite $G \subset T$, the matrix $K_G = \{k(t, t')\}_{t,t' \in G}$ is positive definite).*

Then there exists a gaussian process $f = \{f_t\}_{t \in T}$ with mean function m and covariance k. We write $f \sim \mathcal{GP}(m, k)$. Notably, $f \sim \mathcal{GP}(m, k)$ if and only if, for every finite $G \subset T$, the collection of random variables $f_G \sim \mathcal{N}(m_G, K_G)$.

- Existence and uniqueness are both nontrivial. We briefly mention existence.
- The induced finite marginals are $\mathcal{N}(m_G, K_G)$, with:

$$m_G = \begin{bmatrix} m(t_1) \\ \vdots \\ m(t_{|G|}) \end{bmatrix} \in \mathbb{R}^{|G|}, \quad K_G = \{k(t, t')\}_{t,t' \in G}, \quad \forall G \subset T.$$

Existence

Theorem (Kolmogorov's Extension Theorem): For any set T and universally measurable spaces $\{S_t, \mathcal{B}_t\}_{t \in T}$, and any consistent family of probability laws $\{P_G : G \text{ finite}, G \subset T\}$ (with P_G on S_G), there is a probability measure P_T on S_T with $P_G(\cdot) = P_T\left(h_{TG}^{-1}(\cdot)\right)$ for all finite $G \subset T$. [Dud02, ch. 12]

- An infinite collection of finite random variables defines and is defined by a single infinite dimensional distribution.
- h_{TG} is the projection of T onto G, so this just says that any P_G is a projection of the measure on the process P_T.
- In our context, the space of interest is $\{\mathbb{R}, \mathcal{B}_t\}$ (and $\mathcal{B}$ is the σ-algebra of Borel sets in $\mathbb{R}$), which is universally measurable.
- (we are typically only interest in distributions on real functions, and very often only L_2, as we will see particularly when we get to Hilbert spaces).
- A great deal is hidden in "universally measurable"; we'll skip all that.
- Read [Bog98] to learn about Gaussian measures on all sorts of funky spaces.
- Also something important is hidden in the "consistent family"...

Consistency

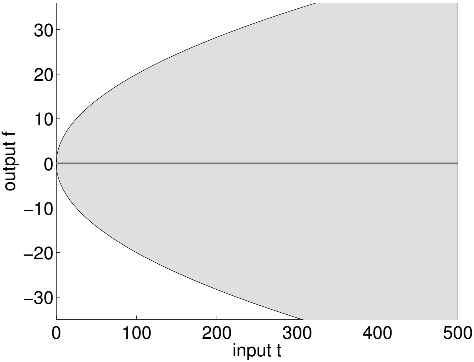
- Consistency essentially means that the marginals of any distribution correspond to the distribution of that lower dimensional object.
- Let $G \subset T$ be a finite subset as before, with $P_G(\cdot) = P_T\left(h_{TG}^{-1}(\cdot)\right)$.
- Then let $G^{-i} \subset G$ be the "one variable integrated out" marginal in the sense of $P_{marg} = \int P_G dP^i$.
- Consistency means $P_{marg} = P_{G^{-i}} = P_T\left(h_{TG^{-i}}^{-1}(\cdot)\right)$.
- Example: if $T = \mathbb{R}$, $G \in \mathbb{R}^d$ with $g_i = g(t_i)$ (so $t \in \mathbb{R}^d$), consistency means:

$$\lim_{t_i \to \infty} P_G(t) = P_{G^{-i}}(t^{-i}).$$

- GP achieve consistency almost trivially by the marginalization property of the normal distribution.

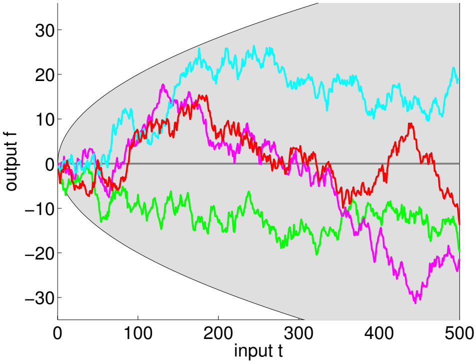
Brownian motion (aka wiener process)

- $m(t) = 0$
- $k(t, t') = \min(t, t')$
- The prior $f \sim \mathcal{GP}(m, k)$:



Brownian motion (aka wiener process)

- $m(t) = 0$
- $k(t, t') = \min(t, t')$
- 4 draws from $f \sim \mathcal{GP}(m, k)$:



Is Brownian motion a gp?

i. input space: $T = \mathbb{R}_+$ (time or some other unidimensional quantity).

ii. mean: $m = 0$ is a map $T \to \mathbb{R} = 0$ (a very common choice).

iii. covariance: $k(t, t') = \min(t, t')$ is positive definite. To see this, order any finite subset $t_q > t_{q-1} > ... \geq 0$, and note $s_i = t_i - t_{i-1} \geq 0$. Then:

$$
\begin{aligned}
f^\top K f &= \sum_{i=1}^{q} \sum_{j=1}^{q} f_i f_j K_{ij} \\
&= \sum_{i=1}^{q} \sum_{j=1}^{q} f_i f_j min(t_i, t_j) \\
&= \sum_{i=1}^{q} \sum_{j=1}^{q} f_i f_j \sum_{k=1}^{\min(i,j)} s_k \\
&= \sum_{k=1}^{q} s_k \sum_{i=k}^{q} \sum_{j=k}^{q} f_i f_j \\
&= \sum_{k=1}^{q} s_k (\sum_{i=k}^{q} f_i)^2 \\
&\geq 0 \quad \forall f \in \mathbb{R}^q.
\end{aligned}
$$

Whither kernel methods?

- §03 of this course will explore different types of kernels, not the underlying Hilbert space structure.
- §09 will go into reproducing kernel Hilbert spaces in depth (theory).
- Thus the beauty underlying kernel functions is only implicitly used during the gp sections.

Outline

Review: multivariate Gaussian

- $f \in \mathbb{R}^n$ is normally distributed means:

$$p(f) = (2\pi)^{-\frac{n}{2}} |K|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(f-m)^\top K^{-1}(f-m)\right\}$$

  for mean vector $m \in \mathbb{R}^n$ and covariance matrix $K \in \mathbb{R}^{n \times n}$.

- shorthand: $f \sim \mathcal{N}(m, K)$

Intuitive definition of a gaussian process

- Loosely, a multivariate Gaussian of uncountably infinite length...
  really long vector $\approx$ function

- f is a Gaussian process if $f(t) = [f(t_1), ..., f(t_n)]'$ has a multivariate normal distribution for all $t = [t_1, ..., t_n]'$:

$$f(t) \sim \mathcal{N}(m(t), k(t, t)).$$

- Let's evaluate $m(t), K(t, t)$...

Intuitive definition of a gaussian process

Mean function $m(t)$:

- any function $m : \mathbb{R} \to \mathbb{R}$ (or $m : \mathbb{R}^d \to \mathbb{R}$)
- very often $m(t) = 0 \;\; \forall \;\; t$ (mean subtract your data)

Kernel (covariance) function:

- any valid Mercer kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$
- Mercer's theorem: every matrix $K(t, t) = \{k(t_i, t_j)\}_{i,j=1\dots n}$ is a positive semidefinite (covariance) matrix $\forall t$:

$$v^T K(t,t) v = \sum_{i=1}^{n}\sum_{j=1}^{n} K_{ij} v_i v_j = \sum_{i=1}^{n}\sum_{j=1}^{n} k(t_i, t_j) v_i v_j \geq 0.$$

- (exactly what we used in the BM example a few minutes ago)

Kernel Function

Example kernel (squared exponential or SE):

$$k(t_i, t_j) = \sigma_f^2 \exp\left\{-\frac{1}{2\ell^2}(t_i - t_j)^2\right\}$$

From kernel to covariance matrix

- Choose some *hyperparameters*: $\sigma_f = 7$, $\ell = 100$

$$t = \begin{bmatrix} 0700 \\ 0800 \\ 1029 \end{bmatrix} \qquad K(t,t) = \{k(t_i, t_j)\}_{i,j} = \begin{bmatrix} 49.0 & 29.7 & 00.2 \\ 29.7 & 49.0 & 03.6 \\ 00.2 & 03.6 & 49.0 \end{bmatrix}$$

Kernel Function

Example kernel (squared exponential or SE):

$$k(t_i, t_j) = \sigma_f^2 \exp\left\{-\frac{1}{2\ell^2}(t_i - t_j)^2\right\}$$

From kernel to covariance matrix

- Choose some *hyperparameters*: $\sigma_f = 7$, $\ell = 500$

$$t = \begin{bmatrix} 0700 \\ 0800 \\ 1029 \end{bmatrix} \qquad K(t,t) = \{k(t_i, t_j)\}_{i,j} = \begin{bmatrix} 49.0 & 48.0 & 39.5 \\ 48.0 & 49.0 & 44.1 \\ 39.5 & 44.1 & 49.0 \end{bmatrix}$$

Kernel Function

Example kernel (squared exponential or SE):

$$k(t_i, t_j) = \sigma_f^2 \exp\left\{-\frac{1}{2\ell^2}(t_i - t_j)^2\right\}$$

From kernel to covariance matrix

- Choose some *hyperparameters*: $\sigma_f = 7$, $\ell = 50$

$$t = \begin{bmatrix} 0700 \\ 0800 \\ 1029 \end{bmatrix} \qquad K(t,t) = \{k(t_i, t_j)\}_{i,j} = \begin{bmatrix} 49.0 & 06.6 & 00.0 \\ 06.6 & 49.0 & 00.0 \\ 00.0 & 00.0 & 49.0 \end{bmatrix}$$

Kernel Function

Example kernel (squared exponential or SE):

$$k(t_i, t_j) = \sigma_f^2 \exp\left\{-\frac{1}{2\ell^2}(t_i - t_j)^2\right\}$$

From kernel to covariance matrix

- Choose some *hyperparameters*: $\sigma_f = 14$, $\ell = 50$

$$t = \begin{bmatrix} 0700 \\ 0800 \\ 1029 \end{bmatrix} \qquad K(t,t) = \{k(t_i, t_j)\}_{i,j} = \begin{bmatrix} 196 & 26.5 & 00.0 \\ 26.5 & 196 & 0.01 \\ 00.0 & 0.01 & 196 \end{bmatrix}$$

Outline

Important gaussian properties (in this context)

- additivity (forming a joint)
- conditioning (inference)
- expectations (posterior and predictive moments)
- marginalisation (marginal likelihood/model selection)
- ...

Additivity (joint)

- prior (or latent) $f \sim \mathcal{N}(m_f, K_{ff})$
- additive iid noise $n \sim \mathcal{N}(0, \sigma_n^2 I)$
- let $y = f + n$, then:

$$p(y, f) = p(y|f)p(f) = \mathcal{N}\left(\begin{bmatrix} f \\ y \end{bmatrix}; \begin{bmatrix} m_f \\ m_y \end{bmatrix}, \begin{bmatrix} K_{ff} & K_{fy} \\ K_{yf} & K_{yy} \end{bmatrix}\right)$$

- where (in this case):

$$K_{fy} = E[(f - m_f)(y - m_y)^T] = K_{ff} \qquad K_{yy} = K_{ff} + \sigma_n^2 I$$

- latent f and noisy observation y are jointly Gaussian

Where did the GP go?

- prior (or latent) $f \sim \mathcal{N}(m_f, K_{ff})$
- additive iid noise $n \sim \mathcal{N}(0, \sigma_n^2 I)$
- let $y = f + n$, then:

$$p(y, f) = p(y|f)p(f) = \mathcal{N}\left(\begin{bmatrix} f \\ y \end{bmatrix}; \begin{bmatrix} m_f \\ m_y \end{bmatrix}, \begin{bmatrix} K_{ff} & K_{fy} \\ K_{yf} & K_{yy} \end{bmatrix}\right)$$

- If f and y are indexed by some vector of inputs $t \in \mathbb{R}^n$:

$$m_f = \begin{bmatrix} m_f(t_1) \\ \vdots \\ m_f(t_n) \end{bmatrix} \qquad K_{ff} = \{k(t_i, t_j)\}_{i,j=1\ldots n} \qquad \ldots$$

Where did the GP go?

- prior (or latent) $f \sim \mathcal{GP}(m_f, k_{ff})$
- additive iid noise $n \sim \mathcal{GP}(0, \sigma_n^2 \delta)$
- let $y = f + n$, then:

$$p(y(t), f(t)) = p(y|f)p(f) = \mathcal{N}\left(\begin{bmatrix} f \\ y \end{bmatrix}; \begin{bmatrix} m_f \\ m_y \end{bmatrix}, \begin{bmatrix} K_{ff} & K_{fy} \\ K_{fy}^T & K_{yy} \end{bmatrix}\right)$$

- If f and y are indexed by some vector of inputs $t \in \mathbb{R}^n$:

$$m_f = \begin{bmatrix} m_f(t_1) \\ \vdots \\ m_f(t_n) \end{bmatrix} \qquad K_{ff} = \{k(t_i, t_j)\}_{i,j=1\ldots n} \qquad \ldots$$

- warning: overloaded notation - f can be infinite (GP) or finite (MVN) depending on context.

Conditioning (inference)

- The joint of f and y:

$$p\left(\begin{bmatrix} f \\ y \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} m_f \\ m_y \end{bmatrix}, \begin{bmatrix} K_{ff} & K_{fy} \\ K_{yf} & K_{yy} \end{bmatrix}\right)$$

- Massively important fact:

$$f|y \sim \mathcal{N}\left(m_f + K_{fy}K_{yy}^{-1}(y - m_y) \;\; , \;\; K_{ff} - K_{fy}K_{yy}^{-1}K_{yf}\right)$$

- Inference of latent GP, given data, is simple linear algebra.
- Tedious proof (Schur complement/LDU):

$$\begin{aligned}
& \log p(f, y) = \log p\left(\begin{bmatrix} f \\ y \end{bmatrix}\right) = \log \mathcal{N}\left(\begin{bmatrix} m_f \\ m_y \end{bmatrix}, \begin{bmatrix} K_{ff} & K_{fy} \\ K_{yf} & K_{yy} \end{bmatrix}\right) \\
\propto \;\; & -\frac{1}{2}\left(\begin{bmatrix} f \\ y \end{bmatrix} - \begin{bmatrix} m_f \\ m_y \end{bmatrix}\right)^\top \begin{bmatrix} K_{ff} & K_{fy} \\ K_{yf} & K_{yy} \end{bmatrix}^{-1} \left(\begin{bmatrix} f \\ y \end{bmatrix} - \begin{bmatrix} m_f \\ m_y \end{bmatrix}\right) \\
= \;\; & \left(\begin{bmatrix} f \\ y \end{bmatrix} - \begin{bmatrix} m_f \\ m_y \end{bmatrix}\right)^\top \begin{bmatrix} I & 0 \\ -K_{yy}^{-1}K_{yf} & I \end{bmatrix} \begin{bmatrix} \left(K_{ff} - K_{fy}K_{yy}^{-1}K_{yf}\right)^{-1} & 0 \\ 0 & K_{yy}^{-1} \end{bmatrix} \begin{bmatrix} I & -K_{fy}K_{yy}^{-1} \\ 0 & I \end{bmatrix} \left(\begin{bmatrix} f \\ y \end{bmatrix} - \begin{bmatrix} m_f \\ m_y \end{bmatrix}\right) \\
= \;\; & \left(\begin{bmatrix} f \\ y \end{bmatrix} - \begin{bmatrix} m_f + K_{fy}K_{yy}^{-1}(y - m_y) \\ m_y \end{bmatrix}\right)^\top \begin{bmatrix} \left(K_{ff} - K_{fy}K_{yy}^{-1}K_{yf}\right)^{-1} & 0 \\ 0 & K_{yy}^{-1} \end{bmatrix} \left(\begin{bmatrix} f \\ y \end{bmatrix} - \begin{bmatrix} m_f + K_{fy}K_{yy}^{-1}(y - m_y) \\ m_y \end{bmatrix}\right) \\
\propto \;\; & \log \mathcal{N}\left(m_y, K_{yy}\right) + \log \mathcal{N}\left(m_f + K_{fy}K_{yy}^{-1}(y - m_y) \;\; , \;\; K_{ff} - K_{fy}K_{yy}^{-1}K_{yf}\right) \\
= \;\; & \log p(y)p(f|y).
\end{aligned}$$

Expectation (posterior and predictive moments)

- Conditioning on data gave us:

$$f|y \sim \mathcal{N}\left(K_{fy}K_{yy}^{-1}(y-m_y)+m_f \quad , \quad K_{ff}-K_{fy}K_{yy}^{-1}K_{yf}\right)$$

- then $E[f|y] = K_{fy}K_{yy}^{-1}(y-m_y)+m_f$ (MAP, posterior mean, ...)
- Predict data observations $y^* = y(t^*)$ for some test point t^*:

$$\begin{bmatrix} y \\ y^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m_y \\ m_{y^*} \end{bmatrix}, \begin{bmatrix} K_{yy} & K_{yy^*} \\ K_{y^*y} & K_{y^*y^*} \end{bmatrix}\right)$$

- no different:

$$y^*|y \sim \mathcal{N}\left(K_{y^*y}K_{yy}^{-1}(y-m_y)+m_{y^*} \quad , \quad K_{y^*y^*}-K_{y^*y}K_{yy}^{-1}K_{yy^*}\right)$$

Marginalisation (marginal likelihood and model selection)

- Again, if:

$$\begin{bmatrix} f \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m_f \\ m_y \end{bmatrix}, \begin{bmatrix} K_{ff} & K_{fy} \\ K_{yf} & K_{yy} \end{bmatrix}\right)$$

- we can marginalize out the latent:

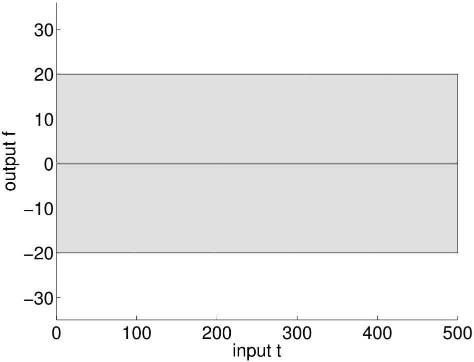
$$p(y) = \int p(y|f)p(f)df \qquad \leftrightarrow \qquad y \sim \mathcal{N}(m_y, K_{yy})$$

- marginal likelihood of the data (or $\log(p(y))$ data log-likelihood)
- In GP context, actually $p(y|\theta) = p(y|\sigma_f, \sigma_n, \ell)$. This can be the basis of model selection (§02 of this course).

Outline

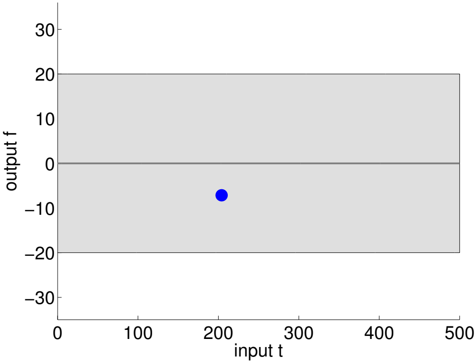
Observations

- the GP prior $p(f)$



Observations

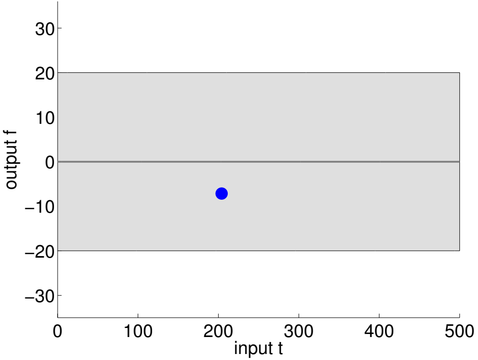
- Observe a single point at $t = 204$:
  $y(204) \sim \mathcal{N}(0, k_{yy}(204, 204)) = \mathcal{N}(0, \sigma_f^2 + \sigma_n^2)$



Observations

- Use conditioning to update the posterior:

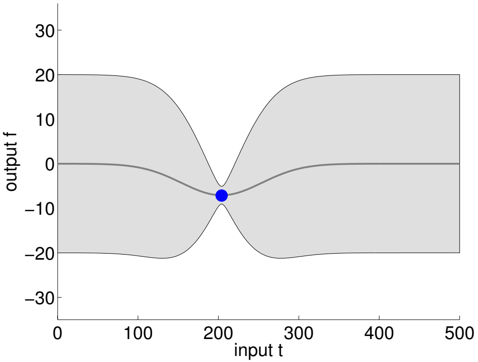
$$f|y(204) \sim \mathcal{N}\left(K_{fy}K_{yy}^{-1}(y(204) - m_y) \quad , \quad K_{ff} - K_{fy}K_{yy}^{-1}K_{fy}^{T}\right)$$



Observations

- Use conditioning to update the posterior:

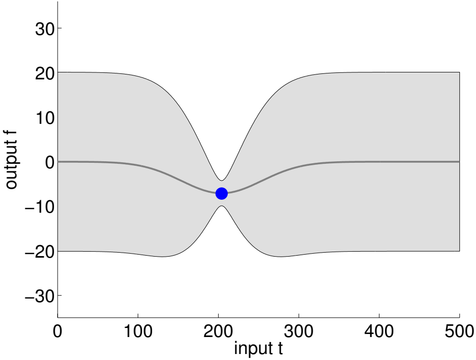
$$f|y(204) \sim \mathcal{N}\left(K_{fy}K_{yy}^{-1}(y(204)-m_y)\quad,\quad K_{ff}-K_{fy}K_{yy}^{-1}K_{fy}^{T}\right)$$



Observations

- ... and the predictive distribution:

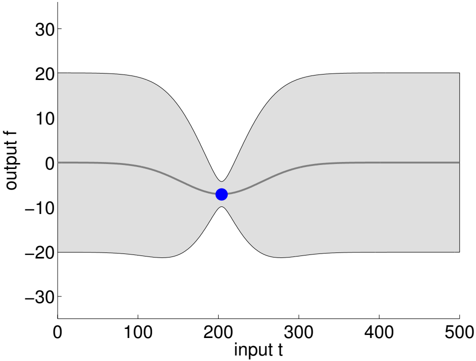
$$y^*|y(204) \sim \mathcal{N}\left(K_{y^*y}K_{yy}^{-1}(y(204) - m_y) \quad , \quad K_{y^*y^*} - K_{y^*y}K_{yy}^{-1}K_{y^*y}^T\right)$$



Observations

- More observations (data vector y):

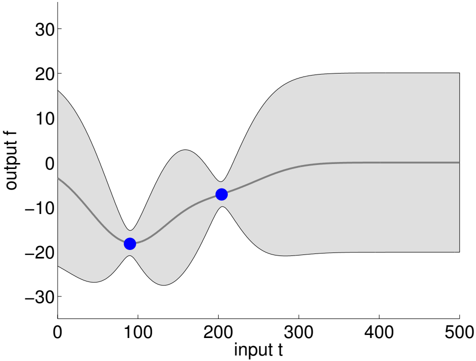
$$y^*|y(\begin{bmatrix}204\\90\end{bmatrix}) \sim \mathcal{N}\left(K_{y^*y}K_{yy}^{-1}\left(y(\begin{bmatrix}204\\90\end{bmatrix}) - m_y\right), K_{y^*y^*} - K_{y^*y}K_{yy}^{-1}K_{y^*y}^T\right)$$



Observations

- More observations (data vector y):

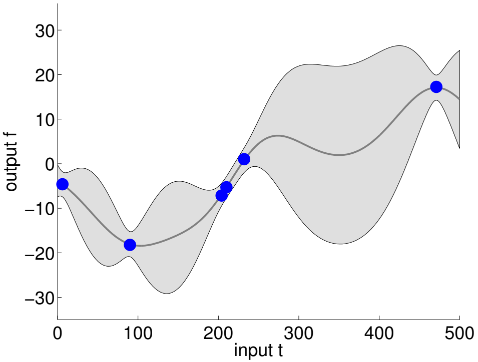
$$y^*|y(\begin{bmatrix}204\\90\end{bmatrix}) \sim \mathcal{N}\left(K_{y^*y}K_{yy}^{-1}\left(y(\begin{bmatrix}204\\90\end{bmatrix}) - m_y\right), K_{y^*y^*} - K_{y^*y}K_{yy}^{-1}K_{y^*y}^T\right)$$



Observations

- More observations (data vector y):

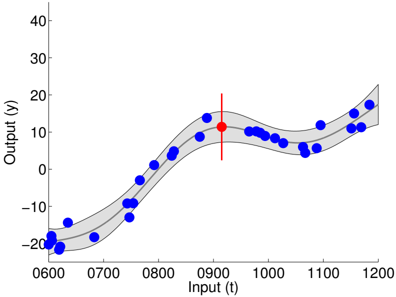
$$y^*|y \sim \mathcal{N}\left(K_{y^*y}K_{yy}^{-1}(y - m_y) \quad , \quad K_{y^*y^*} - K_{y^*y}K_{yy}^{-1}K_{y^*y}^T\right)$$



Observations

- More observations (data vector y):

$$y^*|y \sim \mathcal{N}\left(K_{y^*y}K_{yy}^{-1}(y - m_y) \quad , \quad K_{y^*y^*} - K_{y^*y}K_{yy}^{-1}K_{y^*y}^T\right)$$



Nonparametric Regression

- GP let the data speak for itself... but all the data must speak.

$$y^*|y \sim \mathcal{N}\left(K_{y^*y}K_{yy}^{-1}(y - m_y) \;\;,\;\; K_{y^*y^*} - K_{y^*y}K_{yy}^{-1}K_{y^*y}^T\right)$$

- "nonparametric models have an infinite number of parameters"

Nonparametric Regression

- GP let the data speak for itself... but all the data must speak.

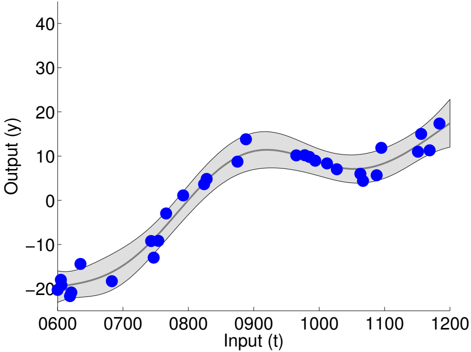
$$y^*|y \sim \mathcal{N}\left(K_{y^*y}K_{yy}^{-1}(y - m_y) \quad , \quad K_{y^*y^*} - K_{y^*y}K_{yy}^{-1}K_{y^*y}^{T}\right)$$

- ~~“nonparametric models have an infinite number of parameters”~~

- “nonparametric models have a finite but unbounded number of parameters that grows with data”

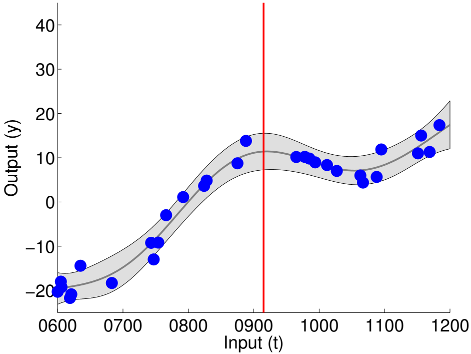
Regression: a few reminders

- denoising/smoothing



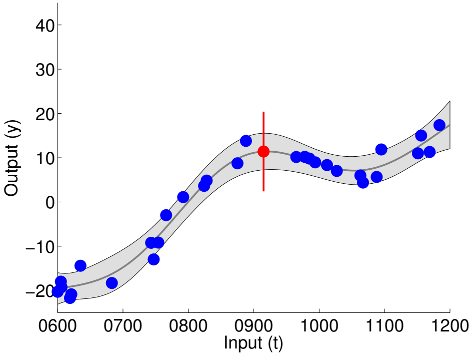
Regression: a few reminders

- denoising/smoothing
- prediction/forecasting



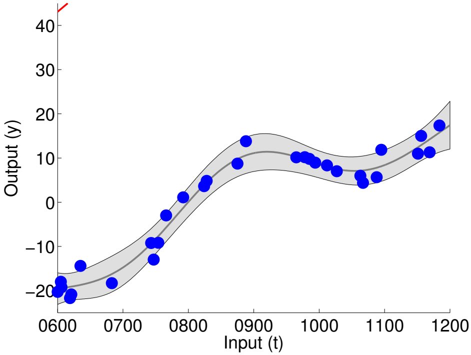
Regression: a few reminders

- denoising/smoothing
- prediction/forecasting



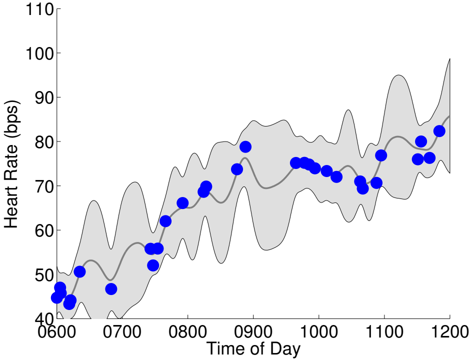
Regression: a few reminders

- denoising/smoothing
- prediction/forecasting
- dangers of parametric models



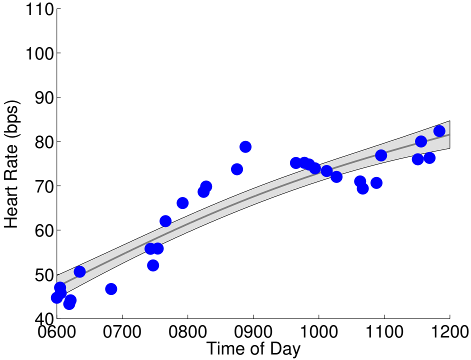
Regression: a few reminders

- denoising/smoothing
- prediction/forecasting
- dangers of parametric models
- dangers of overfitting/underfitting



Regression: a few reminders

- denoising/smoothing
- prediction/forecasting
- dangers of parametric models
- dangers of overfitting/underfitting



Outline

Useful information

- Always start with the syllabus. Highlights...
- Prerequisites (aka, did today make sense to you):
  - Stochastic processes to a basic understanding of gaussian processes
  - Machine learning such as W4400
  - Probability, statistics, linear algebra, basic convex optimization
  - Programming skills
- Grade:
  - **Homework (10%).** Two or three homework sets will be given to ensure students are keeping pace. Homework will contain both written and programming/data analysis elements.
  - **Attendance and Participation (40%).** The course will have a seminar format, and your involvement is critical. This means read in advance, and demonstrate that knowledge.
  - **Course Project (50%).** The course projects will be the focus of the latter half of this course. Projects can take a variety of forms, from contributing to open source machine learning projects, to analyzing data of interest, to advancing a theoretical topic. We will spend substantial time developing ideas for projects, tracking and discussing progress, and presenting final work product. Individual projects are ideal, though projects with groups of two may also be appropriate.

Progress...

Week	Content
1	Introduction to gaussian processes for machine learning
	• Reading: [RW06, ch. 1-2]
	• HW1 out: `https://github.com/cunni/gpkm/blob/master/hw1.ipynb`
2	Model selection
	• Reading: [RW06, ch. 5.1-5.4]; [MA10]; [GOH14, §3 only]
	• HW1 ongoing
3	Approximate inference
	• Reading: [KR05]; [RMC09]; [RW06, ch. 3; 5.5]; [HMG15]
	• HW1 due at the beginning of Monday lecture

References

[Bog98] Vladimir Igorevich Bogachev.
Gaussian measures.
Number 62. American Mathematical Soc., 1998.

[Dud02] Richard M Dudley.
Real analysis and probability, volume 74.
Cambridge University Press, 2002.

[GOH14] Roman Garnett, Michael A Osborne, and Philipp Hennig.
Active learning of linear embeddings for gaussian processes.
UAI, 2014.

[HMG15] James Hensman, Alex Matthews, and Zoubin Ghahramani.
Scalable variational gaussian process classification.
AISTATS, 2015.

[KR05] Malte Kuss and Carl Edward Rasmussen.
Assessing approximate inference for binary gaussian process classification.
The Journal of Machine Learning Research, 6:1679–1704, 2005.

[MA10] Iain Murray and Ryan P Adams.
Slice sampling covariance hyperparameters of latent gaussian models.
In *Advances in Neural Information Processing Systems*, pages 1732–1740, 2010.

[RMC09] Havard Rue, Sara Martino, and Nicolas Chopin.
Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations.
Journal of the royal statistical society: Series b (statistical methodology), 71(2):319–392, 2009.

[RW06] C. E. Rasmussen and C.K.I. Williams.
Gaussian Processes for Machine Learning.
MIT Press, Cambridge, 2006.