UNIVERSITY OF CAMBRIDGE

Machine Learning and the Physical World

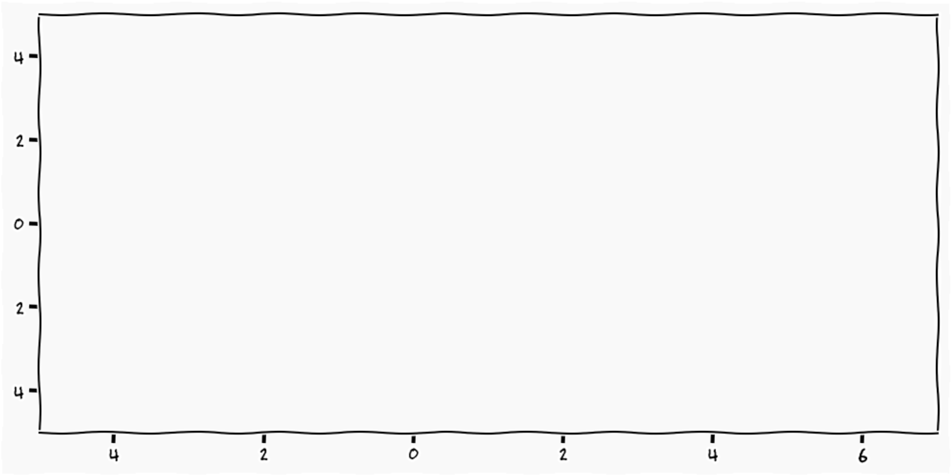
Lecture 4 : Practical Gaussian Processes

Carl Henrik Ek - che29@cam.ac.uk

22th of October, 2024

http://carlhenrik.com

Functions

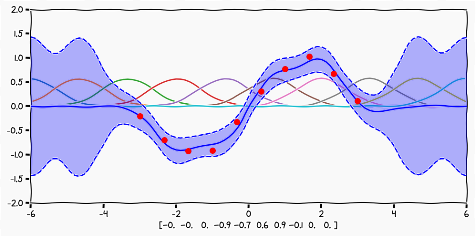


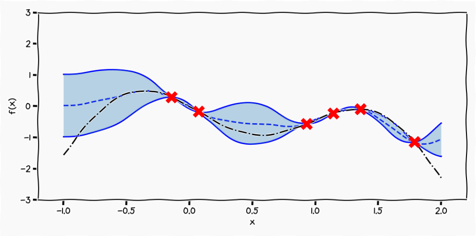
Parametrics vs. Non-parametrics

$$y_i = \underbrace{\mathbf{w}^{\mathrm{T}}}_{\text{random}} \phi(\mathbf{x}_i) + \epsilon_i$$

$$y_i = \underbrace{f_i}_{\text{random}} + \epsilon_i$$

Non-parametrics vs Parametrics





Gaussian Processes: Formalism

$$p(\mathbf{f}) = \mathcal{N}\left(\begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_N \\ \vdots \end{bmatrix} \middle| \begin{bmatrix} \mu(x_1) \\ \mu(x_2) \\ \vdots \\ \mu(x_N) \\ \vdots \end{bmatrix}, \begin{bmatrix} k(x_1,x_1) & k(x_1,x_2) & \dots & k(x_1,x_N) & \dots \\ k(x_2,x_1) & k(x_2,x_2) & \dots & k(x_2,x_N) & \dots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ k(x_N,x_1) & k(x_N,x_2) & \dots & k(x_N,x_N) & \dots \\ \vdots & \vdots & \dots & \vdots & \ddots \end{bmatrix}\right)$$

- Co-variance and mean-function both have parameters

Semantics

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{\underbrace{\int p(y \mid \theta)p(\theta)\mathrm{d}\theta}_{p(y)}}$$

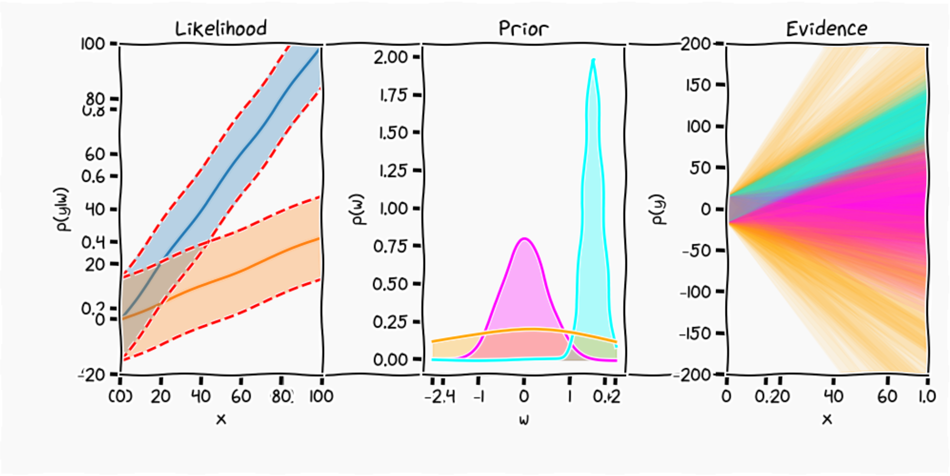
Likelihood How much evidence is there in the data for a specific hypothesis

Prior What are my beliefs about different hypothesis

Posterior What is my updated belief after having seen data

Evidence What is my belief about the data

Learning



Model Selection MacKay, 1991

Learning

$$\boldsymbol{\theta}^* = \text{argmax}_{\theta}\, p(\mathcal{D} \mid \boldsymbol{\theta})$$

Marginal Log-likelihood

df

$$
\begin{aligned}
\log p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}) &= \log \int p(\mathbf{y} \mid \mathbf{f}) p(\mathbf{f} \mid \mathbf{X}, \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \\
&- \frac{1}{2} \mathbf{y}^{\mathrm{T}} \left(k(\mathbf{X}, \mathbf{X} + \beta^{-1}\mathbf{I}) \right)^{-1} \mathbf{y} - \frac{1}{2} \log \det \left(k(\mathbf{X}, \mathbf{X}) + \beta^{-1}\mathbf{I} \right) \\
&- \frac{N}{2} \log 2\pi
\end{aligned}
$$

Today

$$p(y) = \int p(y \mid f) p(f) \mathrm{d}f$$

Code

```
import numpy as np


.....
```

Software Packages

Numerical Stability

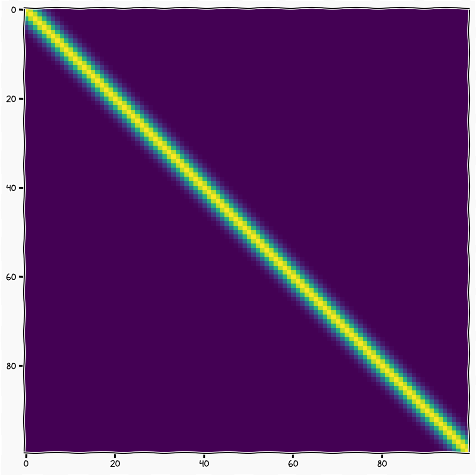
Determinant

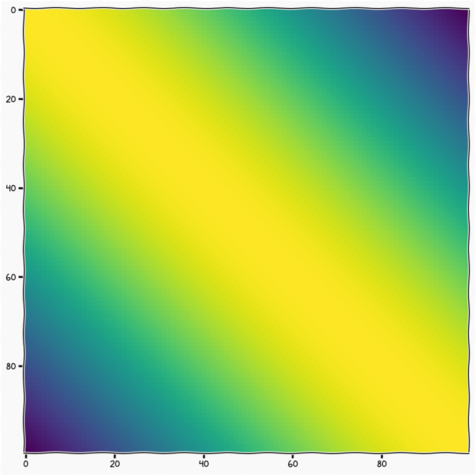
$$\frac{1}{2}\log\det\left(k(\mathbf{X},\mathbf{X})+\beta^{-1}\mathbf{I}\right)$$

Code

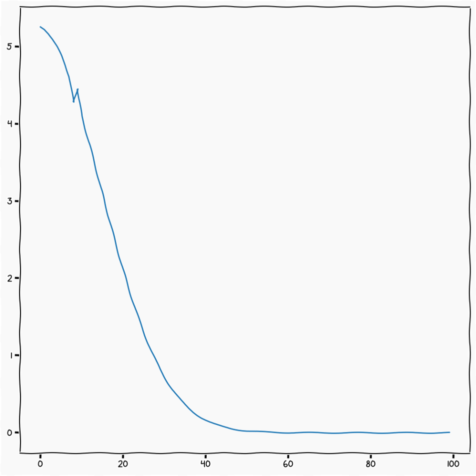
```
0.5*np.log(np.linalg.det(K+1/beta*np.eye(K.shape[0])))
```

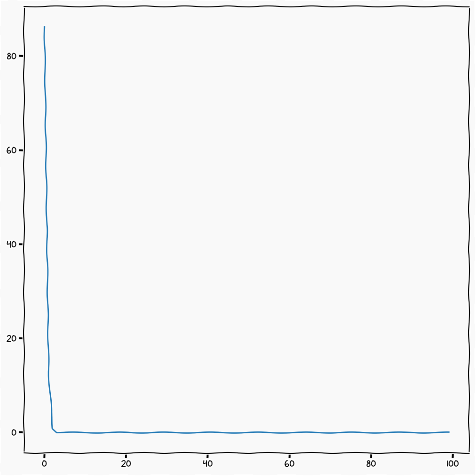
Co-Variance Matrices





Eigen-decomposition





Cholesky Decomposition

$$\mathbf{K} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$$

- Factorisation of a *Hermitian* and *Positive-Semi-Definite* Matrix into the product of two lower-triangular matrices

Determinant from Cholesky

$$\log \det \mathbf{K} = \log \det \left(\mathbf{L}\mathbf{L}^{\mathrm{T}}\right) =$$

$$\log \left(\det \mathbf{L}\right)^2 =$$

$$\log \left(\prod_i^N \ell_{ii}\right)^2 = 2 \sum_i^N \ell_{ii}$$

Inverse

$$\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y}$$

- Matrix inverse have cubic complexity $\mathcal{O}(n^3)$
- Finding the general inverse is numerically tricky
- The matrix is structured and we do not need the explicit matrix

Inverse Using Cholesky

$$
\begin{aligned}
\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} &= \mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{L}^{\mathrm{T}^{-1}}\mathbf{y} \\
&= \mathbf{y}^{\mathrm{T}}\left(\mathbf{L}^{-1}\right)^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{y} \\
&= \left(\mathbf{L}^{-1}\mathbf{y}\right)^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{y} \\
&= \mathbf{z}^{\mathrm{T}}\mathbf{z}.
\end{aligned}
$$

Inverse Using Cholesky

$$\begin{aligned}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} &= \mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{L}^{\mathrm{T}^{-1}}\mathbf{y} \\ &= \mathbf{y}^{\mathrm{T}}\left(\mathbf{L}^{-1}\right)^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{y} \\ &= \left(\mathbf{L}^{-1}\mathbf{y}\right)^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{y} \\ &= \mathbf{z}^{\mathrm{T}}\mathbf{z}. \mathbf{L}\mathbf{z} \qquad = \mathbf{y}\end{aligned}$$

Inverse using Cholesky

$$
\begin{array}{ccccccccc}
\ell_{1,1} z_1 & & & & & & & = & y_1 \\
\ell_{2,1} z_1 & + & \ell_{2,2} z_2 & & & & & = & y_2 \\
\vdots & & & & & & & & \\
\ell_{n,1} z_1 & + & \ell_{n,2} z_2 & + & \dots & + & \ell_{n,n} z_n & = & y_n,
\end{array}
$$

- we can easily solve $z_1 = \frac{y_1}{\ell_{1,1}}$ and $z_2 = \frac{y_1 - \ell_{2,1} z_1}{\ell_{2,2}}$, etc.
- `scipy.linalg.cho_solve`

Numerically Stable Computations

- A numerical method is an "approximation"

Numerically Stable Computations

- A numerical method is an "approximation"
- Our computers have finite precision

Numerically Stable Computations

- A numerical method is an "approximation"
- Our computers have finite precision
- Even "worse" they have floating finite precision

Numerically Stable Computations

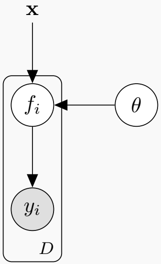
- A numerical method is an "approximation"
- Our computers have finite precision
- Even "worse" they have floating finite precision
- Keep the computer in mind when formulating your problem

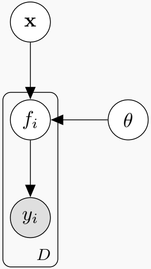
Numerically Stable Computations

- A numerical method is an "approximation"
- Our computers have finite precision
- Even "worse" they have floating finite precision
- Keep the computer in mind when formulating your problem
- There is a "big" forgotten step going from math to code, don't forget your numerical analysis

Intractabilities

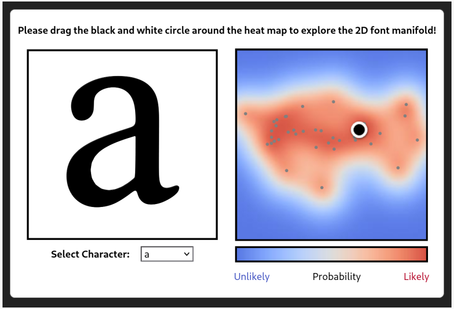
Unsupervised Learning





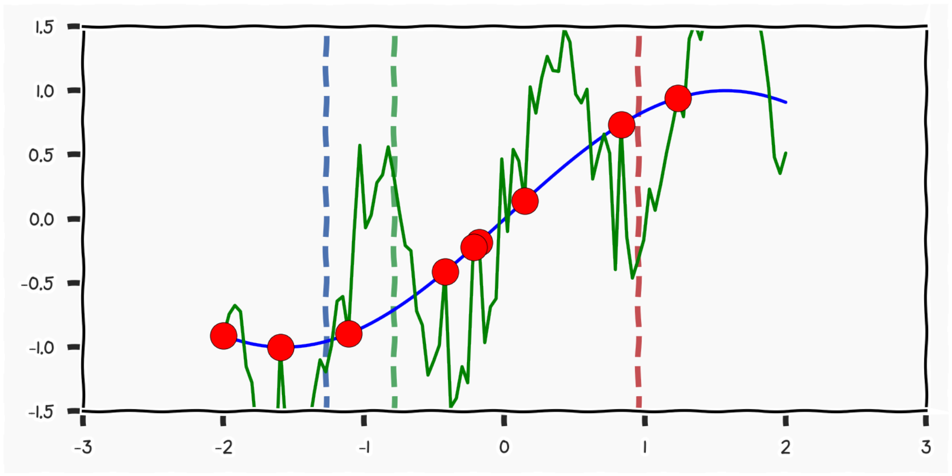
$$p(y|x) = \int p(y \mid f)p(f)\mathrm{d}f$$

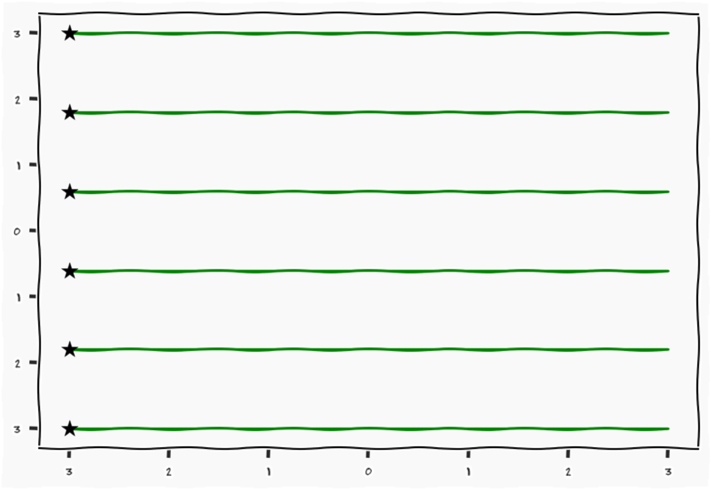
$$p(y) = \int p(y \mid f, x)p(f \mid x)p(x)\mathrm{d}f\mathrm{d}x$$

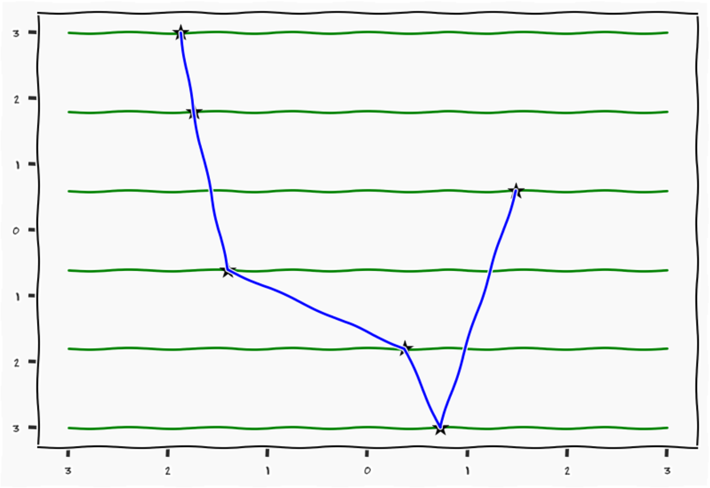


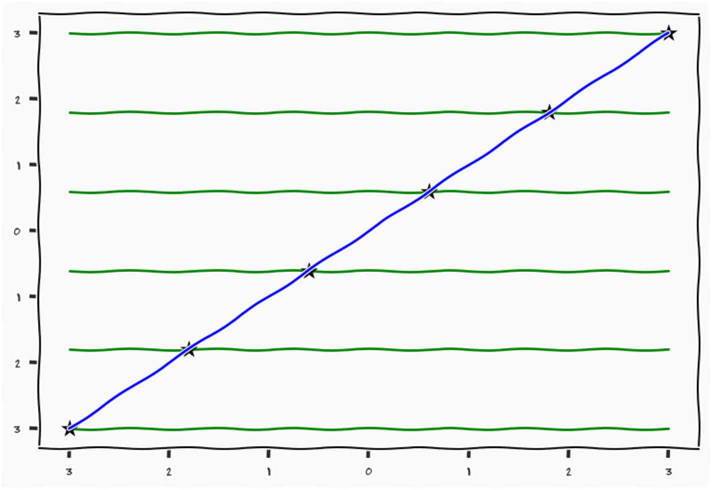
NDF Campbell et al. (July 2014). “**Learning a manifold of fonts.**” In: *ACM Transactions on Graphics (TOG)* 33.4, p. 91

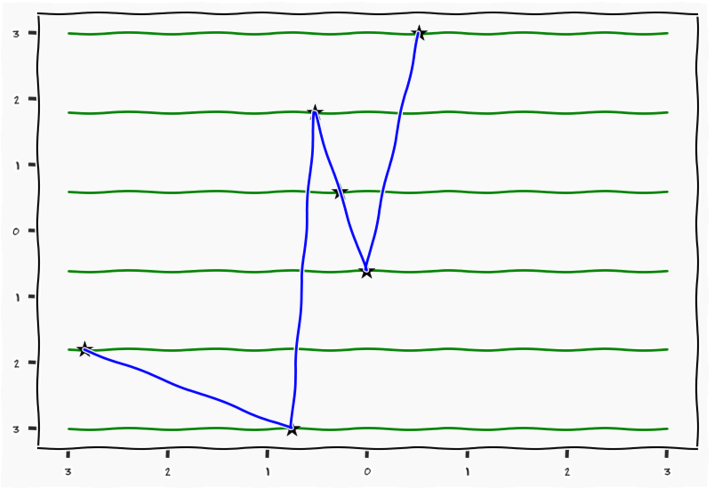
Functions



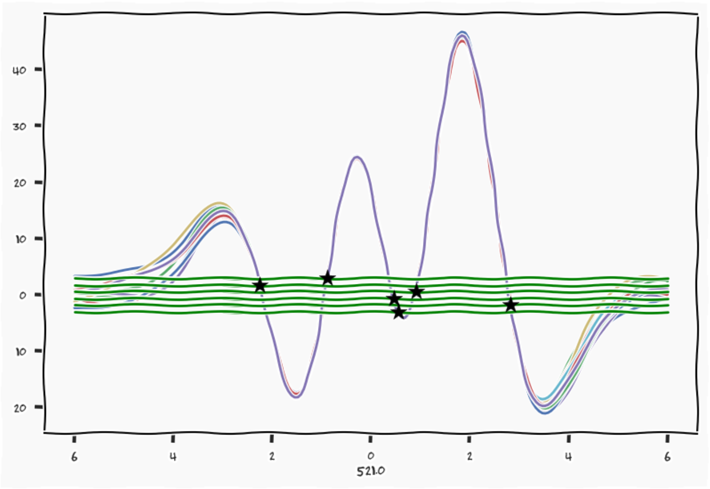

3
2
1
0
1
2
3
3
2
1
0
1
2
3


3
2
1
0
1
2
3
3
2
1
0
1
2
3

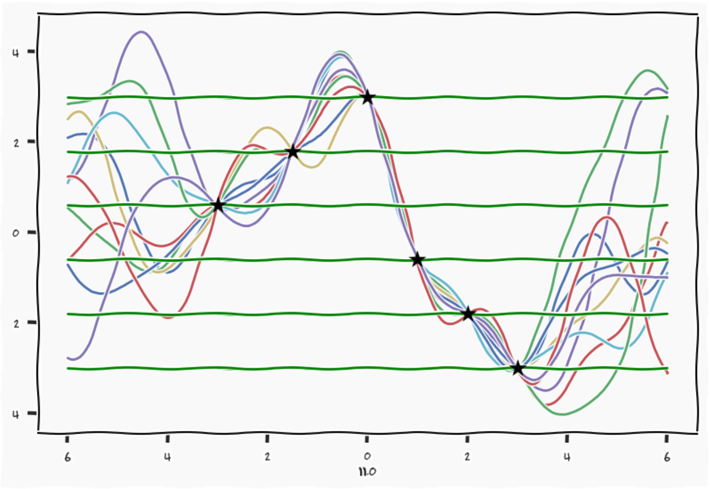

3
2
1
0
1
2
3
3
2
1
0
1
2
3


3
2
1
0
1
2
3
3
2
1
0
1
2
3

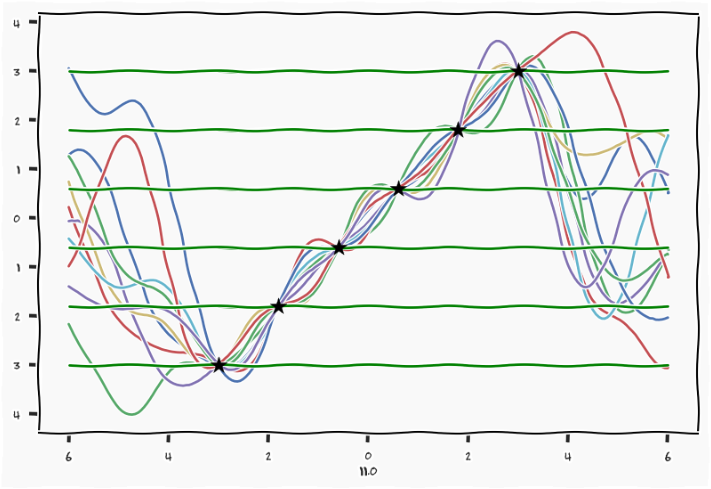
Unsupervised Learning



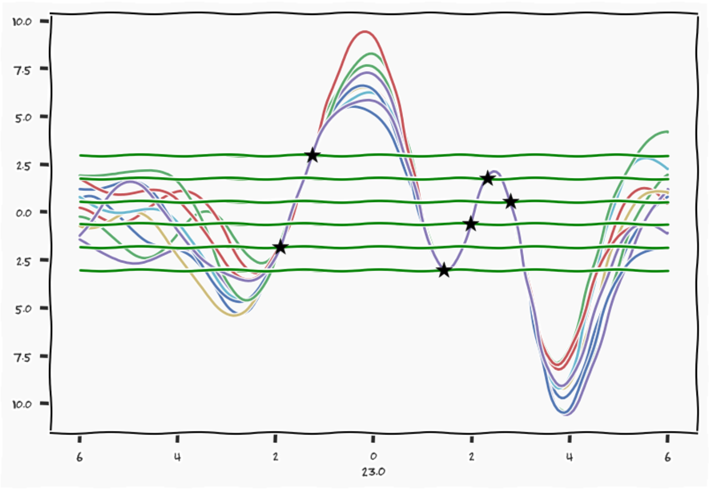
Unsupervised Learning



Unsupervised Learning



Unsupervised Learning



Gaussian Process Latent Variable Model Lawrence, 2004

Regression there are infinite number of possible functions that connects the data equally well. A GP provides a measure over these solutions that makes the problem "well-posed".

Gaussian Process Latent Variable Model Lawrence, 2004

Regression there are infinite number of possible functions that connects the data equally well. A GP provides a measure over these solutions that makes the problem "well-posed".

Unsupervised Learning there are infinite number of possible combinations of input locations and functions that generate the data equally well. A GP and a latent space prior jointly provides a measure over these solutions to make the problem "well-posed"

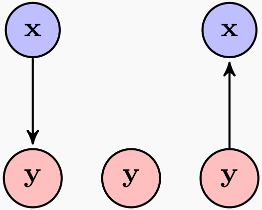
Integration

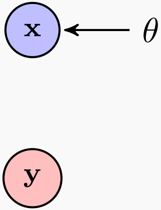
$$p(y) = \int p(y \mid f)p(f \mid x)p(x)\mathrm{d}f\mathrm{d}x$$

- This integral is analytically intractable

Approximate Inference

$$p(y) = \int p(y \mid x) p(x) \mathrm{d}x$$





$$p(y) = \int_x p(y|x)p(x) = \frac{p(y|x)p(x)}{p(x|y)}$$

$$q_\theta(x) \approx p(x|y)$$

Variational Bayes

$$p(y)$$

Variational Bayes

$$\log p(y)$$

Variational Bayes

$$\log p(y) = \log p(y) + \int \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x$$

Variational Bayes

$$
\begin{aligned}
\log p(y) &= \log p(y) + \int \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x \\
&= \int q(x) \log p(y) \mathrm{d}x + \int q(x) \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x
\end{aligned}
$$

Variational Bayes

$$\log p(y) = \log p(y) + \int \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x$$

$$= \int q(x) \log p(y) \mathrm{d}x + \int q(x) \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x$$

$$= \int q(x) \log \frac{p(x|y)p(y)}{p(x|y)} \mathrm{d}x$$

Variational Bayes

$$
\begin{aligned}
\log p(y) &= \log p(y) + \int \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x \\
&= \int q(x) \log p(y) \mathrm{d}x + \int q(x) \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x \\
&= \int q(x) \log \frac{p(x|y)p(y)}{p(x|y)} \mathrm{d}x = \int q(x) \log \frac{p(x,y)}{p(x|y)} \mathrm{d}x
\end{aligned}
$$

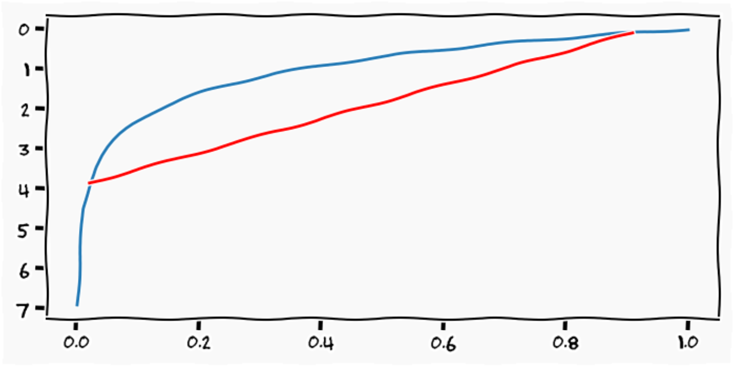
Variational Bayes

$$
\begin{aligned}
\log\, p(y) &= \log\, p(y) + \int \log\, \frac{p(x|y)}{p(x|y)} \mathrm{d}x \\
&= \int q(x) \log\, p(y) \mathrm{d}x + \int q(x) \log\, \frac{p(x|y)}{p(x|y)} \mathrm{d}x \\
&= \int q(x) \log\, \frac{p(x|y)p(y)}{p(x|y)} \mathrm{d}x = \int q(x) \log\, \frac{p(x,y)}{p(x|y)} \mathrm{d}x \\
&= \int q(x) \log\, \frac{q(x)}{q(x)} \mathrm{d}x + \int q(x) \log\, p(x,y) \mathrm{d}x + \int q(x)\, \log\, \frac{1}{p(x|y)} \mathrm{d}x
\end{aligned}
$$

Variational Bayes

$$\begin{aligned}
\log p(y) &= \log p(y) + \int \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x \\
&= \int q(x) \log p(y) \mathrm{d}x + \int q(x) \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x \\
&= \int q(x) \log \frac{p(x|y)p(y)}{p(x|y)} \mathrm{d}x = \int q(x) \log \frac{p(x,y)}{p(x|y)} \mathrm{d}x \\
&= \int q(x) \log \frac{q(x)}{q(x)} \mathrm{d}x + \int q(x) \log p(x,y) \mathrm{d}x + \int q(x) \log \frac{1}{p(x|y)} \mathrm{d}x \\
&= {\color{blue}\int q(x) \log \frac{1}{q(x)} \mathrm{d}x + \int q(x) \log p(x,y) \mathrm{d}x} + {\color{red}\int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x}
\end{aligned}$$

Jensen Inequality



$$f\left(\int g \, \mathrm{d}x\right) \leq \int f \circ g \, \mathrm{d}x,$$

The "posterior" term

$$\int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x$$

The "posterior" term

$$\int q(x)\,\log\frac{q(x)}{p(x|y)}\mathrm{d}x = -\int q(x)\,\log\frac{p(x|y)}{q(x)}\mathrm{d}x$$

The "posterior" term

$$\begin{aligned}\int q(x)\,\log\frac{q(x)}{p(x|y)}\mathrm{d}x &= -\int q(x)\,\log\frac{p(x|y)}{q(x)}\mathrm{d}x \\ &\geq \log\int p(x|y)\mathrm{d}x \\ &= \log 1 = 0\end{aligned}$$

The "posterior" term

$$\int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x$$

The "posterior" term

$$\int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x = \{\text{Lets assume that } q(x) = p(x|y)\}$$

The "posterior" term

$$\int q(x)\,\log\frac{q(x)}{p(x|y)}\mathrm{d}x = \{\text{Lets assume that } q(x) = p(x|y)\}$$

$$= \int p(x|y)\,\log\underbrace{\frac{p(x|y)}{p(x|y)}}_{=1}\mathrm{d}x$$

The "posterior" term

$$\int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x = \{\text{Lets assume that } q(x) = p(x|y)\}$$

$$= \int p(x|y) \log \underbrace{\frac{p(x|y)}{p(x|y)}}_{=1} \mathrm{d}x$$

$$= 0$$

Kullback-Leibler Divergence

$$\mathrm{KL}(q(x)||p(x|y)) = \int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x$$

- $\mathrm{KL}(q(x)||p(x|y)) \geq 0$
- $\mathrm{KL}(q(x)||p(x|y)) = 0 \Leftrightarrow q(x) = p(x|y)$
- Measure of divergence between distributions
- Not a metric (not symmetric)

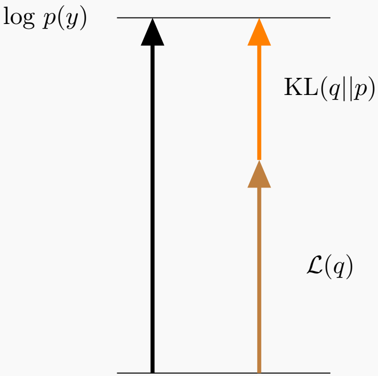
Variational Bayes

$$\log p(y) = \int q(x) \log \frac{1}{q(x)} \mathrm{d}x + \int q(x) \log p(x,y) \mathrm{d}x + \int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x$$

$$\geq - \int q(x) \log q(x) \mathrm{d}x + \int q(x) \log p(x,y) \mathrm{d}x$$

- The Evidence Lower BOnd
- Tight if $q(x) = p(x|y)$

Deterministic Approximation



ELBO

$$\log p(y) \geq -\int q(x) \log q(x) \mathrm{d}x + \int q(x) \log p(x,y) \mathrm{d}x$$
$$= \mathbb{E}_{q(x)}\left[\log p(x,y)\right] - H(q(x)) = \mathcal{L}(q(x))$$

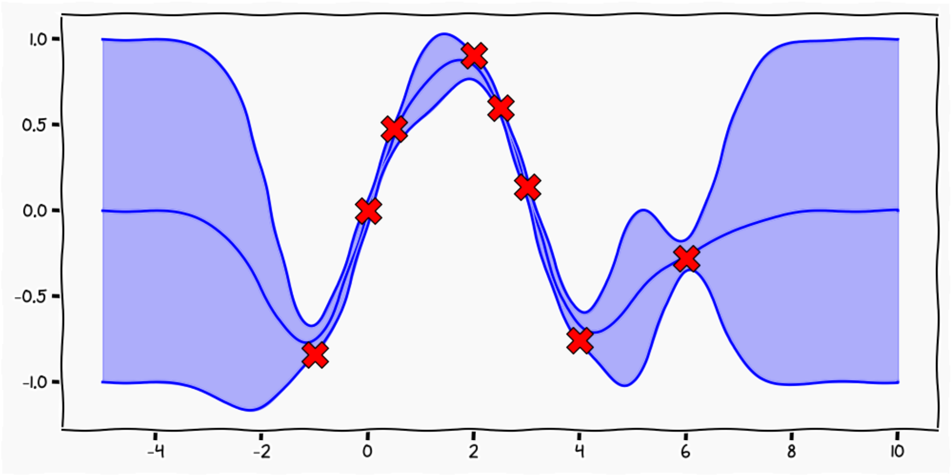
- if we maximise the ELBO we,
  - find an approximate posterior
  - lower bound the marginal likelihood
- *maximising* $p(y)$ is learning
- finding $q(x) \approx p(x|y)$ is prediction

How to choose Q?

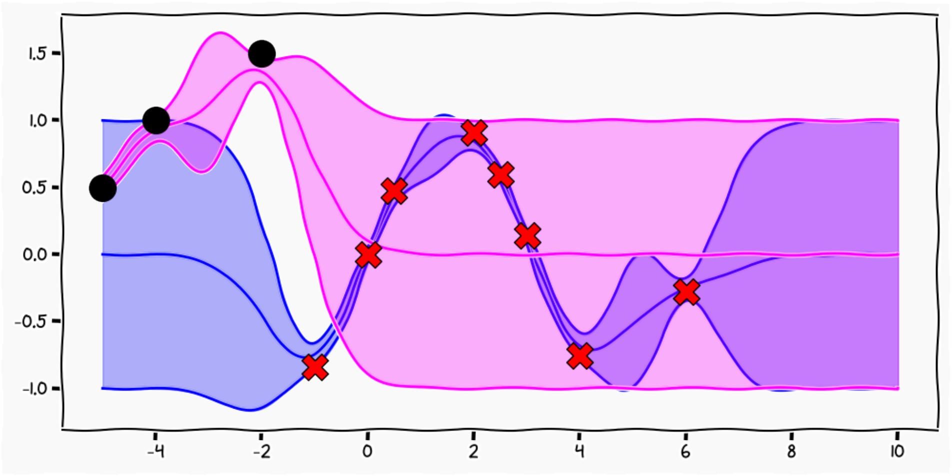
$$\mathcal{L}(q(x)) = \mathbb{E}_{q(x)}\left[\log p(x, y)\right] - H(q(x))$$

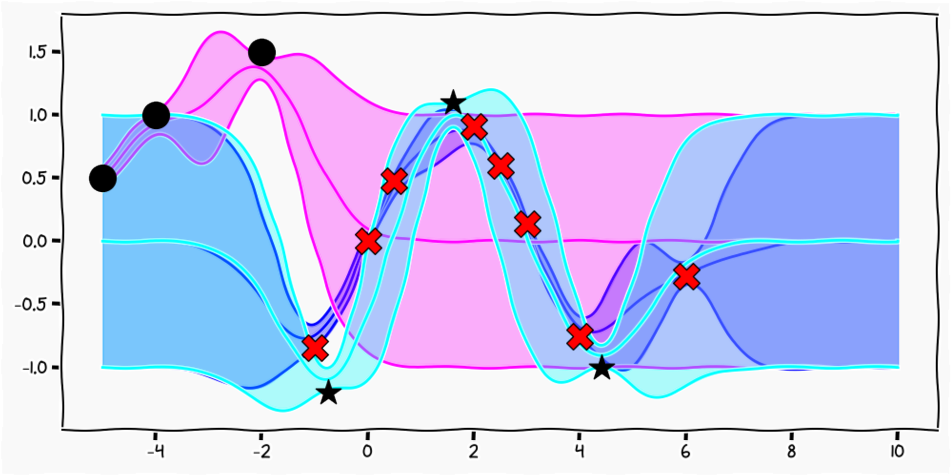
- We have to be able to compute an expectation over the joint distribution
- The second term should be trivial

Gaussian Processes Q



Gaussian Processes Q




1.5
1.0
0.5
0.0
-0.5
-1.0
-4
-2
0
2
4
6
8
10

Lower Bound [1]

$$p(f, u \mid x, z)$$

- Add another set of samples from the same prior
- Conditional distribution

[1] Titsias et al., 2010

Lower Bound [1]

$$p(f, u \mid x, z) = p(f \mid u, x, z)p(u \mid z)$$

- Add another set of samples from the same prior
- Conditional distribution

[1] Titsias et al., 2010

Lower Bound [1]

$$
\begin{aligned}
p(f, u \mid x, z) &= p(f \mid u, x, z)p(u \mid z) \\
&= \mathcal{N}(f \mid K_{fu}K_{uu}^{-1}u, K_{ff} - K_{fu}K_{uu}^{-1}K_{uf})\mathcal{N}(u \mid \mathbf{0}, K_{uu})
\end{aligned}
$$

- Add another set of samples from the same prior
- Conditional distribution

[1] Titsias et al., 2010

Lower Bound

$$p(y, f, u, x \mid z) = p(y \mid f)p(f \mid u, x)p(u \mid z)p(x)$$

- we have done nothing to the model, just project an additional set of marginals from the GP
- *however* we will now interpret u and z not as random variables but variational parameters
- i.e. the variational distribution $q(\cdot)$ is parametrised by these

Lower Bound

- Variational distributions are approximations to intractable posteriors,

$$q(u) \approx p(u \mid y, x, z, f)$$
$$q(f) \approx p(f \mid u, x, z, y)$$
$$q(x) \approx p(x \mid y)$$

Lower Bound

- Variational distributions are approximations to intractable posteriors,

$$q(u) \approx p(u \mid y, x, z, f)$$
$$q(f) \approx p(f \mid u, x, z, y)$$
$$q(x) \approx p(x \mid y)$$

- Bound is tight if u completely represents f i.e. u is sufficient statistics for f

$$q(f) \approx p(f \mid u, x, z, y) = p(f \mid u, x, z)$$

Lower Bound

$$\mathcal{L} = \int_{x,f,u} q(f)q(u)q(x) \log \frac{p(y, f, u \mid x, z)p(x)}{q(f)q(u)q(x)}$$

Lower Bound

$$\mathcal{L} = \int_{x,f,u} q(f)q(u)q(x)\log\frac{p(y,f,u \mid x,z)p(x)}{q(f)q(u)q(x)}$$
$$= \int_{x,f,u} q(f)q(u)q(x)\log\frac{p(y \mid f)p(f \mid u,x,z)p(u \mid z)p(x)}{q(f)q(u)q(x)}$$

- Assume that u is sufficient statistics of f

$$q(f) = p(f \mid u,x,z)$$

Lower Bound

$$\tilde{\mathcal{L}} = \int_{x,f,u} q(f)q(u)q(x) \log \frac{p(y \mid f)p(f \mid u, x, z)p(u \mid z)p(x)}{q(f)q(u)q(x)}$$

Lower Bound

$$\tilde{\mathcal{L}} = \int_{x,f,u} q(f)q(u)q(x)\log\frac{p(y \mid f)p(f \mid u, x, z)p(u \mid z)p(x)}{q(f)q(u)q(x)}$$
$$= \int_{x,f,u} p(f \mid u, x, z)q(u)q(x)\log\frac{p(y \mid f)p(f \mid u, x, z)p(u \mid z)p(x)}{p(f \mid u, x, z)q(u)q(x)}$$

Lower Bound

$$\begin{aligned}\tilde{\mathcal{L}} &= \int_{x,f,u} q(f)q(u)q(x)\log\frac{p(y \mid f)p(f \mid u,x,z)p(u \mid z)p(x)}{q(f)q(u)q(x)} \\ &= \int_{x,f,u} p(f \mid u,x,z)q(u)q(x)\log\frac{p(y \mid f)p(f \mid u,x,z)p(u \mid z)p(x)}{p(f \mid u,x,z)q(u)q(x)}\end{aligned}$$

Lower Bound

$$
\begin{aligned}
\tilde{\mathcal{L}} &= \int_{x,f,u} q(f)q(u)q(x)\log\frac{p(y \mid f)p(f \mid u,x,z)p(u \mid z)p(x)}{q(f)q(u)q(x)} \\
&= \int_{x,f,u} p(f \mid u,x,z)q(u)q(x)\log\frac{p(y \mid f){\color{red}p(f \mid u,x,z)}p(u \mid z)p(x)}{{\color{red}p(f \mid u,x,z)}q(u)q(x)} \\
&= \int_{x,f,u} p(f \mid u,x,z)q(u)q(x)\log\frac{p(y \mid f)p(u \mid z)p(x)}{q(u)q(x)}
\end{aligned}
$$

Lower Bound

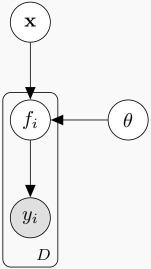
$$\begin{aligned}
\tilde{\mathcal{L}} &= \int_{x,f,u} q(f)q(u)q(x)\log\frac{p(y \mid f)p(f \mid u,x,z)p(u \mid z)p(x)}{q(f)q(u)q(x)} \\
&= \int_{x,f,u} p(f \mid u,x,z)q(u)q(x)\log\frac{p(y \mid f)\color{red}{p(f \mid u,x,z)}\color{black}p(u \mid z)p(x)}{\color{red}{p(f \mid u,x,z)}\color{black}q(u)q(x)} \\
&= \int_{x,f,u} p(f \mid u,x,z)q(u)q(x)\log\frac{p(y \mid f)p(u \mid z)p(x)}{q(u)q(x)} \\
&= \mathbb{E}_{p(f|u,x,z)}[p(y \mid f)] - \mathrm{KL}(q(u) \parallel p(u \mid z)) - \mathrm{KL}(q(x) \parallel p(x))
\end{aligned}$$

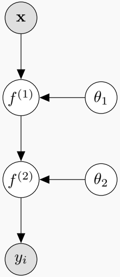
Lower Bound

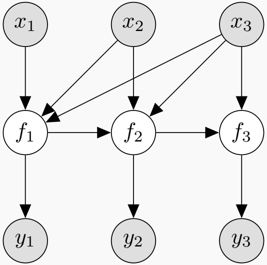
$$\mathcal{L} = \mathbb{E}_{p(f \mid u,x,z)}[p(y \mid f)] - \mathrm{KL}(q(u) \parallel p(u \mid z)) - \mathrm{KL}(q(x) \parallel p(x))$$

- Expectation tractable (for some co-variances) Titsias et al., 2010
- Stochastic inference Hensman et al., 2013
- Importantly $p(x)$ only appears in $\mathrm{KL}(\cdot \parallel \cdot)$ term!

Interesting Models







Summary

Summary

- Hopefully this gave you a flavour of the "practical" part of working with probabilistic models
- You are not expected to know this, but having it in the back of your mind
- Remember the no-free lunch, any result is relative to the assumptions that you put in
- Computations and implementations makes up a huge part of your assumptions

eof

References

References

- Campbell, NDF and J Kautz (July 2014). "**Learning a manifold of fonts.**" In: *ACM Transactions on Graphics (TOG)* 33.4, p. 91.
- Hensman, James, N Fusi, and Neil D Lawrence (2013). "**Gaussian Processes for Big Data.**" In: *Uncertainty in Artificial Intelligence*.
- Lawrence, Neil D. (2004). "**Gaussian Process Models for Visualisation of High Dimensional Data.**" In: *Advances in Neural Information Processing Systems*. Ed. by Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf. Vol. 16. Cambridge, MA: MIT Press, pp. 329–336.

MacKay, D. J. C. (1991). “**Bayesian Methods for Adaptive Models.**” PhD thesis. California Institute of Technology.

Titsias, Michalis and Neil D Lawrence (2010). “**Bayesian Gaussian Process Latent Variable Model.**” In: *International Conference on Airtificial Inteligence and Statistical Learning*, pp. 844–851.