Stochastic Processes

Week 05 (Version 1.0)

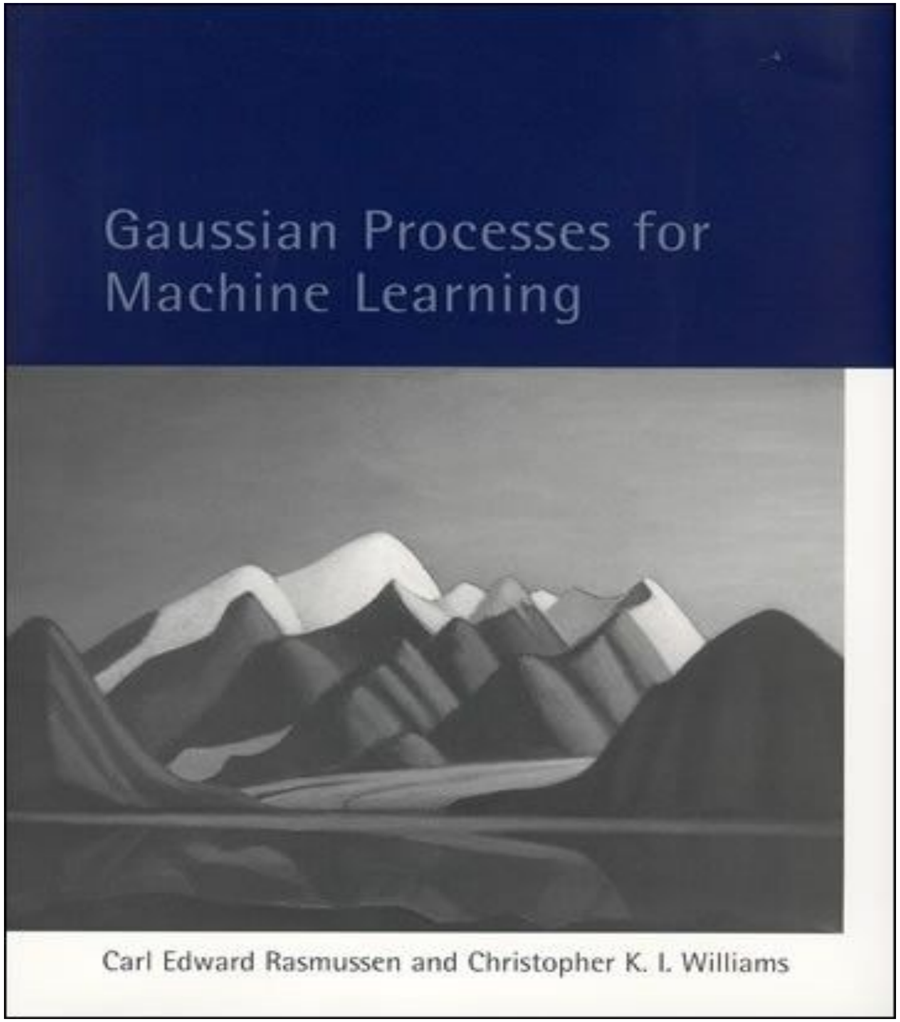
Gaussian Processes

Hamid R. Rabiee

Fall 2024

Main Reference Textbook

http://gaussianprocess.org/gpml/chapters/RW.pdf



Outline

Gaussian Process Motivation

- In many real-world applications we are confronted by the need for a model.
- We apply our knowledge of the physics involved to deduce a specific model form.
- Often our knowledge of the underlying physical processes is useless or involves too many assumptions or variables we cannot measure.
- In these cases machine learning attempts to solve the problem by learning relationships from existing data or measurements (empirical models)
- Impressive results from Deep Neural Networks (DNN). However, they are black box (not interpretable) in general, and are vulnerable to adversarial attacks.

Gaussian Process Motivation

- DNN solutions are still far from commonplace in some applied engineering for a variety of reasons such as data availability, processing complexity, robustness, …
- Gaussian Processes can be considered as a suitable model in many applications.

What is a Gaussian Process (GP)

- A Gaussian Process (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution.
- With GP we seek to find the most likely function that could produce our data using a Bayesian approach.
- Gaussian Processes provide a machine learning approach where uncertainty in the model is concretely available and the model can be used to gain physical insight into the process.
- Gaussian Processes can be considered as neural networks with infinitely many weights where a distribution is assigned to each weight (GP is a distribution over functions)

GP as Distribution over Functions

- A Gaussian process $f(x) \sim \mathcal{GP}(m(x), k(x, x'))$
- is completely specified by a mean and covariance function:

$$
\begin{aligned}
m(x) &= \mathbb{E}[f(x)] \\
k(x, x') &= \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))]
\end{aligned}
$$

Outline

1. Motivation
2. The Gaussian Distribution
3. Covariance Functions
4. Gaussian Process
5. Basis Function Representations
6. Constructing Covariance
7. Gaussian Process Limitations
8. Conclusion

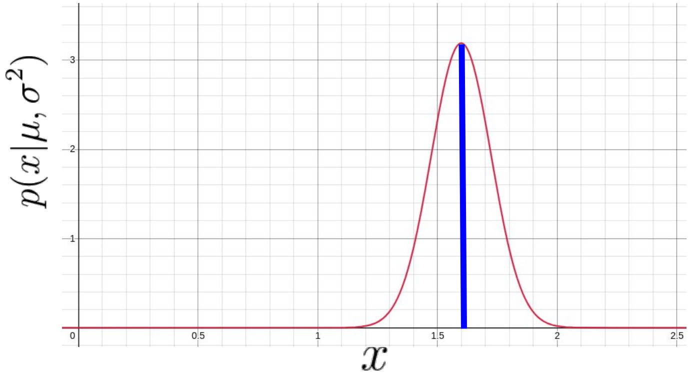
Gaussian Density Function

Gaussian: The most common probability density function. It is completely specified by its mean and variance:

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

Gaussian PDF with mean 1.6 and variance 0.125.

Blue vertical line shows the mean.



Important Gaussian Properties

- Sum of independent Gaussian variables is also Gaussian:

$$y_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \quad 1 \leq i \leq n$$

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma^2\right)$$

- As sum increases, sum of non-Gaussian, finite variance variables is also Gaussian [Central Limit Theorem].

Important Gaussian Properties

- Scaling a Gaussian leads to a Gaussian.

$$y \sim \mathcal{N}(\mu, \sigma^2)$$
$$ay \sim \mathcal{N}(a\mu, a^2\sigma^2)$$

Multivariate Gaussian Density Function

The multivariate normal distribution of a k-dimensional random vector $\mathbf{x} = (X_1, \ldots, X_k)^{\mathrm{T}}$ can be written as:

$$\mathbf{X}\ \sim\ \mathcal{N}(\boldsymbol{\mu},\, \boldsymbol{\Sigma})$$

Where:

$$\boldsymbol{\mu} = \mathrm{E}[\mathbf{X}] = (\mathrm{E}[X_1], \mathrm{E}[X_2], \ldots, \mathrm{E}[X_k])^{\mathbf{T}}$$

$$\Sigma_{i,j} = \mathrm{E}[(X_i - \mu_i)(X_j - \mu_j)] = \mathrm{Cov}[X_i, X_j]$$

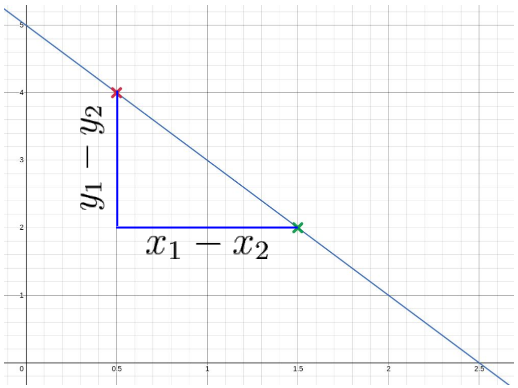
$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

Case Study:
System of Equations
Random Lines

Two Simultaneous Equations

A system of **two** equations with **two** unknowns.

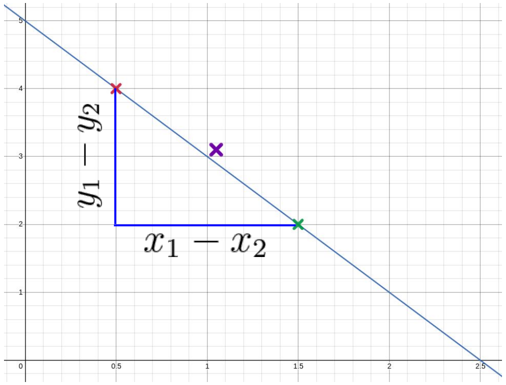
$$y_1 = mx_1 + c$$
$$y_2 = mx_2 + c$$
$$y_1 - y_2 = m(x_1 - x_2)$$
$$m = \frac{y_1 - y_2}{x_1 - x_2}$$
$$c = y_1 - mx_1$$



Three Simultaneous Equations

How do we deal with **three** simultaneous equations with only **two** unknowns?

$$y_1 = mx_1 + c$$
$$y_2 = mx_2 + c$$
$$y_3 = mx_3 + c$$



Overdetermined Systems

With two unknowns and two observations:

$$y_1 = mx_1 + c$$
$$y_2 = mx_2 + c$$

Additional observation leads to overdetermined system:

$$y_3 = mx_3 + c$$

This problem is solved through a noise model

$$y_1 = mx_1 + c + \epsilon_1$$
$$y_2 = mx_2 + c + \epsilon_2$$
$$y_3 = mx_3 + c + \epsilon_3$$

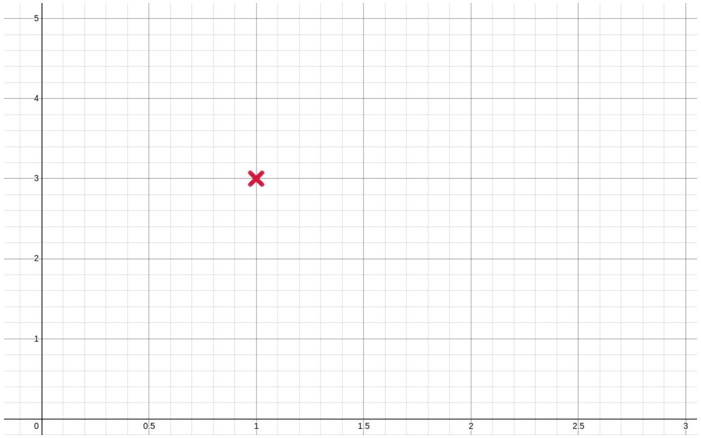
Noise Models

- We aren’t modeling the entire system.
- Noise model gives mismatch between model and data.
- Gaussian model justified by appeal to central limit theorem.
- Other models also possible (Student-t for heavy tails).
- Maximum likelihood with Gaussian noise leads to least squares.

Underdetermined Systems

What about **two** unknowns and **one** observation?

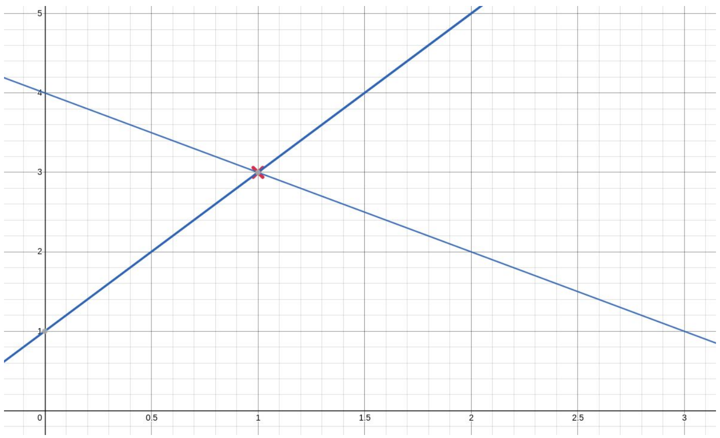
$$y_1 = mx_1 + c$$



Underdetermined System

We can compute m given c:

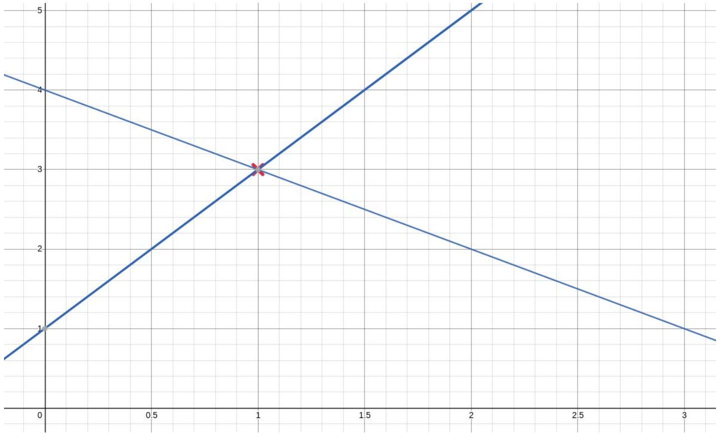
$$c = 4 \Rightarrow m = -1$$



Underdetermined System

We can compute m given c:

$$c = 1 \Rightarrow m = 2$$

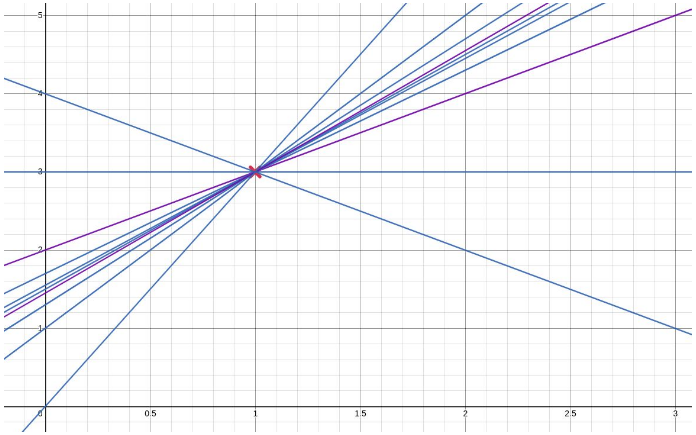


Underdetermined System

We can compute m given c.

Assume:

$$c \sim \mathcal{N}(1.5, 0.75)$$



We find a distribution of solutions.

Probability for Under- and Overdetermined Systems

- To deal with overdetermined system, introduced probability distribution for **variable**, ϵ_i .
- For underdetermined system, introduced probability distribution for **parameter**, c.
- We can solve this problem with a GP model (the random line example).
- This is known as a Bayesian treatment.

Probability for Under- and Overdetermined Systems

- For general Bayesian inference we need multivariate priors.
- E.g. for multivariate linear regression:

$$y_i = \sum_j w_{i,j} x_j + \epsilon_i$$

$$y_i - \mathbf{w}^T \mathbf{x}_{i,:} = \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

(where we've dropped c for convenience)

- We need distribution over parameters (w) and variables (ϵ_i). This motivates a multivariate Gaussian density.

Multivariate Regression Likelihood

- Multivariate regression likelihood:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x}_{i,:})^2\right)$$

- Now we use a multivariate Gaussian prior:

$$p(\mathbf{w}) = \frac{1}{(2\pi\alpha)^{\frac{p}{2}}} \exp\left(-\frac{1}{2\alpha} \mathbf{w}^T \mathbf{w}\right)$$

Posterior Density

- If compute the posterior, we get to Gaussian distribution again:

$$\log p(\mathbf{w}|\mathbf{y},\mathbf{X}) = \log p(\mathbf{y}|\mathbf{w},\mathbf{X}) + \log p(\mathbf{w}) + const.$$

$$= -\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2 + \frac{2}{2\sigma^2}\sum_{i=1}^{n} y_i \mathbf{x}_{i,:}^T \mathbf{w}$$

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n} \mathbf{w}^T \mathbf{x}_{i,:} \mathbf{x}_{i,:}^T \mathbf{w} - \frac{1}{2\alpha}\mathbf{w}^T\mathbf{w} + const.$$

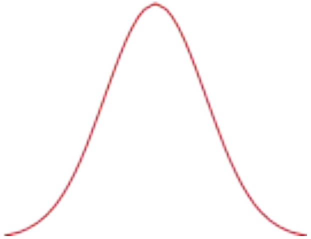
$$p(\mathbf{w}|\mathbf{y},\mathbf{X}) = \mathcal{N}(\mathbf{w}|\mu_{\mathbf{w}}, \mathbf{C}_{\mathbf{w}})$$

$$\mathbf{C}_{\mathbf{w}} = (\sigma^{-2}\mathbf{X}^T\mathbf{X} + \alpha^{-1})^{-1} \qquad \mu_{\mathbf{w}} = \mathbf{C}_{\mathbf{w}}\sigma^{-2}\mathbf{X}^T\mathbf{y}$$

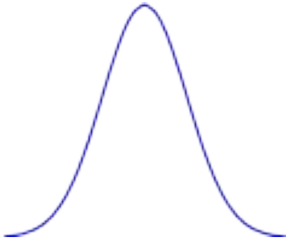
Case Study:
Two Dimensional Gaussian
Height vs Weight

Height and Weight Models

Gaussian distributions for weight and height:



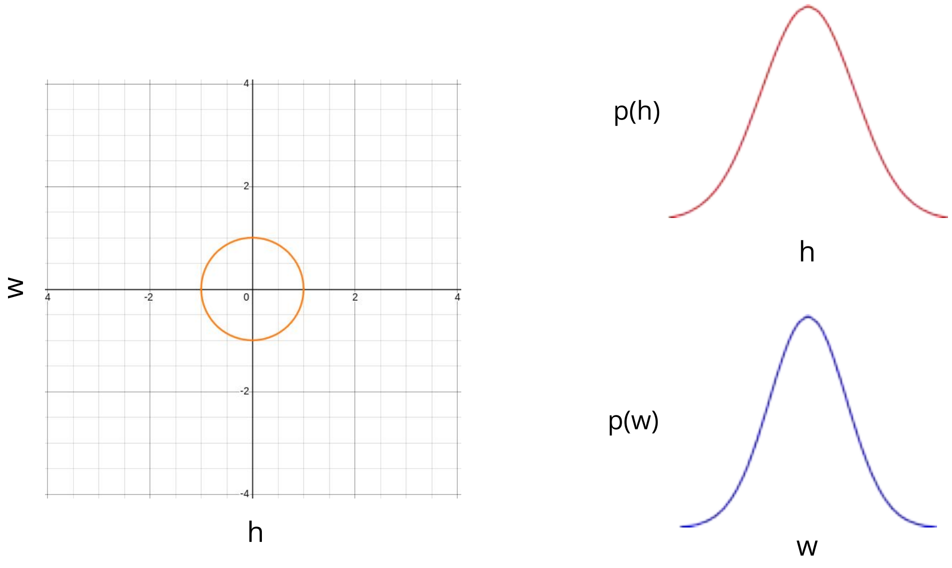
Height distribution



Weight distribution

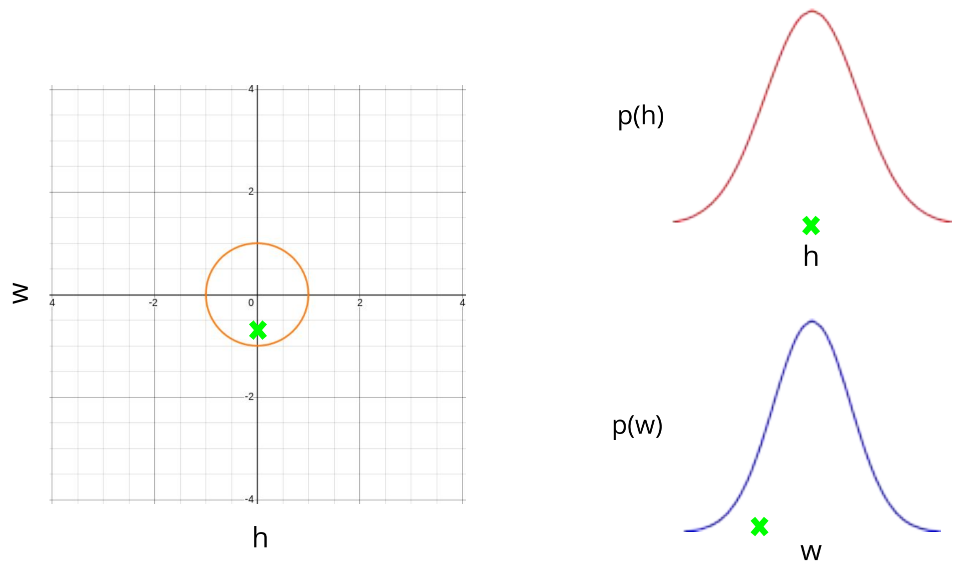
Sampling Two Dimensional Variables

Sample height and weight one after the other and plot against each other:



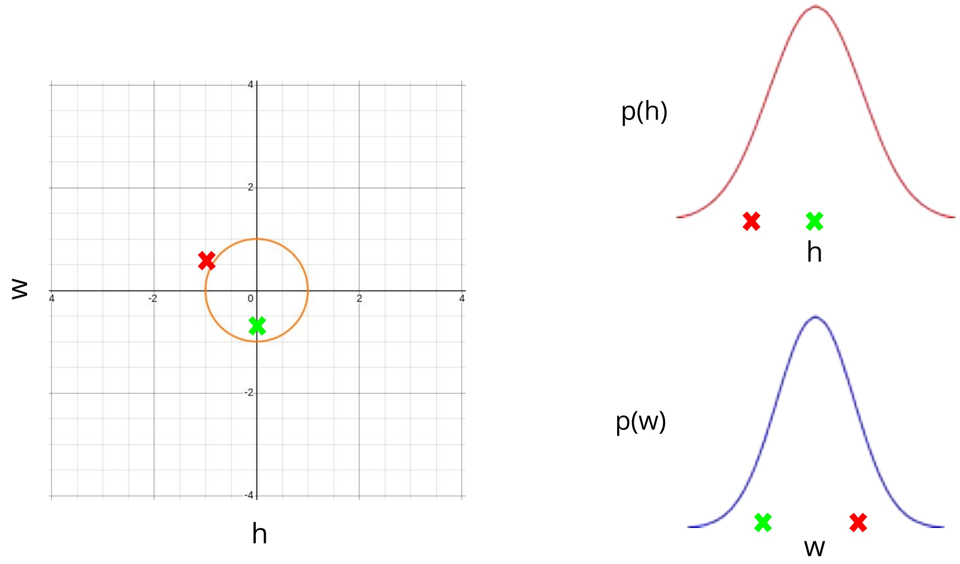
Sampling Two Dimensional Variables

Sample height and weight one after the other and plot against each other:



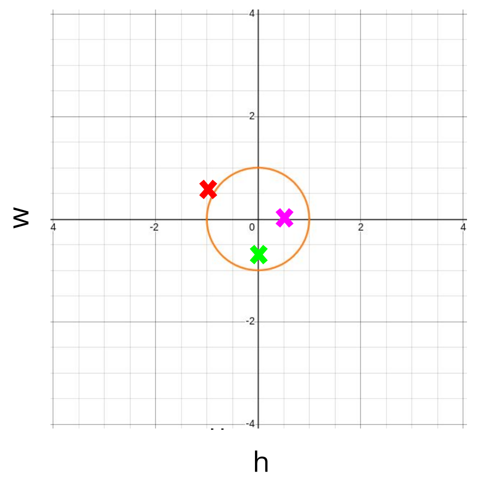
Sampling Two Dimensional Variables

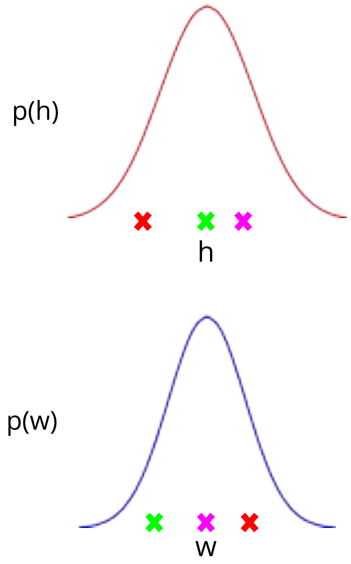
Sample height and weight one after the other and plot against each other:



Sampling Two Dimensional Variables

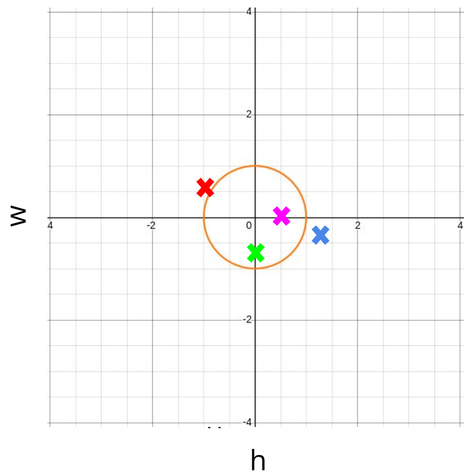
Sample height and weight one after the other and plot against each other:

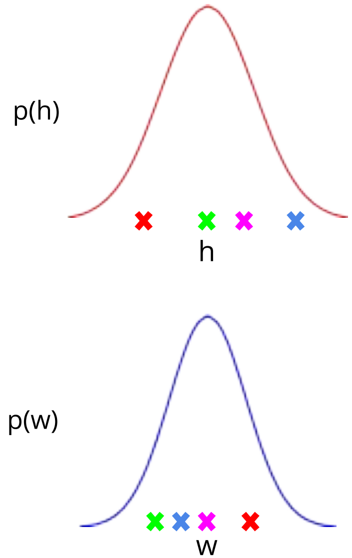




Sampling Two Dimensional Variables

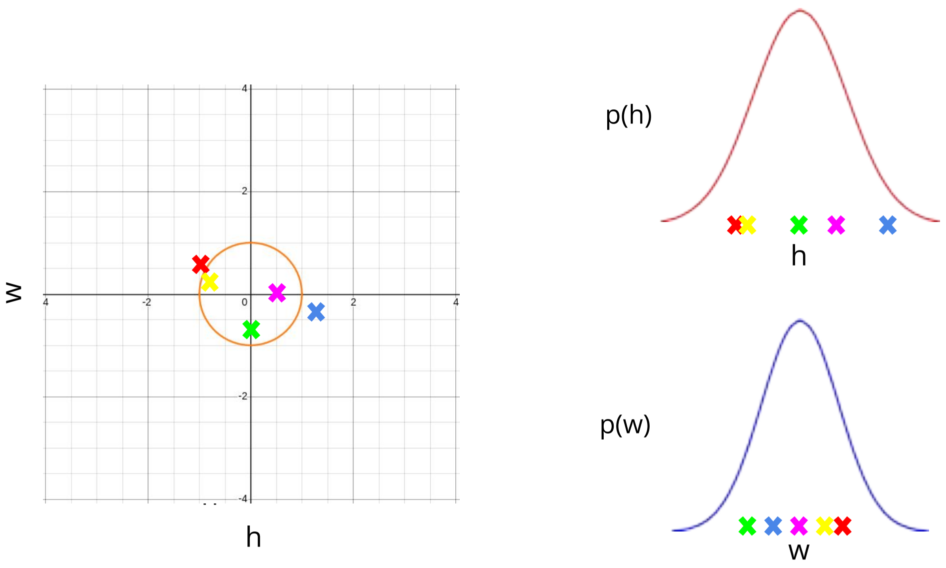
Sample height and weight one after the other and plot against each other:





Sampling Two Dimensional Variables

Sample height and weight one after the other and plot against each other:



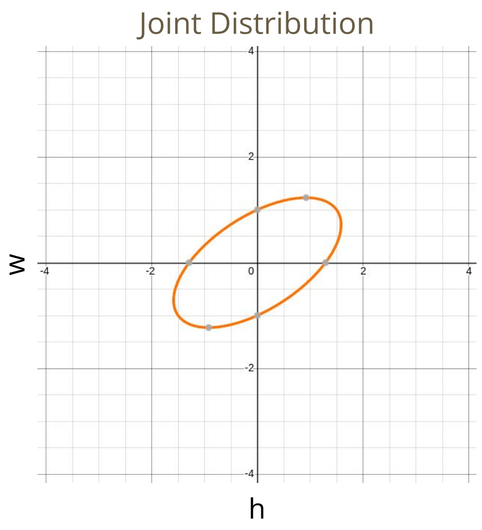
Independence Assumption

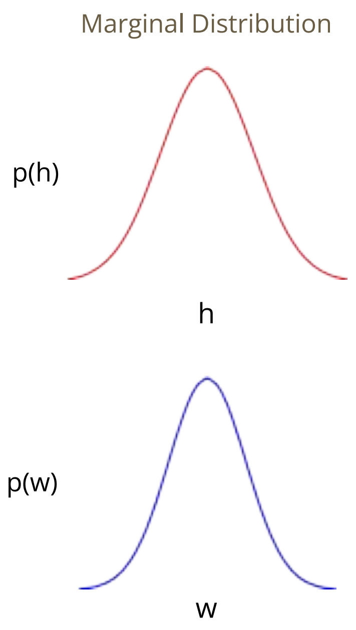
This assumes height and weight are independent.

$$p(h, w) = p(h)p(w)$$

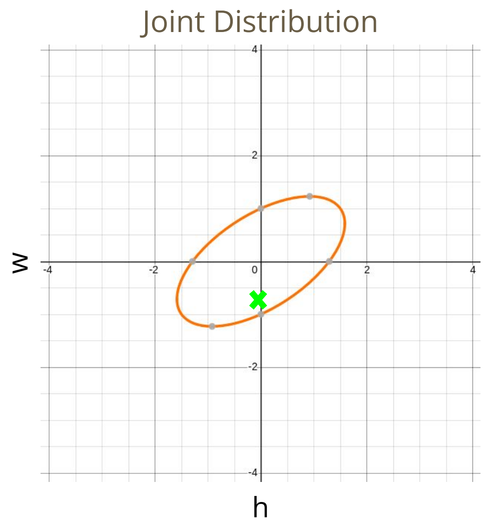
In reality they are dependent (body mass index = $\frac{w}{h^2}$) .

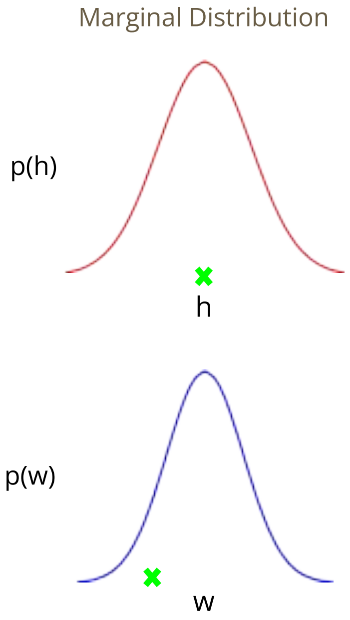
Sampling Two Dimensional Variables



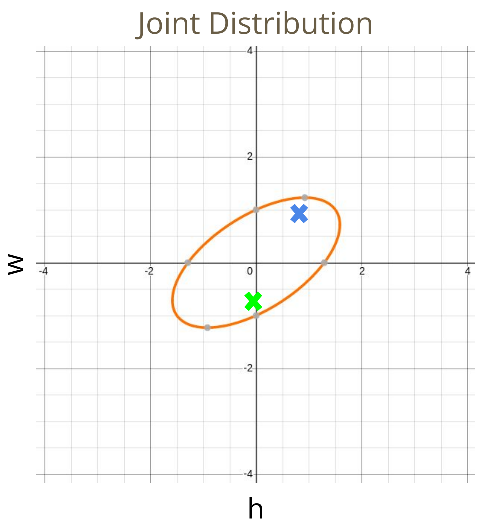


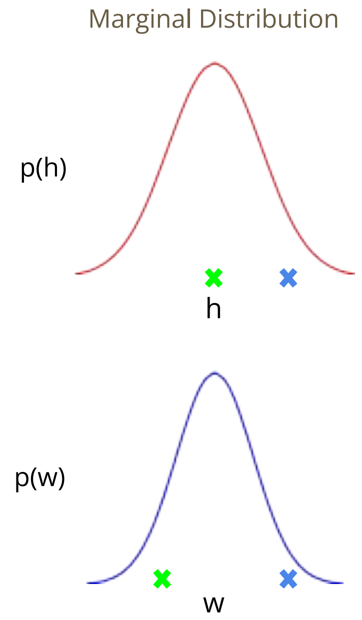
Sampling Two Dimensional Variables



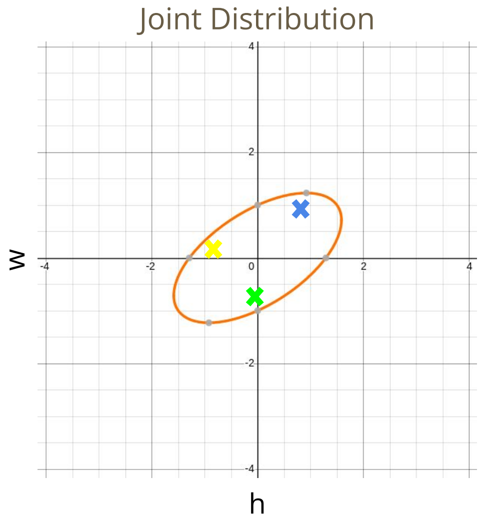


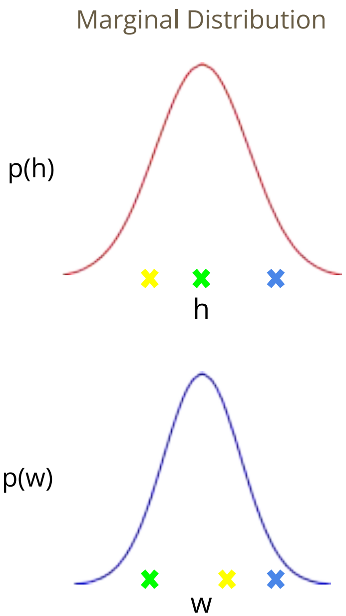
Sampling Two Dimensional Variables



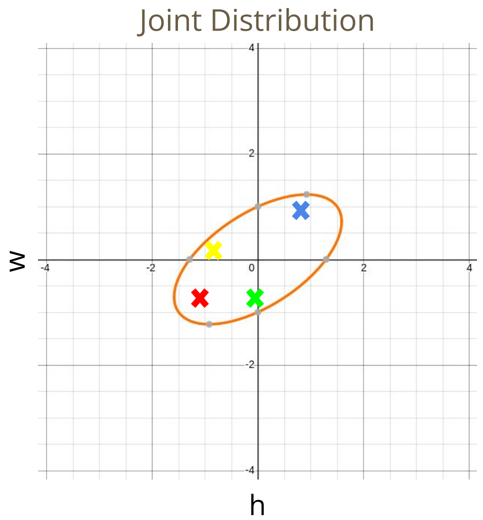


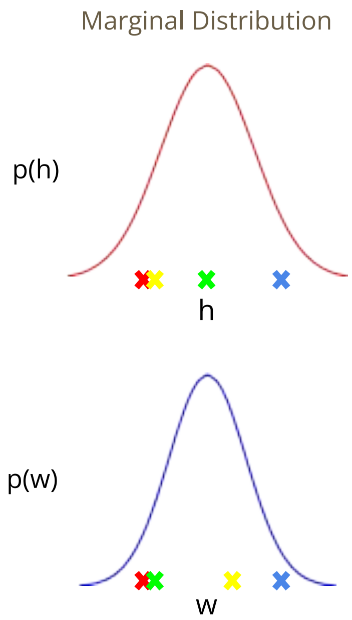
Sampling Two Dimensional Variables



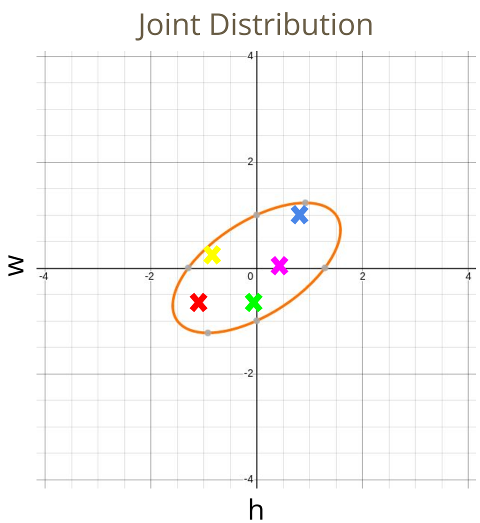


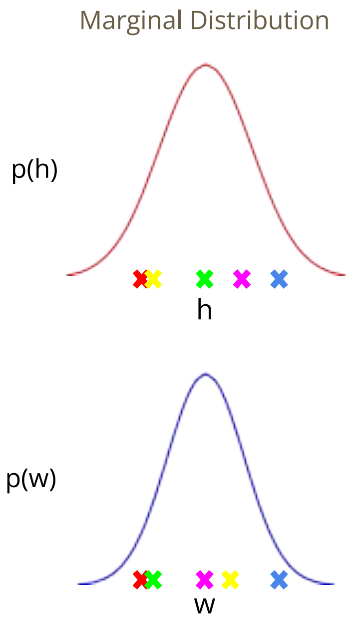
Sampling Two Dimensional Variables





Sampling Two Dimensional Variables





Bivariate Independent Gaussians

$$p(h,w) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} \exp\left(-\frac{1}{2}\begin{pmatrix} h-\mu_1 \\ w-\mu_2 \end{pmatrix}^T \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} h-\mu_1 \\ w-\mu_2 \end{pmatrix}\right)$$

$$p(\mathbf{y}) = \frac{1}{(2\pi)^{k/2}|\mathbf{D}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\mu)^T\mathbf{D}^{-\mathbf{1}}(\mathbf{y}-\mu)\right)$$

Correlated Gaussians

Obtained from original by rotating the data space using matrix **R**.

$$p(\mathbf{y}) = \frac{1}{(2\pi)^{k/2}|\mathbf{C}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\mu)^T \mathbf{C}^{-1}(\mathbf{y}-\mu)\right)$$

This gives a covariance matrix:

$$\mathrm{C} = \mathrm{RDR}^T$$

Recall Univariate Gaussian Properties (slides 10-11)
Multivariate Consequence:

If we have:

$$\mathbf{x} \sim \mathcal{N}(\mu, \boldsymbol{\Sigma})$$

And:

$$\mathbf{y} = \mathbf{W}\mathbf{x}$$

Then:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{W}\mu, \mathbf{W}\boldsymbol{\Sigma}\mathbf{W}^T)$$

Prediction with Correlated Gaussians

Suppose a zero-mean 2-dimensional Gaussian variable:

$$p(x_1, x_2) = \mathcal{N}(\mathbf{0}, \mathbf{K})$$

Prediction of x_2 from x_1 requires conditional density. Conditional density is also Gaussian:

$$p(x_2|x_1) \sim \mathcal{N}\left(\frac{\mathbf{K}_{1,2}}{\mathbf{K}_{1,1}} x_1, \mathbf{K}_{2,2} - \frac{\mathbf{K}_{1,2}^2}{\mathbf{K}_{1,1}}\right)$$

Prediction with Correlated Gaussians

General case (still zero-mean):

$$p(\mathbf{x}, x_*) = \mathcal{N}\left(0, \begin{pmatrix} \mathbf{K}_{\mathbf{x},\mathbf{x}} & \mathbf{k}_{\mathbf{x},*} \\ \mathbf{k}_{*,\mathbf{x}} & k_{*,*} \end{pmatrix}\right)$$

Prediction of x_* from $\mathbf{x}$ requires conditional density. Conditional density is also Gaussian:

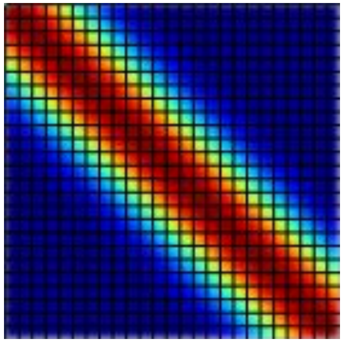
$$p(x_*|\mathbf{x}) = \mathcal{N}\left(\mathbf{k}_{*,\mathbf{x}}\mathbf{K}_{\mathbf{x},\mathbf{x}}^{-1}\mathbf{x}, k_{*,*} - \mathbf{k}_{*,\mathbf{x}}\mathbf{K}_{\mathbf{x},\mathbf{x}}^{-1}\mathbf{k}_{\mathbf{x},*}\right)$$

Outline

1. Motivation
2. The Gaussian Distribution
3. Covariance Functions
4. Gaussian Process
5. Basis Function Representations
6. Constructing Covariance
7. Gaussian Process Limitations
8. Conclusion

Covariance Functions

- Covariance **matrix** is built by getting values from covariance **function**.
- The covariance function is also known as a kernel.
- Covariance functions are building blocks of covariance matrices.



Covariance matrix

Covariance Functions: Example

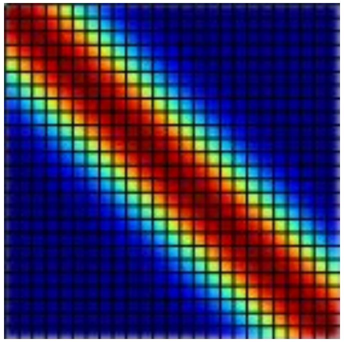
Covariance **matrix** is built by getting values from covariance **function**.

Exponentiated Quadratic Kernel Function:

(also known as RBF, Squared Exponential, Gaussian)

$$k(x_1, x_2) = \alpha \exp\left(-\frac{(x_1 - x_2)^2}{2\gamma^2}\right)$$

$$\longrightarrow M_{i,j} = k(i, j)$$



Covariance matrix

Covariance Functions: Example

$$k(x_1, x_2) = \alpha \exp\left(-\frac{(x_1 - x_2)^2}{2\gamma^2}\right)$$

$$x_1 = -3.0, x_2 = 1.20, x_3 = 1.40$$
$$\gamma = 2.00, \alpha = 1.00$$

$$\begin{bmatrix} 1.0 & & \\ & & \\ & & \end{bmatrix}$$

$$k_{1,1} = k(x_1, x_1)$$

$$= k(-3.0, -3.0) = 1.0 \times \exp\left(-\frac{(-3.0 - -3.0)^2}{2 \times 2.00^2}\right) = 1.0$$

Covariance Functions: Example

$$k(x_1, x_2) = \alpha \exp\left(-\frac{(x_1 - x_2)^2}{2\gamma^2}\right)$$

$$x_1 = -3.0, x_2 = 1.20, x_3 = 1.40$$

$$\gamma = 2.00, \alpha = 1.00$$

$$k_{1,2} = k(x_1, x_2)$$

$$= k(-3.0, 1.2) = 1.0 \times \exp\left(-\frac{(-3.0 - 1.2)^2}{2 \times 2.00^2}\right) = 0.11$$

$$\begin{bmatrix} 1.0 & \boxed{0.11} & \\ & & \\ & & \end{bmatrix}$$

Covariance Functions: Example

$$k(x_1, x_2) = \alpha \exp\left(-\frac{(x_1 - x_2)^2}{2\gamma^2}\right)$$

$$x_1 = -3.0, x_2 = 1.20, x_3 = 1.40$$

$$\gamma = 2.00, \alpha = 1.00$$

$$k_{1,3} = k(x_1, x_3)$$

$$= k(-3.0, 1.4) = 1.0 \times \exp\left(-\frac{(-3.0 - 1.4)^2}{2 \times 2.00^2}\right) = 0.089$$

$$\begin{bmatrix} 1.0 & 0.11 & \boxed{0.089} \\ & & \\ & & \end{bmatrix}$$

Covariance Functions: Example

$$k(x_1, x_2) = \alpha \exp\left(-\frac{(x_1 - x_2)^2}{2\gamma^2}\right)$$

$$x_1 = -3.0, x_2 = 1.20, x_3 = 1.40$$

$$\gamma = 2.00, \alpha = 1.00$$

$$k_{2,1} = k(x_2, x_1)$$

$$= k(1.2, -3.0) = 1.0 \times \exp\left(-\frac{(1.2 - -3.0)^2}{2 \times 2.00^2}\right) = 0.11$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 \\ \boxed{0.11} & & \\ & & \end{bmatrix}$$

Covariance Functions: Example

$$k(x_1, x_2) = \alpha \exp\left(-\frac{(x_1 - x_2)^2}{2\gamma^2}\right)$$

$$x_1 = -3.0, x_2 = 1.20, x_3 = 1.40$$
$$\gamma = 2.00, \alpha = 1.00$$

$$k_{2,2} = k(x_2, x_2)$$
$$= k(1.2, 1.2) = 1.0 \times \exp\left(-\frac{(1.2 - 1.2)^2}{2 \times 2.00^2}\right) = 1.0$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 \\ 0.11 & \boxed{1.0} & \\ & & \end{bmatrix}$$

Covariance Functions: Example

$$k(x_1, x_2) = \alpha \exp\left(-\frac{(x_1 - x_2)^2}{2\gamma^2}\right)$$

$$x_1 = -3.0, x_2 = 1.20, x_3 = 1.40$$

$$\gamma = 2.00, \alpha = 1.00$$

$$k_{2,3} = k(x_2, x_3)$$

$$= k(1.2, 1.4) = 1.0 \times \exp\left(-\frac{(1.2 - 1.4)^2}{2 \times 2.00^2}\right) = 0.995$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 \\ 0.11 & 1.0 & \boxed{0.995} \\ & & \end{bmatrix}$$

Covariance Functions: Example

$$k(x_1, x_2) = \alpha \exp\left(-\frac{(x_1 - x_2)^2}{2\gamma^2}\right)$$

$$x_1 = -3.0, x_2 = 1.20, x_3 = 1.40$$
$$\gamma = 2.00, \alpha = 1.00$$

$$k_{3,1} = k(x_3, x_1)$$
$$= k(1.4, -3.0) = 1.0 \times \exp\left(-\frac{(1.4 - -3.0)^2}{2 \times 2.00^2}\right) = 0.089$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 \\ 0.11 & 1.0 & 0.995 \\ 0.089 & & \end{bmatrix}$$

Covariance Functions: Example

$$k(x_1, x_2) = \alpha \exp\left(-\frac{(x_1 - x_2)^2}{2\gamma^2}\right)$$

$$x_1 = -3.0, x_2 = 1.20, x_3 = 1.40$$

$$\gamma = 2.00, \alpha = 1.00$$

$$k_{3,2} = k(x_3, x_2)$$

$$= k(1.4, 1.2) = 1.0 \times \exp\left(-\frac{(1.4 - 1.2)^2}{2 \times 2.00^2}\right) = 0.995$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 \\ 0.11 & 1.0 & 0.995 \\ 0.089 & \boxed{0.995} & \end{bmatrix}$$

Covariance Functions: Example

$$k(x_1, x_2) = \alpha \exp\left(-\frac{(x_1 - x_2)^2}{2\gamma^2}\right)$$

$$x_1 = -3.0, x_2 = 1.20, x_3 = 1.40$$
$$\gamma = 2.00, \alpha = 1.00$$

$$k_{3,3} = k(x_3, x_3)$$

$$= k(1.4, 1.4) = 1.0 \times \exp\left(-\frac{(1.4 - 1.4)^2}{2 \times 2.00^2}\right) = 1.0$$

$$\begin{bmatrix} 1.0 & 0.11 & 0.089 \\ 0.11 & 1.0 & 0.995 \\ 0.089 & 0.995 & \boxed{1.0} \end{bmatrix}$$

Outline

1. Motivation
2. The Gaussian Distribution
3. Covariance Functions
4. Gaussian Process
5. Basis Function Representations
6. Constructing Covariance
7. Gaussian Process Limitations
8. Conclusion

Gaussian Process

Stochastic process X indexed on some space $\mathcal{X}$ is called a Gaussian process (GP) with mean function μ and covariance function k if for every finite subset of $\mathcal{X}$ such as $\{x_1, x_2, ..., x_n\}$ the joint distribution of X on this subset is a multivariate Gaussian variable with mean μ and covariance generated from k :

$$P(X(x_1), ..., X(x_n)) = \mathcal{N}\left(\begin{pmatrix} \mu(x_1) \\ \mu(x_2) \\ . \\ . \\ \mu(x_n) \end{pmatrix}, \begin{pmatrix} k(x_1, x_1) & k(x_1, x_2) & ... & k(x_1, x_n) \\ k(x_2, x_1) & ... & & k(x_2, x_n) \\ . & . & . & . \\ . & . & . & . \\ k(x_n, x_1) & ... & & k(x_n, x_n) \end{pmatrix} \right)$$

Outline

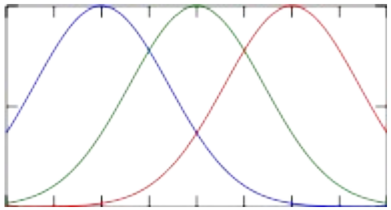
1. Motivation
2. The Gaussian Distribution
3. Covariance Functions
4. Gaussian Process
5. Basis Function Representations
6. Constructing Covariance
7. Gaussian Process Limitations
8. Conclusion

Basis Function Form

Radial basis functions commonly have the form:

$$\varphi_k(\mathbf{x}_i) = \exp\left(-\frac{|\mathbf{x} - \mu_k|^2}{2\gamma^2}\right)$$

Basis function maps data into a **feature space** in which a linear sum is a nonlinear function.



A set of radial basis functions

Basis Function Representations

Represent a function by a linear sum over a basis.

$$f(\mathbf{x}_i; \mathbf{w}) = \sum_{k=1}^{m} w_k \varphi_k(\mathbf{x}_i)$$

Where $\varphi_k(\cdot)$ are basis functions.

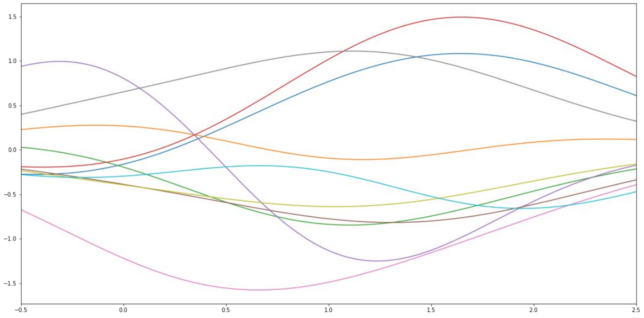
Random Functions

Functions derived using:

$$f(\mathbf{x}) = \sum_{k=1}^{m} w_k \varphi_k(\mathbf{x})$$

where **w** is sampled from a Gaussian density:

$$w_i \sim \mathcal{N}(0, \alpha)$$



Each line is a separate sample, generated by a weighted sum of the basis set. The weights, w are sampled from a Gaussian density with variance 1.

Direct Construction of Covariance Matrix

Using matrix notation to write function:

$$f(\mathbf{x}_i; \mathbf{w}) = \sum_{k=1}^{m} w_k \varphi_k(\mathbf{x}_i)$$

computed at training data gives a vector:

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

- w and f are only related by a inner product.
- Φ is fixed and non-stochastic for a given training set.
- f is Gaussian distributed.
- It is straightforward to compute distribution for f.

Infinite Feature Space

- A RBF model with infinite basis functions is a Gaussian process. The covariance function is the exponentiated quadratic.
- Note: The functional form for the covariance function and basis functions are similar.
  - This is a special case,
  - In general they are very different

Nonparametric Gaussian Processes

- Gaussian processes are generally non-parametric: combine data with covariance function to get model.
- This representation cannot be summarized by a parameter vector of a fixed size.

The Parametric Bottleneck

Parametric models have a representation that does not respond to increasing training set size.

Bayesian posterior distributions over parameters contain the information about the training data.

1. Use Bayes' rule from training data to estimate parameters: p (w|y, X),
2. Make predictions on test data using estimated parameters:

$$p(y_*|X_*, y, X) = \int p(y_*|w, X_*)p(w|y, X)dw$$

The Parametric Bottleneck

$$p(y_*|X_*, y, X) = \int p(y_*|w, X_*)p(w|y, X)dw$$

w becomes a bottleneck for information about the training set to pass to the test set.

Solution: increase m (dimension of w) so that the bottleneck is so large that it no longer presents a problem.

How big is big enough for m? Non-parametric says:

$$m \to \infty$$

The Parametric Bottleneck: Nonparametric Solution

Now no longer possible to manipulate the model through the standard parametric form.

However, it is possible to express *parametric* as GPs:

$$k(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$$

These are known as degenerate covariance matrices.

No matter how big training data is, their rank is bounded by m. Instead, non-parametric models have full rank covariance matrices.

Most well known is the **linear kernel**:

$$k(x_i, x_j) = x_i^T x_j$$

Making Predictions

- For non-parametrics prediction at new points x_* is made by conditioning on x in the joint distribution.
- In GPs this involves combining the training data with the covariance function and the mean function.
- Parametric is a special case when conditional prediction can be summarized in a fixed number of parameters.
- Complexity of parametric model remains fixed regardless of the size of our training data set.
- For a non-parametric model the required number of parameters grows with the size of the training data.

Outline

1. Motivation
2. The Gaussian Distribution
3. Covariance Functions
4. Gaussian Process
5. Basis Function Representations
6. Constructing Covariance
7. Gaussian Process Limitations
8. Conclusion

Constructing Covariance Functions

Sum of two covariances is also a covariance function.

$$k(x, x') = k_1(x, x') + k_2(x, x')$$

Constructing Covariance Functions

Product of two covariances is also a covariance function:

$$k(x, x') = k_1(x, x') \cdot k_2(x, x')$$

If f(x) is a Gaussian process, and g (x) is a deterministic function and:

$$h(x) = f(x)g(x)$$

Then:

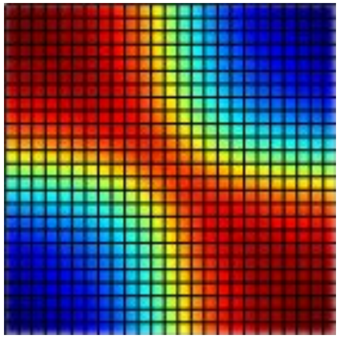
$$k_h(x, x') = g(x)k_f(x, x')g(x')$$

Covariance Functions

MLP covariance function:

$$k(\mathbf{x}, \mathbf{x}') = \alpha \arcsin\left(\frac{w\mathbf{x}^T\mathbf{x}' + b}{\sqrt{(w\mathbf{x}^T\mathbf{x} + b + 1)(w\mathbf{x}'^T\mathbf{x}' + b + 1)}}\right)$$

Based on the infinite neural network model.



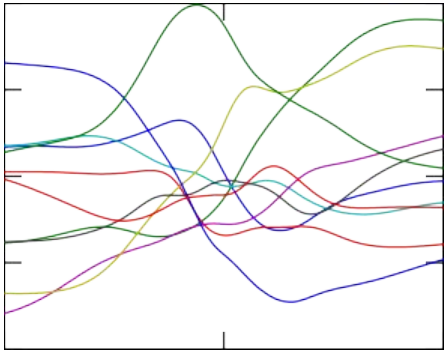
MLP covariance matrix

Covariance Functions

MLP covariance function:

$$k(\mathbf{x}, \mathbf{x}') = \alpha \arcsin\left(\frac{w\mathbf{x}^T\mathbf{x}' + b}{\sqrt{(w\mathbf{x}^T\mathbf{x} + b + 1)(w\mathbf{x}'^T\mathbf{x}' + b + 1)}}\right)$$

Based on infinite neural network model:



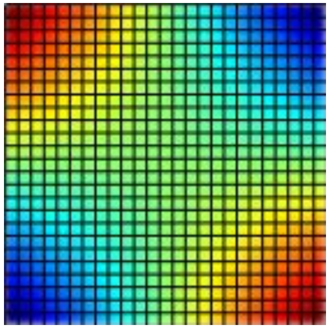
Samples of GP generated by MLP covariance

Covariance Functions

Linear covariance function:

$$k(\mathbf{x}, \mathbf{x}') = \alpha \mathbf{x}^T \mathbf{x}'$$

Bayesian linear regression.



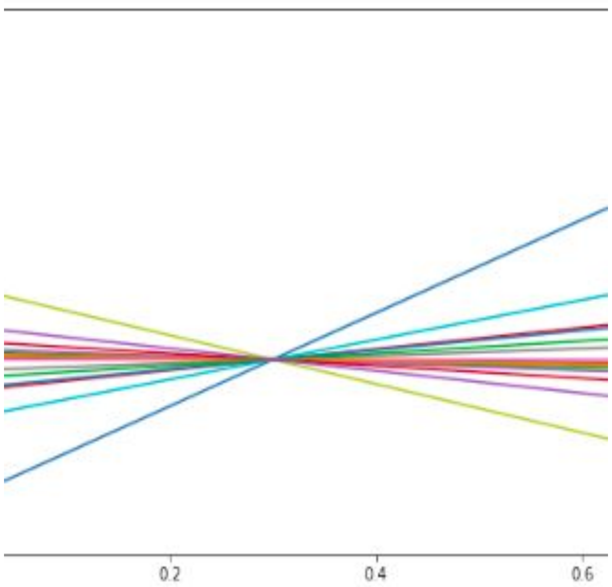
Linear covariance matrix

Covariance Functions

Linear covariance function:

$$k(\mathbf{x}, \mathbf{x}') = \alpha \mathbf{x}^T \mathbf{x}'$$

Bayesian linear regression.



Samples of GP generated by Linear covariance

Gaussian Noise

Gaussian noise model,

$$P(y_i|x_i) = \mathcal{N}(y_i|x_i, \sigma^2)$$

Where σ^2 is the variance of the noise.

Equivalent to a covariance function of the form

$$k(x_i, x_j) = \delta_{i,j}\sigma^2$$

where $\delta_{i,j}$ is the Kronecker delta function.

Additive nature of Gaussians means we can simply add this term to existing covariance matrices.

Outline

1. Motivation
2. The Gaussian Distribution
3. Covariance Functions
4. Gaussian Process
5. Basis Function Representations
6. Constructing Covariance
7. Gaussian Process Limitations
8. Conclusion

Limitations of Gaussian Processes

- Inference is $O(N^3)$ due to matrix inverse (in practice use Cholesky).
- Gaussian processes don't deal well with discontinuities (financial crises, phosphorylation, collisions, edges in images).
- Widely used exponentiated quadratic covariance (RBF) can be too smooth in practice (but there are many alternatives.)

Outline

1. Motivation
2. The Gaussian Distribution
3. Covariance Functions
4. Gaussian Process
5. Basis Function Representations
6. Constructing Covariance
7. Gaussian Process Limitations
8. Conclusion

Summary

- Broad introduction to Gaussian processes.

  Started with Gaussian distribution.

  Motivated Gaussian processes through the multivariate density.
- Emphasized the role of the covariance (not the mean). Performs nonlinear regression with error bars.
- Parameters of the covariance function (kernel) are easily optimized with maximum likelihood.
- **Demos:**
- https://edward-rees.com/gp
- http://chifeng.scripts.mit.edu/stuff/gp-demo/

References

G. Della Gatta, M. Bansal, A. Ambesi-Impiombato, D. Antonini, C. Missero, and D. di Bernardo. Direct targets of the trp63 transcription factor revealed by a combination of gene expression profiling and reverse engineering. Genome Research, 18(6): 939–948, Jun 2008.

A. A. Kalaitzis and N. D. Lawrence. A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression. BMC Bioinformatics, 12(180), 2011.

R. M. Neal. Bayesian Learning for Neural Networks. Springer, 1996. Lecture Notes in Statistics 118.

J. Oakley and A. O'Hagan. Bayesian inference for the uncertainty distribution of computer model outputs. Biometrika, 89(4): 769–784, 2002.

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

C. K. I. Williams. Computation with infinite neural networks. Neural Computation, 10(5):1203–1216, 1998.

Next Week:

Point Estimation

Have a good day!