EECS 545: Machine Learning

Lecture 25. Gaussian Process Regression

Honglak Lee and Michał Dereziński

04/11/2021

Logistics

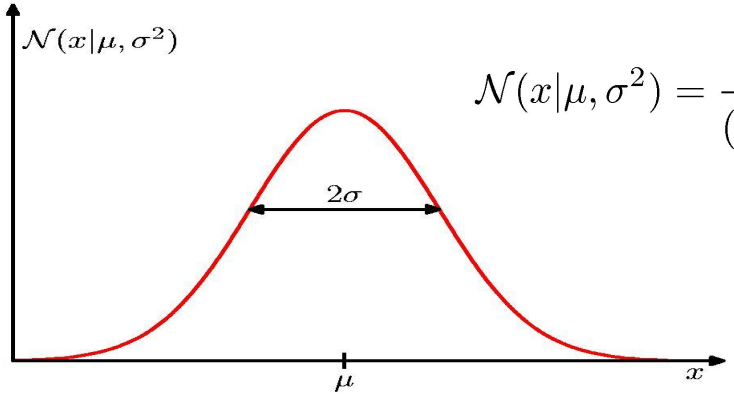
- HW6 due April 19 (last day of classes)
- Project presentations on April 21
  - In-person
  - If you are unable to be there in person, please email us and specify which group you are in
- Project final report due April 28
  - Use NeurIPS format (provided in project info document, can also be found online)
  - Type of content should be same / similar as those from the progress report

Outline

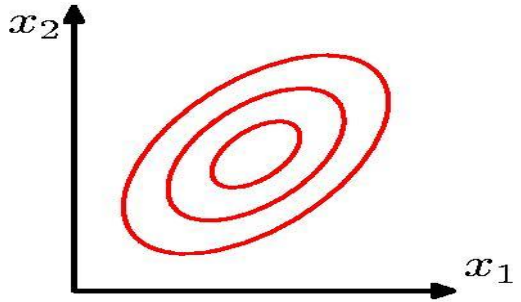
- Multivariate Gaussians
  - Marginal and conditional distributions
- Gaussian Processes
  - GP for Regression

Multivariate Gaussians – marginal and conditional distributions

The Gaussian Distribution



$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Multivariate Gaussian (mean)

$$\mathbb{E}[\mathbf{x}] = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \int \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \mathbf{x}\,\mathrm{d}\mathbf{x}$$

$$= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \int \exp\left\{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{z}\right\} (\mathbf{z}+\boldsymbol{\mu})\,\mathrm{d}\mathbf{z}$$

$$= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \int \left(\exp\left\{-\frac{1}{2}\mathbf{z}^{T}\Sigma^{-1}\mathbf{z}\right\}\mathbf{z} + \exp\left\{-\frac{1}{2}\mathbf{z}^{T}\Sigma^{-1}\mathbf{z}\right\}\mu \right) \mathbf{dz}$$

odd wrt **z** even wrt **z**

Because the integral is from (-∞, ∞),

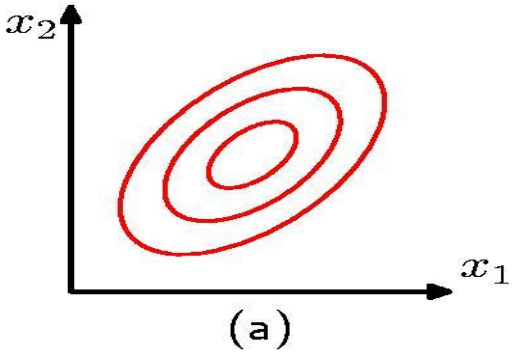
$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$$

Multivariate Gaussian (covariance)

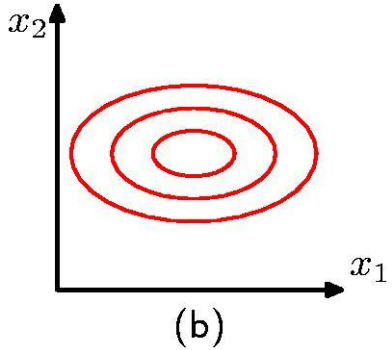
$$\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}$$

$$\mathrm{cov}[\mathbf{x}] = \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^{\mathrm{T}}\right] = \boldsymbol{\Sigma}$$

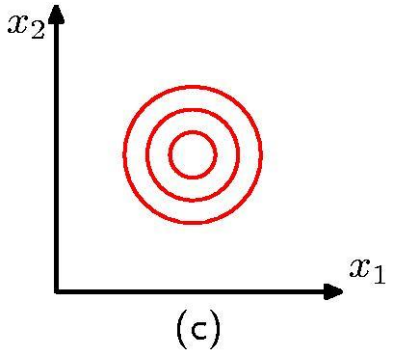
Types of covariance matrices:



General



Diagonal



Spherical
(const*Identity)

Partitioned Gaussian Distributions

- Multivariate Gaussian distribution for x

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Partitioning x into x_a and x_b.
  - Mean and covariance

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$$

  - Precision matrix

$$\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1} \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

Useful Formula for Matrix Inversion

- Woodbury Matrix Inversion Lemma

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \end{pmatrix}$$

  - where $\mathbf{M} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}$.
  - $\mathbf{M}^{-1}$ is known as Schur complement of the matrix $\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}$.

Partitioned Conditionals and Marginals

- Conditional distribution

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$$

$$\begin{aligned}
\boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \\
\boldsymbol{\mu}_{a|b} &= \boldsymbol{\Sigma}_{a|b}\left\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\right\} \\
&= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\
&= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b)
\end{aligned}$$

- Marginal distribution

$$\begin{aligned}
p(\mathbf{x}_a) &= \int p(\mathbf{x}_a, \mathbf{x}_b)\,\mathrm{d}\mathbf{x}_b \\
&= \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})
\end{aligned}$$

Derivation of conditional Gaussian

- From the exponent of the Gaussian

$$
\begin{aligned}
-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \\
-\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b) \\
-\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b).
\end{aligned}
$$

- Note that general Gaussians have the following form:

$$
-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \text{const}
$$

- Treat $\mathbf{x}_b$ as constant and rearrange terms for $\mathbf{x}_a$.

Derivation of conditional Gaussian

- Second order term:

$$-\frac{1}{2}\mathbf{x}_a^{\mathrm{T}}\boldsymbol{\Lambda}_{aa}\mathbf{x}_a \quad \Rightarrow \quad \Sigma_{a|b} = \Lambda_{aa}^{-1}.$$

- First order term:

$$\mathbf{x}_a^{\mathrm{T}}\left\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\right\}$$

– From $\Sigma_{a|b}^{-1}\mu_{a|b} = \Lambda_{aa}\mu_a - \Lambda_{ab}(x_b - \mu_b)$

$$\begin{aligned}\boldsymbol{\mu}_{a|b} &= \Sigma_{a|b}\left\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\right\} \\ &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\end{aligned}$$

Derivation of conditional Gaussian

- Expressing in terms of covariance matrix

$$\begin{aligned}\boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\ \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}.\end{aligned}$$

  - where

$$\begin{pmatrix}\boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb}\end{pmatrix}^{-1} = \begin{pmatrix}\boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb}\end{pmatrix}$$

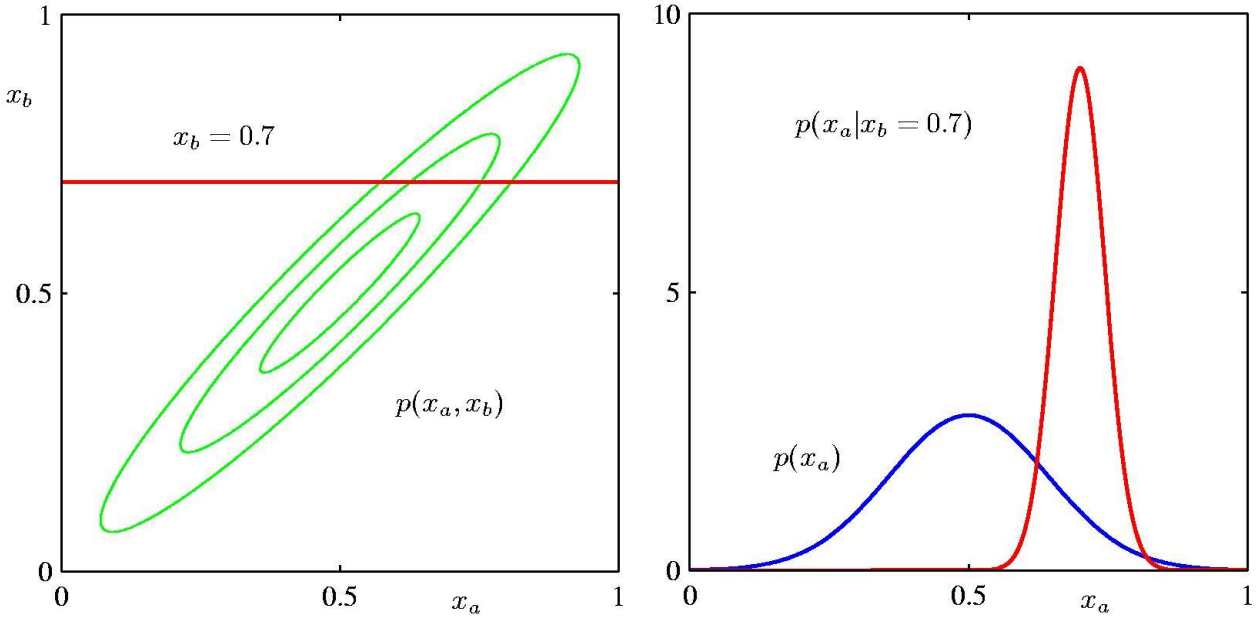
$$\begin{aligned}\boldsymbol{\Lambda}_{aa} &= (\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba})^{-1} \\ \boldsymbol{\Lambda}_{ab} &= -(\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba})^{-1}\boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}.\end{aligned}$$

- Here, we used matrix inversion lemma

$$\begin{pmatrix}\mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D}\end{pmatrix}^{-1} = \begin{pmatrix}\mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1}\end{pmatrix}$$

  - where $\mathbf{M} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}.$

Partitioned Conditionals and Marginals



Linear Gaussian Distributions

- Linear combination of Gaussian random vector also has a Gaussian distribution
  - For Gaussian random vector $X \sim \mathcal{N}(\mu_X, \Sigma_X)$, we have

$$AX + b \sim \mathcal{N}(b + A\mu_X, A\Sigma_X A^T)$$

    for matrix **A**, vector b

  - Marginal and conditional distributions are also Gaussian

Bayes' Theorem for Gaussian Variables

- Given

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right)$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b}, \mathbf{L}^{-1}\right)$$

- we arrive at the joint distribution

$$\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\mu\\ \mathbf{A}\mathbf{x}+\mathbf{b}\end{pmatrix}, \begin{pmatrix}\Lambda^{-1} & \Lambda^{-1}A^T\\ A(\Lambda^{-1})^T & A\Lambda^{-1}A^T + L^{-1}\end{pmatrix}\right)$$

- and then we derive (using derivations for conditional and marginal distributions)

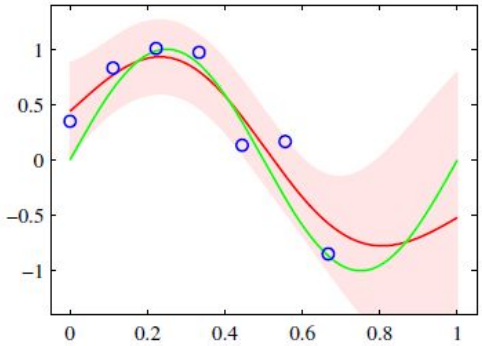
$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}+\mathbf{b}, \mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}})$$
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma})$$

  - where $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}$

Gaussian Processes

Why GPs?

- Here are some data points. What function did they come from?



- GPs are a nice way of expressing this “prior on functions” idea.

- Applications:
  - Regression
  - Classification

Definition of GP

- A Gaussian process is defined as a probability distribution over functions h(**x**), such that the set of values of h(**x**) evaluated at an arbitrary set of points $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}$ jointly have a Gaussian distribution.
  - Any finite subset of indices defines a multivariate Gaussian distribution (i.e., $h(\mathbf{x}^{(1)}), h(\mathbf{x}^{(2)}), \ldots, h(\mathbf{x}^{(n)})$) is a multivariate Gaussian.

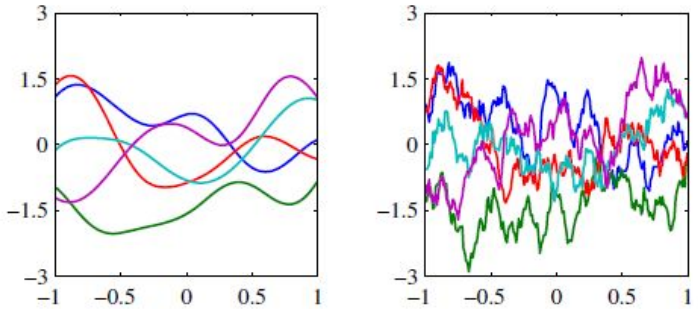
- What determines the GP is
  - The mean function μ(**x**) = E(h(**x**))
  - The covariance function (kernel)

    $$k(\mathbf{x},\mathbf{x}')=E(h(\mathbf{x})h(\mathbf{x}'))$$
  - In most applications, we take μ(x)=0. Hence the prior is represented by the kernel.

Covariance function of GP defines prior

- The figures show samples of functions drawn from Gaussian processes for two different choices of kernel functions
  - Gaussian kernel (left) vs. exponential kernel (right)

$$k(\mathbf{x}, \mathbf{x}') = \exp(-\theta \|\mathbf{x} - \mathbf{x}'\|_2{}^2)$$



$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\theta ||\mathbf{x} - \mathbf{x}'||_2\right)$$

Linear Regression Revisited – Bayesian regression

- Bayesian linear regression model: combination of M fixed basis functions given by $\phi(\mathbf{x})$, so that

$$h(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x})$$

- Prior distribution: $p(\mathbf{w}) = N(\mathbf{w}|0, \alpha^{-1}\mathbf{I})$

- Given training data points $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}$, what is the joint distribution of $h(\mathbf{x}^{(1)}), h(\mathbf{x}^{(2)}), \ldots, h(\mathbf{x}^{(n)})$?
  - **h** is the vector with elements $h^{(i)} = h(\mathbf{x}^{(i)})$, which is given by

$$\mathbf{h} = \Phi\mathbf{w}$$

  - Where Φ is the data matrix with elements $\Phi_{nk} = \Phi_k(\mathbf{x}^{(n)})$

Linear Regression Revisited – Bayesian regression

- $\mathbf{h} = \Phi\mathbf{w}$: **h** is a linear combination of Gaussian distributed variables **w**, hence itself is Gaussian.

- Mean and covariance

$$\mathrm{E}[\mathbf{h}] = \Phi\mathrm{E}[\mathbf{w}]$$

$$\mathrm{cov}[\mathbf{h}] = \mathrm{E}[\mathbf{h}\mathbf{h}^{\mathrm{T}}] = \Phi\mathrm{E}[\mathbf{w}\mathbf{w}^{\mathrm{T}}]\Phi^{\mathrm{T}} = \frac{1}{\alpha}\Phi\Phi^{\mathrm{T}} = K$$

where *K* is the Gram matrix with elements

$$\mathrm{K}_{nm} = k(\mathbf{x}^{(n)}, \mathbf{x}^{(m)}) = \frac{1}{\alpha}\Phi(\mathbf{x}^{(n)})^{T}\Phi(\mathbf{x}^{(m)})$$

and k(x,x') is the kernel function

Bayesian Linear Regression and GP

- In summary, Bayesian linear regression is a special instance of a Gaussian Process
- It is defined by the linear regression model

$$h(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x})$$

  – with a weight prior

$$p(\mathbf{w}) = N(\mathbf{w}|0, \alpha^{-1}\mathbf{I})$$

- The kernel function is given by

$$k(\mathbf{x}^{(n)}, \mathbf{x}^{(m)}) = \frac{1}{\alpha}\phi(\mathbf{x}^{(n)})^T \phi(\mathbf{x}^{(m)})$$

Features in Bayesian LR ⇔ kernel functions in GP.

GP for regression

- Consider the noise on the observed target values

$$y^{(n)} = h^{(n)} + \epsilon^{(n)}$$

  - where $\epsilon^{(n)}$ is a random noise.

- Equivalently, consider a noise process:

$$p(y^{(n)}|h^{(n)}) = \mathcal{N}(y^{(n)}|h^{(n)}, \beta^{-1})$$

  - where β is a hyperparameter (precision of the noise)

- Since $\epsilon^{(n)}$ is independent, this can be represented as multivariate Gaussian

$$p(\mathbf{y}|\mathbf{h}) = \mathcal{N}(\mathbf{y}|\mathbf{h}, \beta^{-1}\mathbf{I})$$

GP for regression

- From the definition of GP, the marginal distribution p(**h**) is given by

$$p(\mathbf{h}) = \mathcal{N}(\mathbf{h}|0, K)$$

  - Note: p(**h**) and p(**y**|**h**) form linear Gaussian

- Then, the marginal distribution of y is given by

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{h})p(\mathbf{h})d\mathbf{h} = \mathcal{N}(\mathbf{y}|0, C)$$

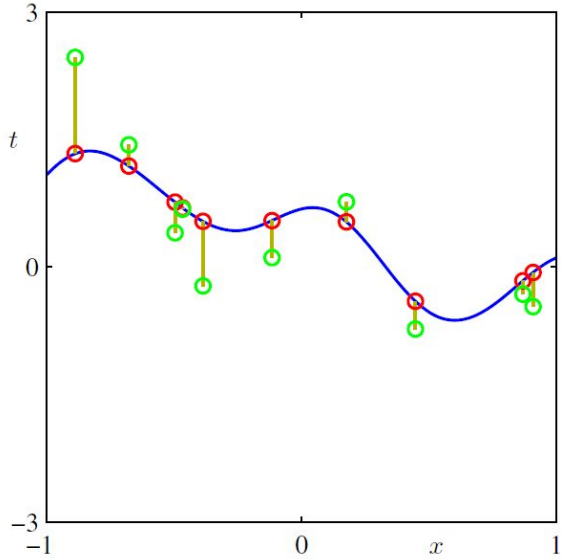
  - where the covariance matrix C has elements

Kronecker delta

$$C(\mathbf{x}^{(n)}, \mathbf{x}^{(m)}) = k(\mathbf{x}^{(n)}, \mathbf{x}^{(m)}) + \beta^{-1}\delta_{nm}$$

Example: sampling data points

- Sample function from GP (blue): $\mathrm{h}(\mathbf{x}) \sim GP(\mu, \mathrm{K})$
- Sample points from GP (red): $\left(\mathbf{x}^{(1)}, h\left(\mathbf{x}^{(1)}\right)\right), \left(\mathbf{x}^{(2)}, h\left(\mathbf{x}^{(2)}\right)\right), \left(\mathbf{x}^{(N)}, h\left(\mathbf{x}^{(N)}\right)\right) \sim GP(\mu, K)$
- Add noise (green): $y^{(n)} \sim h(\mathbf{x}^{(n)}) + \mathcal{N}(0, \beta^{-1})$



GP for regression

- We have used GP to build a model of the joint distribution over sets of data points
- Goal: Given data $\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right)$ and target values $\mathbf{y} = \left(y^{(1)}, \ldots, y^{(n)}\right)$, predict $y^{(n+1)}$ for query point $\mathbf{x}^{(n+1)}$.
- Idea: GP assumes that

$$p\left(y^{(1)}, \ldots y^{(n)}, y^{(n+1)}\right) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C}_{n+1})$$

  - where $\boldsymbol{C}_{n+1}$ is a (n+1)x(n+1) matrix

$$C_{n+1} = \begin{pmatrix} C_n & k \\ k^T & \mathbf{c} \end{pmatrix}, \text{ where } C_n \text{ is } \mathbf{n} \times \mathbf{n} \text{ matrix, and } c = k(\mathbf{x}^{(n+1)}, \mathbf{x}^{(n+1)}) + \beta^{-1}$$

$$C_n[i,j] = k(x^{(i)}, x^{(j)}) + \beta^{-1}\delta_{ij}$$

k is n-dim vector where $k_i = k(x^{(n+1)}, x^{(i)})$

GP for regression

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$$
$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b)$$
$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}$$

- The conditional distribution $p(y^{(n+1)}|\boldsymbol{y})$ is a Gaussian distribution with mean and covariance given by

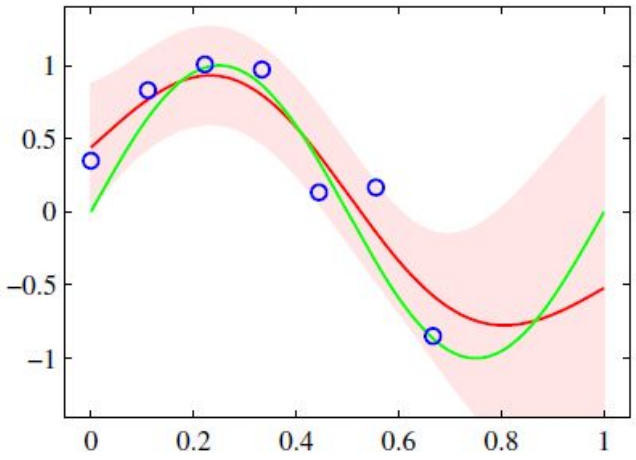
$$m(\mathbf{x}^{(n+1)}) = k^T C_n^{-1}\mathbf{y}$$
$$\sigma^2(\mathbf{x}^{(n+1)}) = c - k^T C_n^{-1} k$$

$$C_{n+1} = \begin{pmatrix} C_n & k \\ k^T & \mathrm{c} \end{pmatrix}, \text{ where } C_n \text{ is n} \times \text{n matrix, and } c = k(\mathbf{x}^{(n+1)}, \mathbf{x}^{(n+1)}) + \beta^{-1}$$

- These are the key results that define Gaussian process regression.
- The predictive distribution is a Gaussian whose mean and variance both depend on $\boldsymbol{x}^{(n+1)}$

An Example of GP regression

- Green: underlying function (sine function)
- Blue: samples from GP (with noise)
- Red: prediction from GP regression with “error bars”



GP for Regression

- The only restriction on the kernel is that the covariance matrix given by

$$C(\mathbf{x}^{(n)}, \mathbf{x}^{(m)}) = k(\mathbf{x}^{(n)}, \mathbf{x}^{(m)}) + \beta^{-1}\delta_{nm}$$

must be positive definite.

- GP will involve a matrix of size N*N, for which require $O(N^3)$ computations.

Learning Hyperparameters

- We can learn hyperparameters $\boldsymbol{\theta}$ as coefficients for the kernel function
  - e.g. exponential quadratic kernel

$$k(\mathbf{x}^{(n)}, \mathbf{x}^{(m)}) = \theta_0 \exp\left\{ -\frac{\theta_1}{2} ||\mathbf{x}^{(n)} - \mathbf{x}^{(m)}||^2 \right\} + \theta_2 + \theta_3 (\mathbf{x}^{(n)})^T \mathbf{x}^{(m)}$$

- Log likelihood

$$\ln p(\mathbf{y}|\theta) = -\frac{1}{2}\ln|C_N| - \frac{1}{2}\mathbf{y}^T C_N^{-1}\mathbf{y} - \frac{N}{2}\ln(2\pi)$$

- Gradient Ascent for parameter θ

$$\frac{\partial}{\partial \theta_i}\ln p(\mathbf{y}|\theta) = -\frac{1}{2}\mathrm{Tr}\left(C_N^{-1}\frac{\partial C_N}{\partial \theta_i}\right) + \frac{1}{2}\mathbf{y}^T C_N^{-1}\frac{\partial C_N}{\partial \theta_i}C_N^{-1}\mathbf{y}$$

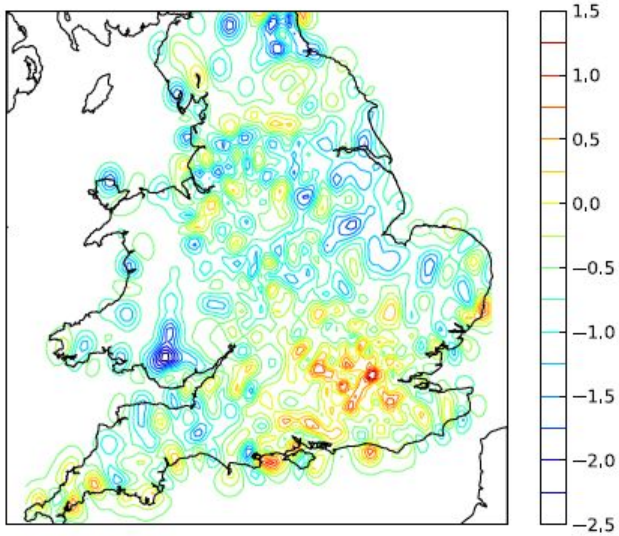
  – where we used the following:

$$\frac{\partial}{\partial x}\left(\mathbf{A}^{-1}\right) = -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}\mathbf{A}^{-1}$$

$$\frac{\partial}{\partial x}\ln|\mathbf{A}| = \mathrm{Tr}\left(\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}\right)$$

GP applications

- Modeling variability of the apartment price



James Hensman, Nicolo Fusi, and Neil D. Lawrence.
"Gaussian processes for big data." *arXiv preprint arXiv:1309.6835* (2013).

GP applications

- Modeling atmospheric CO2 concentrations

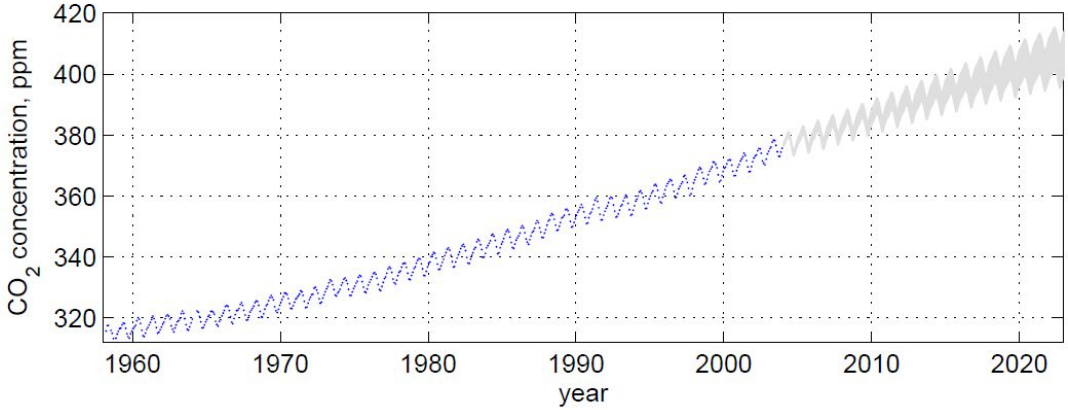


Figure 5.6: The 545 observations of monthly averages of the atmospheric concentration of CO_2 made between 1958 and the end of 2003, together with 95% predictive confidence region for a Gaussian process regression model, 20 years into the future. Rising trend and seasonal variations are clearly visible. Note also that the confidence interval gets wider the further the predictions are extrapolated.

C. E. Rasmussen & C. K. I. Williams
"Gaussian processes for machine learning." *the MIT Press* 2.3 (2006): 4.

Summary of GP

- **Practically,** you can use GP to model mean and uncertainty of your prediction.

- GP is a prior distribution over functions
- GP generates data points that are jointly a Gaussian distribution

- Most interesting structure is in k(**x**,**x**'), the kernel.
- GP can be used for regression to predict the target for a new input example.

Quiz

Link here or QR code below