Probabilistic Modelling and Reasoning: A Machine Learning Approach

Introduction to Probabilistic Modelling

Edwin V. Bonilla

Principal Research Scientist, CSIRO's Data61
Associate Professor (Hon.), Australian National University

December 14th, 2021

Suggested Readings

Machine Learning: A Probabilistic Perspective
Kevin P. Murphy, 2012

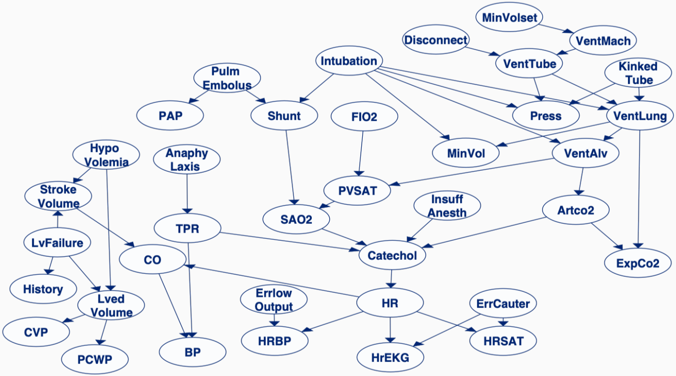
Bayesian Reasoning and Machine Learning
David Barber, 2012

Pattern Recognition and Machine Learning
Christopher Bishop, 2006

Gaussian Processes for Machine Learning
Carl E. Rasmussen and Christopher K. I. Williams, 2006

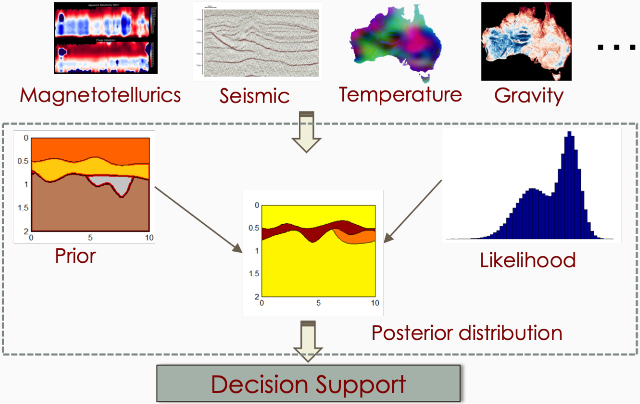
Motivation (1)

- Medical diagnosis in an intensive care unit



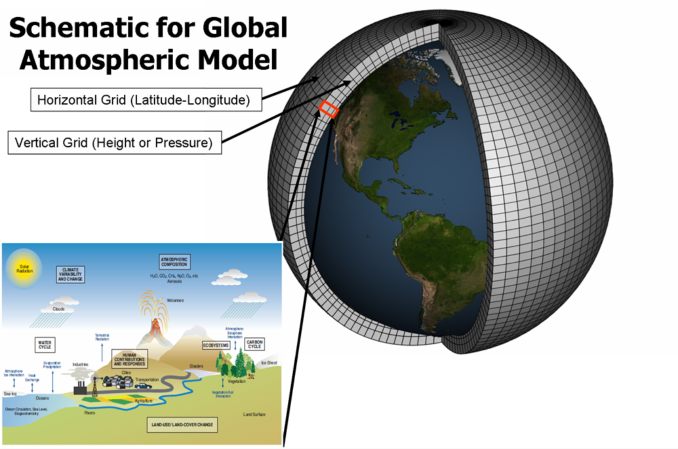
Motivation (2)

- Data fusion for geothermal energy exploration



Motivation (3)

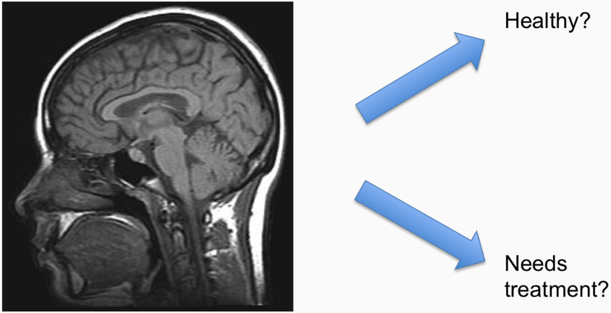
- Quantification of Uncertainty with Expensive Computational Models: Climate modelling



Kennedy and O'Hagan, *RSS-B*, 2001

Motivation (4)

- Quantification of Uncertainty with No Models: Classification and progression modelling of neurodegenerative diseases



Filippone et al., *AoAS*, 2012

A Unified Framework

A model might be expensive to simulate/inaccurate

- Emulate model/discrepancy using a surrogate

A Unified Framework

A model might be expensive to simulate/inaccurate

- Emulate model/discrepancy using a surrogate

A model might not even be available

- Make use of a flexible model, e.g., Neural Nets

A Unified Framework

A model might be expensive to simulate/inaccurate

- Emulate model/discrepancy using a surrogate

A model might not even be available

- Make use of a flexible model, e.g., Neural Nets

Quantification of Uncertainty

- Bayesian neural nets
- Gaussian Processes

Three Lectures: Outline

1. Introduction to probabilistic modelling
   - Machine Learning and Probability Theory
   - Bayesian Linear Regression
2. Gaussian Processes
   - Gaussian Processes for Regression
   - Model Approximations
3. Advanced Topics
   - Approximate Inference
   - Applications, Challenges & Opportunities

This Lecture: Outline

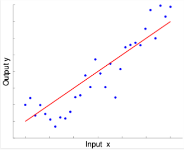
1. Basic Machine Learning Concepts
2. Probability Theory Refresher
3. Bayesian Linear Regression

Basic Machine Learning Concepts

Basic Machine Learning Concepts (1)

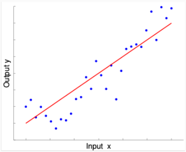
Types of machine learning

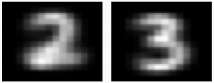
- Supervised
  - Classification
  - Regression

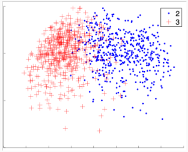


Types of machine learning

- Supervised
  - Classification
  - Regression
- Unsupervised
  - Dimensionality reduction
  - Clustering
  - Latent variable modelling
  - Density estimation



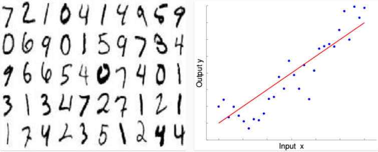


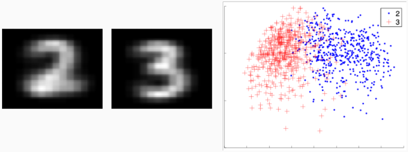


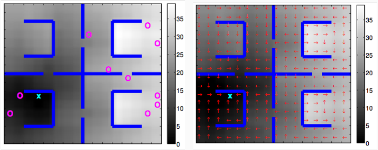
Basic Machine Learning Concepts (1)

Types of machine learning

- Supervised
  - Classification
  - Regression
- Unsupervised
  - Dimensionality reduction
  - Clustering
  - Latent variable modelling
  - Density estimation
- Reinforcement learning
  - Delayed reward
  - Acting and planning

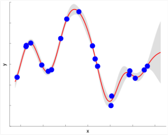






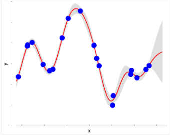
Basic Machine Learning Concepts (2)

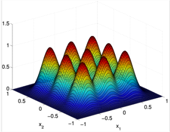
- The need for probabilistic predictions
  - Risk assessment, decision theory
  - Active learning
  - Reinforcement learning



Basic Machine Learning Concepts (2)

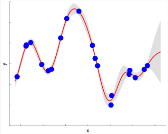
- The need for probabilistic predictions
  - Risk assessment, decision theory
  - Active learning
  - Reinforcement learning



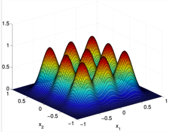


- The curse of dimensionality

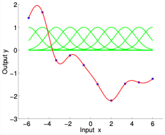
- The need for probabilistic predictions
  - Risk assessment, decision theory
  - Active learning
  - Reinforcement learning

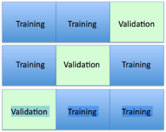


- The curse of dimensionality



- Generalisation
  - Overfitting, model selection
  - Validation set, cross validation
  - No free lunch theorem





Probability Theory Refresher

Discrete Random Variables

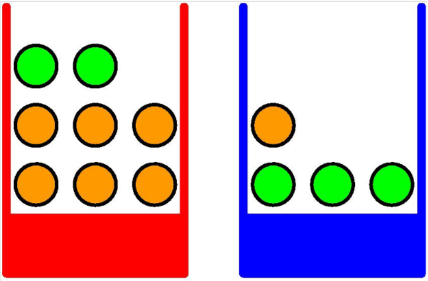
- $X \in \mathcal{X}$: Random variable (r.v.) X can take on any value from $\mathcal{X}$
- $p(X = x)$ or simply $p(x)$: Probability that $X = x$
- Probability mass function (pmf):

$$0 \leq p(x) \leq 1, \sum_{x \in \mathcal{X}} p(x) = 1$$

Discrete Random Variables

- $X \in \mathcal{X}$: Random variable (r.v.) X can take on any value from $\mathcal{X}$
- $p(X = x)$ or simply $p(x)$: Probability that $X = x$
- Probability mass function (pmf):

$$0 \leq p(x) \leq 1, \sum_{x \in \mathcal{X}} p(x) = 1$$

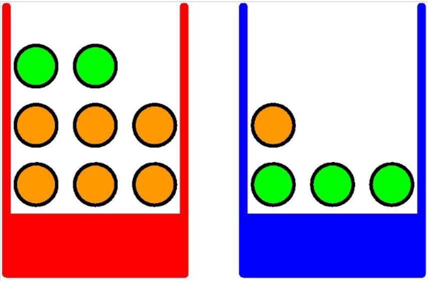


- $B \in \{r, b\}$: r.v. for the box taking on values red or blue
- $F \in \{a, o\}$: r.v. for the fruit taking on values apple or orange

Discrete Random Variables

- $X \in \mathcal{X}$: Random variable (r.v.) X can take on any value from $\mathcal{X}$
- $p(X = x)$ or simply $p(x)$: Probability that $X = x$
- Probability mass function (pmf):

$$0 \leq p(x) \leq 1, \sum_{x \in \mathcal{X}} p(x) = 1$$

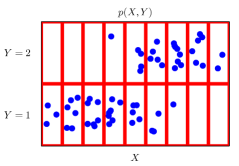


- $B \in \{r, b\}$: r.v. for the box taking on values red or blue
- $F \in \{a, o\}$: r.v. for the fruit taking on values apple or orange

We can specify a joint distribution $p(B, F) = P(B)P(F|B)$

The Rules of Probability and Terminology

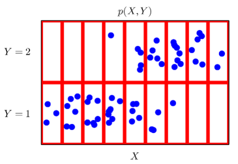
- Joint $p(X = x, Y = x)$

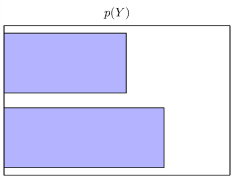


The Rules of Probability and Terminology

- Joint $p(X = x, Y = x)$
- Marginal (using the sum rule):

$$p(Y = y) = \sum_{x \in \mathcal{X}} p(X = x, Y = y)$$





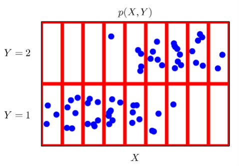
The Rules of Probability and Terminology

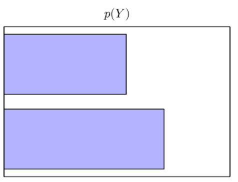
- Joint $p(X = x, Y = x)$
- Marginal (using the sum rule):

$$p(Y = y) = \sum_{x \in \mathcal{X}} p(X = x, Y = y)$$

- Product rule:

$$\begin{aligned} p(X, Y) &= p(Y)p(X \mid Y) \\ &= p(X)p(Y \mid X) \end{aligned}$$





The Rules of Probability and Terminology

- Joint $p(X = x, Y = x)$
- Marginal (using the sum rule):

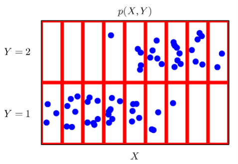
$$p(Y = y) = \sum_{x \in \mathcal{X}} p(X = x, Y = y)$$

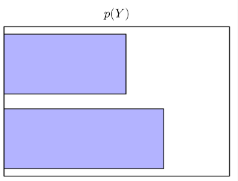
- Product rule:

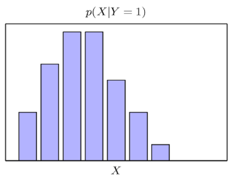
$$\begin{aligned} p(X, Y) &= p(Y)p(X \mid Y) \\ &= p(X)p(Y \mid X) \end{aligned}$$

- Conditional:

$$p(x = x \mid Y = y) = \frac{p(X = x, Y = Y)}{p(Y = y)}$$







How to Update our Beliefs Given New Data

Bayesian Inference

Bayesian inference provides us with a a mathematical framework explaining how to change our (prior) beliefs in the light of new evidence.

$$\overbrace{p(X = x \mid Y = y)}^{\text{posterior}} = \frac{\overbrace{p(X = x)}^{\text{prior}}\overbrace{p(Y = y \mid X = x)}^{\text{likelihood}}}{\underbrace{p(Y = y)}_{\text{evidence:}\, p(Y=y)=\sum_{x'} p(X=x')p(Y=y \mid X=x')}}$$

How to Update our Beliefs Given New Data

Bayesian Inference

Bayesian inference provides us with a a mathematical framework explaining how to change our (prior) beliefs in the light of new evidence.

$$\overbrace{p(X = x \mid Y = y)}^{\text{posterior}} = \frac{\overbrace{p(X = x)}^{\text{prior}}\overbrace{p(Y = y \mid X = x)}^{\text{likelihood}}}{\underbrace{p(Y = y)}_{\text{evidence:}p(Y=y)=\sum_{x'} p(X=x')p(Y=y \mid X=x')}}$$

Example: Suppose you have been tested positive for a disease; what is the probability that you actually have the disease?

- $X \in \{0, 1\}$: Whether you have the disease
- $Y \in \{0, 1\}$: Outcome of the test

How to Update our Beliefs Given New Data

Bayesian Inference

Bayesian inference provides us with a a mathematical framework explaining how to change our (prior) beliefs in the light of new evidence.

$$\overbrace{p(X = x \mid Y = y)}^{\text{posterior}} = \frac{\overbrace{p(X = x)}^{\text{prior}}\overbrace{p(Y = y \mid X = x)}^{\text{likelihood}}}{\underbrace{p(Y = y)}_{\text{evidence:}p(Y=y)=\sum_{x'} p(X=x')p(Y=y \mid X=x')}}$$

Example: Suppose you have been tested positive for a disease; what is the probability that you actually have the disease?

- $X \in \{0, 1\}$: Whether you have the disease
- $Y \in \{0, 1\}$: Outcome of the test

Computational challenges

Statistical Independence

In our fruit-box example, suppose that both boxes (red and blue) contain the same proportion of apples and oranges, say:

$$p(F = a \,|\, B = r) = p(F = a \,|\, B = b) = 0.2$$

$$p(F = o \,|\, B = r) = p(F = o \,|\, B = b) = 0.8$$

The probability of selecting an apple (or an orange) is independent of the box that is chosen.

Statistical Independence

In our fruit-box example, suppose that both boxes (red and blue) contain the same proportion of apples and oranges, say:

$$p(F = a \mid B = r) = p(F = a \mid B = b) = 0.2$$
$$p(F = o \mid B = r) = p(F = o \mid B = b) = 0.8$$

The probability of selecting an apple (or an orange) is independent of the box that is chosen.

Independent Variables

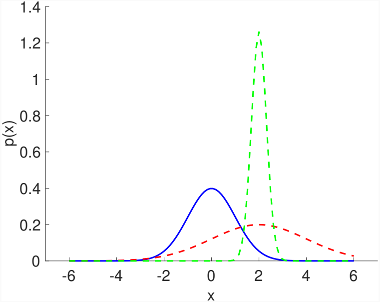
Two variables X and Y are statistically independent iff their joint distribution factorises into the product of their marginals:

$$X \perp\!\!\!\perp Y \leftrightarrow p(X, Y) = P(X)p(Y)$$

This definition generalises to more than two variables.

Continuous Random Variables

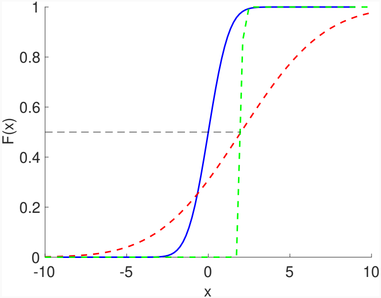
Probability density function (pdf) $p(x)$:



$$p(x) \geq 0, \int_{-\infty}^{\infty} p(x)dx = 1$$

$$p(a < x < b) = \int_{a}^{b} p(x)dx$$

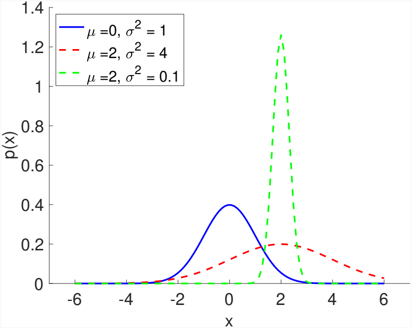
Cumulative distribution function (cdf) $F(x)$:



$$F(x) = p(X \leq x) = \int_{-\infty}^{x} p(z)dz$$

The Gaussian Distribution: 1-dimensional Case

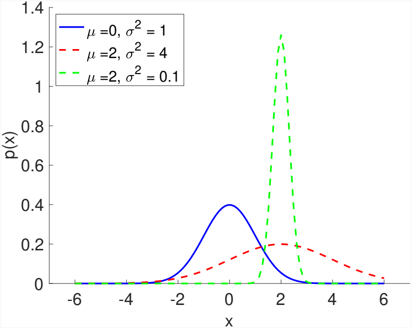
$$p(x) = \mathcal{N}(x; \mu, \sigma^2)$$

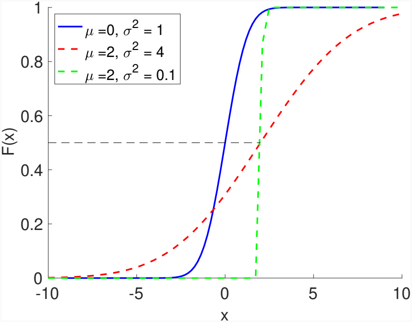


$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

The Gaussian Distribution: 1-dimensional Case

$p(x) = \mathcal{N}(x; \mu, \sigma^2)$



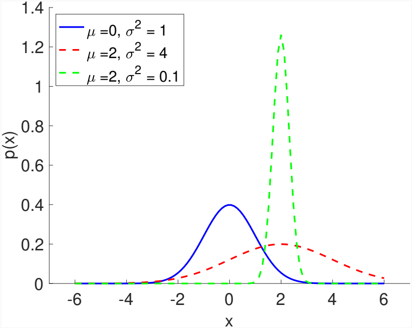


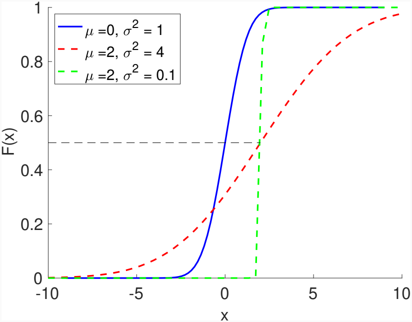
$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$

$F(x) = \int_{-\infty}^{x} \mathcal{N}(z; \mu, \sigma^2) dz$

The Gaussian Distribution: 1-dimensional Case

$p(x) = \mathcal{N}(x; \mu, \sigma^2)$



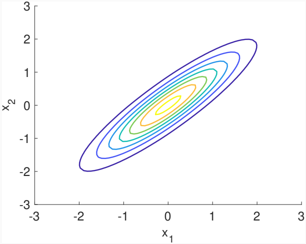


$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right)$

$F(x) = \int_{-\infty}^{x} \mathcal{N}(z; \mu, \sigma^2) dz$

For a standard Normal, $\mu = 0, \sigma^2 = 1$

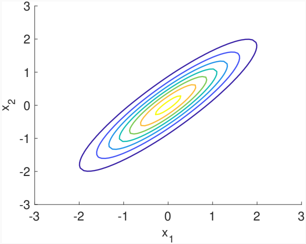
The Gaussian Distribution: 2-dimensional Case



$$p(x_1, x_2) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$$

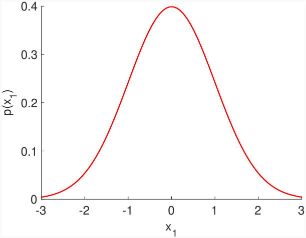
Joint

The Gaussian Distribution: 2-dimensional Case



$p(x_1, x_2) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$

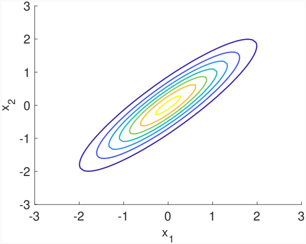
Joint



$p(x_1)$

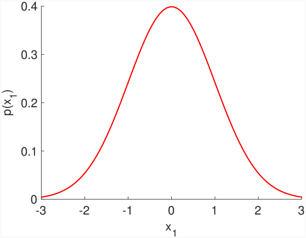
Marginal

The Gaussian Distribution: 2-dimensional Case



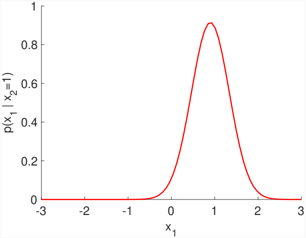
$p(x_1, x_2) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$

Joint



$p(x_1)$

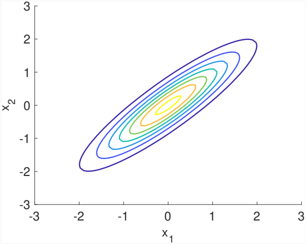
Marginal



$p(x_1|x_2)$

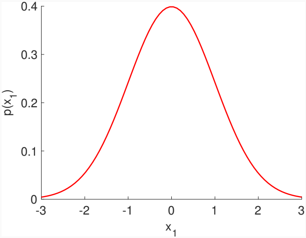
Conditional

The Gaussian Distribution: 2-dimensional Case



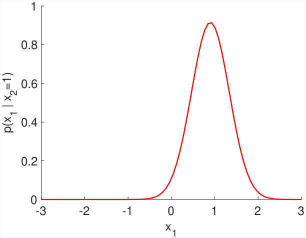
$p(x_1, x_2) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$

Joint



$p(x_1)$

Marginal



$p(x_1|x_2)$

Conditional

The marginal and the conditional distributions are also Gaussians

In general:

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

- $\boldsymbol{\Sigma}^{-1}$: precision matrix

The Gaussian Distribution

In general:

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

- Σ^{-1}: precision matrix
  - $\Sigma_{ij}^{-1} = 0$: x_i, x_j are *conditionally independent* given all the others

The Gaussian Distribution

In general:

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

- Σ^{-1}: precision matrix
  - $\Sigma_{ij}^{-1} = 0$: x_i, x_j are *conditionally independent* given all the others
- Σ: covariance matrix

The Gaussian Distribution

In general:

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

- Σ^{-1}: precision matrix
  - $\Sigma_{ij}^{-1} = 0$: x_i, x_j are *conditionally independent* given all the others
- Σ: covariance matrix
  - $\Sigma_{ij} = 0$: x_i, x_j are *marginally independent*

The Gaussian Distribution

In general:

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

- $\boldsymbol{\Sigma}^{-1}$: precision matrix
  - $\Sigma_{ij}^{-1} = 0$: x_i, x_j are *conditionally independent* given all the others
- $\boldsymbol{\Sigma}$: covariance matrix
  - $\Sigma_{ij} = 0$: x_i, x_j are *marginally independent*
- Marginalizing out a variable

The Gaussian Distribution

In general:

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

- $\boldsymbol{\Sigma}^{-1}$: precision matrix
  - $\Sigma_{ij}^{-1} = 0$: x_i, x_j are *conditionally independent* given all the others
- $\boldsymbol{\Sigma}$: covariance matrix
  - $\Sigma_{ij} = 0$: x_i, x_j are *marginally independent*
- Marginalizing out a variable
  - Leaves $\boldsymbol{\Sigma}$ unchanged but changes $\boldsymbol{\Sigma}^{-1}$

The Gaussian Distribution

In general:

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

- $\boldsymbol{\Sigma}^{-1}$: precision matrix
  - $\Sigma_{ij}^{-1} = 0$: x_i, x_j are *conditionally independent* given all the others
- $\boldsymbol{\Sigma}$: covariance matrix
  - $\Sigma_{ij} = 0$: x_i, x_j are *marginally independent*
- Marginalizing out a variable
  - Leaves $\boldsymbol{\Sigma}$ unchanged but changes $\boldsymbol{\Sigma}^{-1}$
  - This is crucial when parameterizing a Gaussian process

The Rules of Probability: Continuous Case

Consider two continuous random variables x and y with $p(x, y)$

- Sum rule:

$$p(x) = \int p(x, y) dy$$

- Product rule:

$$p(x, y) = p(y)p(x \mid y) = p(x)p(y \mid x)$$

The Rules of Probability: Continuous Case

Consider two continuous random variables x and y with $p(x, y)$

- Sum rule:

$$p(x) = \int p(x, y) dy$$

- Product rule:

$$p(x, y) = p(y)p(x \mid y) = p(x)p(y \mid x)$$

- Bayes' rule:

$$p(x \mid y) = \frac{p(x)p(y \mid x)}{p(y)}$$

Expectation, Variance and Quantiles

- Expectation: $\mathbb{E}[X] \stackrel{\text{def}}{=} \int_{x \in \mathcal{X}} x p(x) dx$.

Expectation, Variance and Quantiles

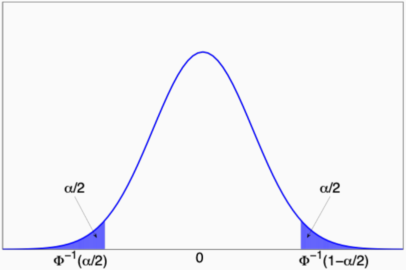
- Expectation: $\mathbb{E}[X] \stackrel{\text{def}}{=} \int_{x \in \mathcal{X}} x p(x) dx.$
- More generally, $\mathbb{E}_{p(X)}[g(X)] \stackrel{\text{def}}{=} \int_{x \in \mathcal{X}} g(x) p(x) dx$

Expectation, Variance and Quantiles

- Expectation: $\mathbb{E}[X] \stackrel{\text{def}}{=} \int_{x \in \mathcal{X}} x p(x) dx$.
- More generally, $\mathbb{E}_{p(X)}[g(X)] \stackrel{\text{def}}{=} \int_{x \in \mathcal{X}} g(x) p(x) dx$
- Variance: $\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$

Expectation, Variance and Quantiles

- Expectation: $\mathbb{E}[X] \stackrel{\text{def}}{=} \int_{x \in \mathcal{X}} x p(x) dx$.
- More generally, $\mathbb{E}_{p(X)}[g(X)] \stackrel{\text{def}}{=} \int_{x \in \mathcal{X}} g(x) p(x) dx$
- Variance: $\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$
- α-quantile: $x_\alpha = F^{-1}(\alpha)$ such that $p(X \leq x_\alpha) = \alpha$

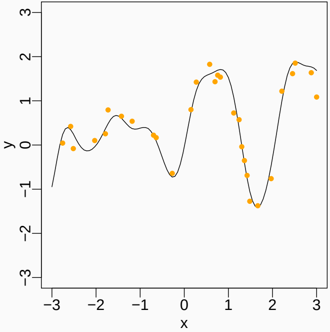


- For a $\mathcal{N}(\mu, \sigma^2)$:
  - 95% interval: $(\mu - 1.96\sigma, \mu + 1.96\sigma)$

Bayesian Linear Regression

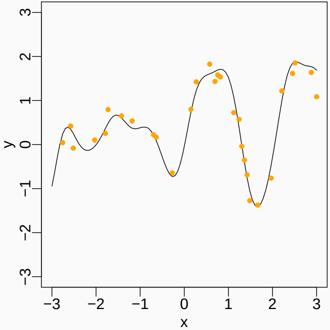
Learning from Data: Function Estimation

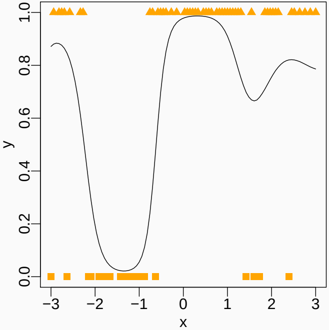
- Take these two examples



Learning from Data: Function Estimation

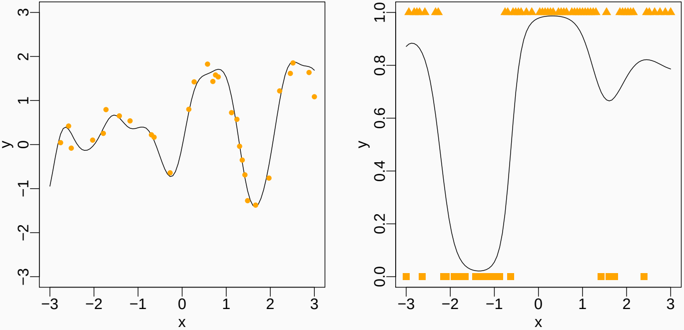
- Take these two examples





Learning from Data: Function Estimation

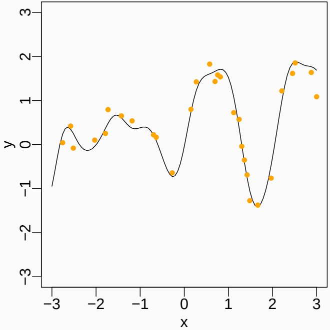
- Take these two examples

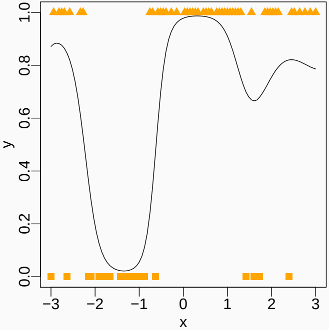


- We are interested in estimating a function $f(x)$ from data

Learning from Data: Function Estimation

- Take these two examples

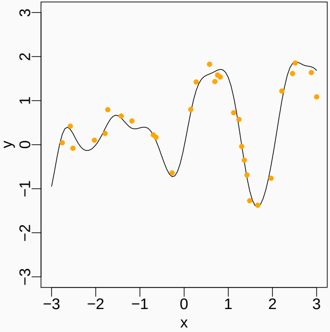


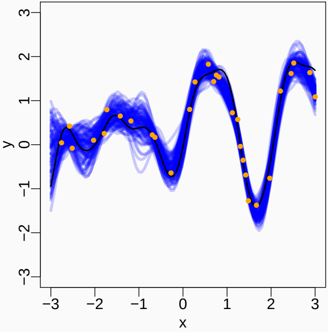


- We are interested in estimating a function $f(x)$ from data
- Most problems in Machine Learning can be cast this way!

What do Bayesian Models Have to Offer?

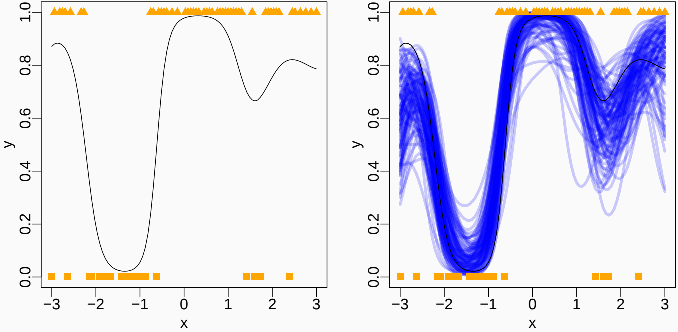
- Regression example





What do Bayesian Models Have to Offer?

- Classification example



Linear-in–the-Parameters Models: Problem Formulation

- Data: $\mathcal{D} = \{\mathbf{x}^{(n)}, y^{(n)}\}_{n=1}^{N}$, $\mathbf{x}^{(n)} \in \mathbb{R}^{D_x}$, $y^{(n)} \in \mathbb{R}$

Linear-in–the-Parameters Models: Problem Formulation

- Data: $\mathcal{D} = \{\mathsf{x}^{(n)}, y^{(n)}\}_{n=1}^{N}$, $\mathsf{x}^{(n)} \in \mathbb{R}^{D_x}$, $y^{(n)} \in \mathbb{R}$
- Inputs : $\mathsf{X} = (\mathsf{x}^{(1)}, \ldots, \mathsf{x}^{(N)})^{\top}$

Linear-in–the-Parameters Models: Problem Formulation

- Data: $\mathcal{D} = \{\mathbf{x}^{(n)}, y^{(n)}\}_{n=1}^{N}$, $\mathbf{x}^{(n)} \in \mathbb{R}^{D_x}$, $y^{(n)} \in \mathbb{R}$
- $\text{Inputs} : \mathbf{X} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)})^\top$
- $\text{Labels} : \mathbf{y} = (y^{(1)}, \ldots, y^{(N)})^\top$

Linear-in–the-Parameters Models: Problem Formulation

- Data: $\mathcal{D} = \{\mathbf{x}^{(n)}, y^{(n)}\}_{n=1}^{N}$, $\mathbf{x}^{(n)} \in \mathbb{R}^{D_x}$, $y^{(n)} \in \mathbb{R}$
- Inputs : $\mathbf{X} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)})^\top$
- Labels : $\mathbf{y} = (y^{(1)}, \ldots, y^{(N)})^\top$
- Goal: : $\mathbf{x} \stackrel{f(\mathbf{x})}{\to} y$

Linear-in–the-Parameters Models: Problem Formulation

- Data: $\mathcal{D} = \{\mathbf{x}^{(n)}, y^{(n)}\}_{n=1}^{N}$, $\mathbf{x}^{(n)} \in \mathbb{R}^{D_x}$, $y^{(n)} \in \mathbb{R}$
- Inputs : $\mathbf{X} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)})^\top$
- Labels : $\mathbf{y} = (y^{(1)}, \ldots, y^{(N)})^\top$
- Goal: : $\mathbf{x} \overset{f(\mathbf{x})}{\to} y$
- Implement a linear combination of basis functions

$$f(\mathbf{x}) = \sum_{j=1}^{D} w_j \varphi_j(\mathbf{x}) = \mathbf{w}^\top \boldsymbol{\varphi}(\mathbf{x})$$

with

$$\boldsymbol{\varphi}(\mathbf{x}) = (\varphi_1(\mathbf{x}), \ldots, \varphi_D(\mathbf{x}))^\top$$

Linear-in–the-Parameters Models: Problem Formulation

- Data: $\mathcal{D} = \{\mathbf{x}^{(n)}, y^{(n)}\}_{n=1}^{N}$, $\mathbf{x}^{(n)} \in \mathbb{R}^{D_x}$, $y^{(n)} \in \mathbb{R}$
- Inputs : $\mathbf{X} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)})^\top$
- Labels : $\mathbf{y} = (y^{(1)}, \ldots, y^{(N)})^\top$
- Goal: : $\mathbf{x} \stackrel{f(\mathbf{x})}{\to} y$

- Implement a linear combination of basis functions

$$f(\mathbf{x}) = \sum_{j=1}^{D} w_j \varphi_j(\mathbf{x}) = \mathbf{w}^\top \boldsymbol{\varphi}(\mathbf{x})$$

with

$$\boldsymbol{\varphi}(\mathbf{x}) = (\varphi_1(\mathbf{x}), \ldots, \varphi_D(\mathbf{x}))^\top$$

- Each $\varphi_i(\mathbf{x})$ is a (non-linear) feature on $\mathbf{x}$, e.g. $x_1, x_2, x_1^2, x_2^2, x_1 x_2 \ldots$

Linear-in–the-Parameters Models: Problem Formulation

- Data: $\mathcal{D} = \{\mathbf{x}^{(n)}, y^{(n)}\}_{n=1}^{N}$, $\mathbf{x}^{(n)} \in \mathbb{R}^{D_x}$, $y^{(n)} \in \mathbb{R}$
- Inputs : $\mathbf{X} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)})^\top$
- Labels : $\mathbf{y} = (y^{(1)}, \ldots, y^{(N)})^\top$
- Goal: : $\mathbf{x} \overset{f(\mathbf{x})}{\to} y$
- Implement a linear combination of basis functions

$$f(\mathbf{x}) = \sum_{j=1}^{D} w_j \varphi_j(\mathbf{x}) = \mathbf{w}^\top \boldsymbol{\varphi}(\mathbf{x})$$

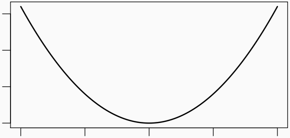
with

$$\boldsymbol{\varphi}(\mathbf{x}) = (\varphi_1(\mathbf{x}), \ldots, \varphi_D(\mathbf{x}))^\top$$

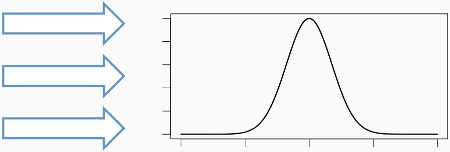
- Each $\varphi_i(\mathbf{x})$ is a (non-linear) feature on $\mathbf{x}$, e.g. $x_1, x_2, x_1^2, x_2^2, x_1 x_2 \ldots$
- Weights : $\mathbf{w} = (w_1, \ldots, w_D)^\top \to$ parameters to *estimate* from data

Probabilistic Interpretation of Loss Minimization

Quadratic Loss



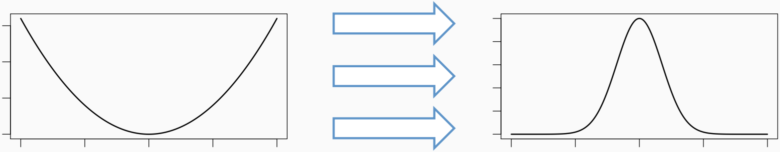
$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) \propto \exp(-\mathrm{Loss})$



- Minimization of a loss function

Probabilistic Interpretation of Loss Minimization

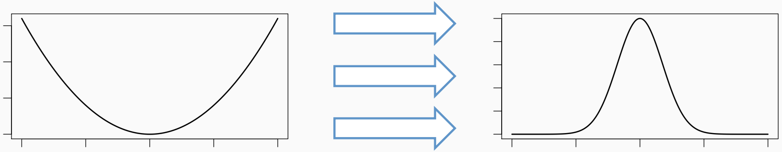
Quadratic Loss $\qquad\qquad$ $p(\mathbf{y}|\mathbf{X}, \mathbf{w}) \propto \exp(-\text{Loss})$



- Minimization of a loss function
- Maximization of conditional likelihood $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$

Probabilistic Interpretation of Loss Minimization

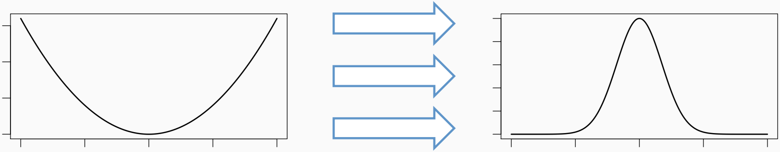
Quadratic Loss $\qquad$ $p(\mathbf{y}|\mathbf{X}, \mathbf{w}) \propto \exp(-\mathrm{Loss})$



- Minimization of a loss function
- Maximization of conditional likelihood $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$
- Assume $p(y \mid \mathbf{w}, \mathbf{x}) = \mathcal{N}(y; \mathbf{w}^\top \boldsymbol{\varphi}(\mathbf{x}), \sigma^2)$

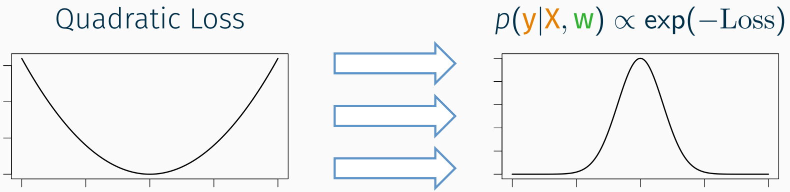
Probabilistic Interpretation of Loss Minimization

Quadratic Loss $\qquad\qquad p(\mathbf{y}|\mathbf{X}, \mathbf{w}) \propto \exp(-\mathrm{Loss})$



- Minimization of a loss function
- Maximization of conditional likelihood $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$
- Assume $p(y \,|\, \mathbf{w}, \mathbf{x}) = \mathcal{N}(y; \mathbf{w}^\top \boldsymbol{\varphi}(\mathbf{x}), \sigma^2)$
- Assume iid observations, i.e., $p(\mathbf{y} \,|\, \mathbf{X}, \mathbf{w}) = \prod_{n=1}^{N} p(y^{(n)} \,|\, \mathbf{x}^{(n)}, \mathbf{w})$

Probabilistic Interpretation of Loss Minimization

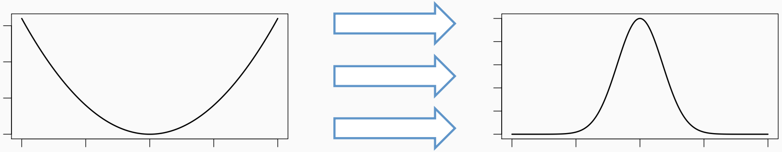


- Minimization of a loss function
- Maximization of conditional likelihood $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$
- Assume $p(y \mid \mathbf{w}, \mathbf{x}) = \mathcal{N}(y; \mathbf{w}^\top \boldsymbol{\varphi}(\mathbf{x}), \sigma^2)$
- Assume iid observations, i.e., $p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = \prod_{n=1}^{N} p(y^{(n)} \mid \mathbf{x}^{(n)}, \mathbf{w})$
- Estimate

$$\hat{\mathbf{w}}_{\mathrm{ML}} = \arg\max_{\mathbf{w}} \log p(\mathbf{y} \mid \mathbf{X}, \mathbf{w})$$

Probabilistic Interpretation of Loss Minimization

Quadratic Loss $\qquad\qquad p(\mathbf{y}|\mathbf{X}, \mathbf{w}) \propto \exp(-\mathrm{Loss})$



- Minimization of a loss function
- Maximization of conditional likelihood $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$
- Assume $p(y \mid \mathbf{w}, \mathbf{x}) = \mathcal{N}(y; \mathbf{w}^\top \boldsymbol{\varphi}(\mathbf{x}), \sigma^2)$
- Assume iid observations, i.e., $p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = \prod_{n=1}^{N} p(y^{(n)} \mid \mathbf{x}^{(n)}, \mathbf{w})$
- Estimate

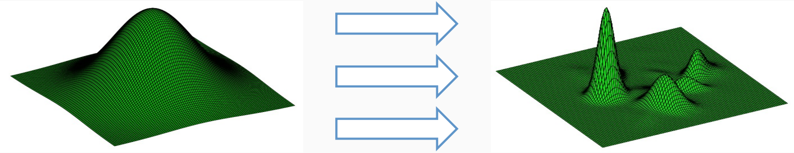
$$\hat{\mathbf{w}}_{\mathrm{ML}} = \arg\max_{\mathbf{w}} \log p(\mathbf{y} \mid \mathbf{X}, \mathbf{w})$$

We will incorporate uncertainty about the weights instead

Bayesian Inference

- Inputs : $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^\top$
- Labels : $\mathbf{y} = (y_1, \ldots, y_N)^\top$
- Weights : $\mathbf{w} = (w_1, \ldots, w_D)^\top$

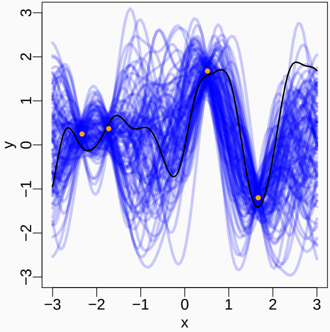
$p(\mathbf{w})$ $\qquad\qquad$ $p(\mathbf{w}|\mathbf{y}, \mathbf{X})$



$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{\int p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})d\mathbf{w}}$$

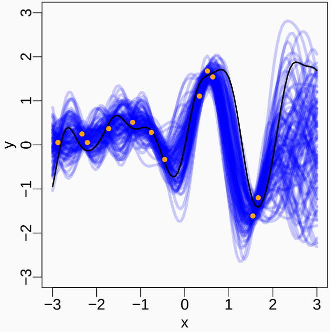
Bayesian Linear Models in Action

- Today's posterior is tomorrow's prior



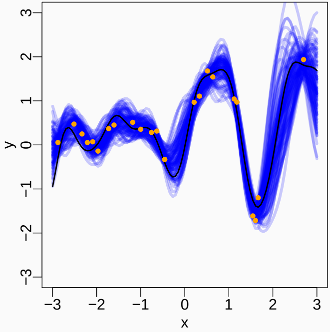
Bayesian Linear Models in Action

- Today's posterior is tomorrow's prior



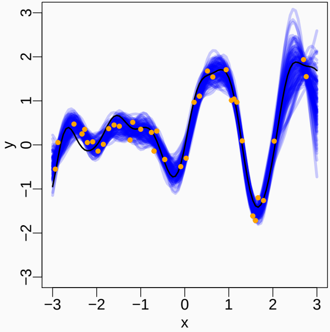
Bayesian Linear Models in Action

- Today's posterior is tomorrow's prior



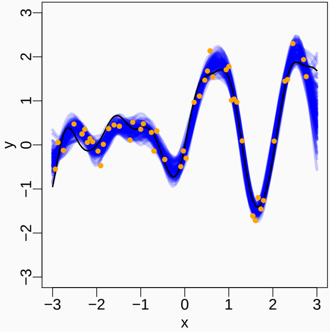
Bayesian Linear Models in Action

- Today's posterior is tomorrow's prior



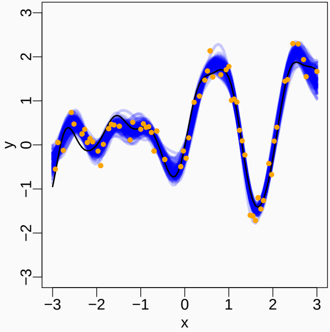
Bayesian Linear Models in Action

- Today's posterior is tomorrow's prior



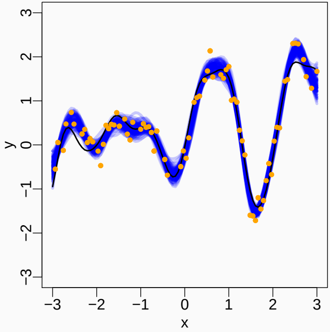
Bayesian Linear Models in Action

- Today's posterior is tomorrow's prior



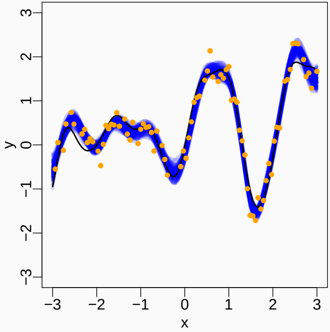
Bayesian Linear Models in Action

- Today's posterior is tomorrow's prior



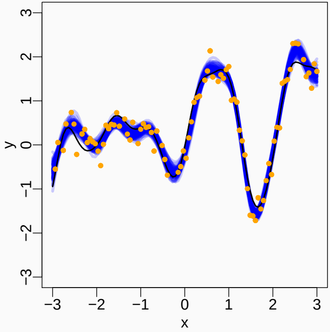
Bayesian Linear Models in Action

- Today's posterior is tomorrow's prior



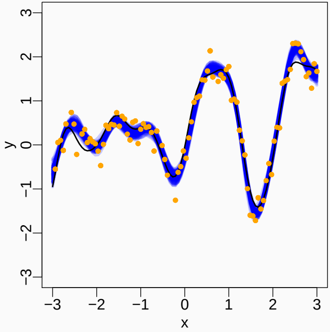
Bayesian Linear Models in Action

- Today's posterior is tomorrow's prior



Bayesian Linear Models in Action

- Today's posterior is tomorrow's prior



Bayesian Linear Regression

- Modelling observations as noisy realizations of a linear combination of the features:

$$p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \sigma^2) = \mathcal{N}(\mathbf{\Phi}\mathbf{w}, \sigma^2\mathbf{I})$$

- $\mathbf{\Phi} = \mathbf{\Phi}(\mathbf{X})$ has entries

$$\mathbf{\Phi} = \begin{bmatrix} \varphi_1(\mathbf{x}_1) & \dots & \varphi_D(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ \varphi_1(\mathbf{x}_N) & \dots & \varphi_D(\mathbf{x}_N) \end{bmatrix}$$

Bayesian Linear Regression

- Modelling observations as noisy realizations of a linear combination of the features:

$$p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \sigma^2) = \mathcal{N}(\Phi\mathbf{w}, \sigma^2\mathbf{I})$$

- $\Phi = \Phi(\mathbf{X})$ has entries

$$\Phi = \begin{bmatrix} \varphi_1(\mathbf{x}_1) & \dots & \varphi_D(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ \varphi_1(\mathbf{x}_N) & \dots & \varphi_D(\mathbf{x}_N) \end{bmatrix}$$

- Gaussian prior over model parameters:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \mathbf{S})$$

Bayesian Linear Regression: Posterior Distribution

- Bayes rule:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{w})p(\mathbf{y}|\mathbf{X}, \mathbf{w})}{\int p(\mathbf{w})p(\mathbf{y}|\mathbf{X}, \mathbf{w})d\mathbf{w}} = \frac{p(\mathbf{w})p(\mathbf{y}|\mathbf{X}, \mathbf{w})}{p(\mathbf{y}|\mathbf{X})}$$

Bayesian Linear Regression: Posterior Distribution

- Bayes rule:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{w})p(\mathbf{y}|\mathbf{X}, \mathbf{w})}{\int p(\mathbf{w})p(\mathbf{y}|\mathbf{X}, \mathbf{w})d\mathbf{w}} = \frac{p(\mathbf{w})p(\mathbf{y}|\mathbf{X}, \mathbf{w})}{p(\mathbf{y}|\mathbf{X})}$$

- **Prior density**: $p(\mathbf{w})$

Bayesian Linear Regression: Posterior Distribution

- Bayes rule:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{w})p(\mathbf{y}|\mathbf{X}, \mathbf{w})}{\int p(\mathbf{w})p(\mathbf{y}|\mathbf{X}, \mathbf{w})d\mathbf{w}} = \frac{p(\mathbf{w})p(\mathbf{y}|\mathbf{X}, \mathbf{w})}{p(\mathbf{y}|\mathbf{X})}$$

- **Prior density**: $p(\mathbf{w})$
  - Anything we know about parameters *before* we see any data

Bayesian Linear Regression: Posterior Distribution

- Bayes rule:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{w})p(\mathbf{y}|\mathbf{X}, \mathbf{w})}{\int p(\mathbf{w})p(\mathbf{y}|\mathbf{X}, \mathbf{w})d\mathbf{w}} = \frac{p(\mathbf{w})p(\mathbf{y}|\mathbf{X}, \mathbf{w})}{p(\mathbf{y}|\mathbf{X})}$$

- **Prior density**: $p(\mathbf{w})$
  - Anything we know about parameters *before* we see any data
- **Conditional Likelihood** : $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$

Bayesian Linear Regression: Posterior Distribution

- Bayes rule:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{w})p(\mathbf{y}|\mathbf{X}, \mathbf{w})}{\int p(\mathbf{w})p(\mathbf{y}|\mathbf{X}, \mathbf{w})d\mathbf{w}} = \frac{p(\mathbf{w})p(\mathbf{y}|\mathbf{X}, \mathbf{w})}{p(\mathbf{y}|\mathbf{X})}$$

- **Prior density**: $p(\mathbf{w})$
  - Anything we know about parameters *before* we see any data
- **Conditional Likelihood** : $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$
  - Measure of "fitness"

Bayesian Linear Regression: Posterior Distribution

- Bayes rule:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{w})p(\mathbf{y}|\mathbf{X}, \mathbf{w})}{\int p(\mathbf{w})p(\mathbf{y}|\mathbf{X}, \mathbf{w})d\mathbf{w}} = \frac{p(\mathbf{w})p(\mathbf{y}|\mathbf{X}, \mathbf{w})}{p(\mathbf{y}|\mathbf{X})}$$

- **Prior density**: $p(\mathbf{w})$
  - Anything we know about parameters *before* we see any data
- **Conditional Likelihood** : $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$
  - Measure of “fitness”
- **Marginal likelihood**: $p(\mathbf{y}|\mathbf{X})$

Bayesian Linear Regression: Posterior Distribution

- Bayes rule:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{w})p(\mathbf{y}|\mathbf{X}, \mathbf{w})}{\int p(\mathbf{w})p(\mathbf{y}|\mathbf{X}, \mathbf{w})d\mathbf{w}} = \frac{p(\mathbf{w})p(\mathbf{y}|\mathbf{X}, \mathbf{w})}{p(\mathbf{y}|\mathbf{X})}$$

- **Prior density**: $p(\mathbf{w})$
  - Anything we know about parameters *before* we see any data
- **Conditional Likelihood** : $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$
  - Measure of “fitness”
- **Marginal likelihood**: $p(\mathbf{y}|\mathbf{X})$
  - It is a normalization constant—ensures $\int p(\mathbf{w}|\mathbf{X}, \mathbf{y})\, d\mathbf{w} = 1$.

Bayesian Linear Regression: Posterior Distribution

- Bayes rule:

$$p(\mathbf{w}|\mathbf{X},\mathbf{y}) = \frac{p(\mathbf{w})p(\mathbf{y}|\mathbf{X},\mathbf{w})}{\int p(\mathbf{w})p(\mathbf{y}|\mathbf{X},\mathbf{w})d\mathbf{w}} = \frac{p(\mathbf{w})p(\mathbf{y}|\mathbf{X},\mathbf{w})}{p(\mathbf{y}|\mathbf{X})}$$

- **Prior density**: $p(\mathbf{w})$
  - Anything we know about parameters *before* we see any data
- **Conditional Likelihood** : $p(\mathbf{y}|\mathbf{X},\mathbf{w})$
  - Measure of "fitness"
- **Marginal likelihood**: $p(\mathbf{y}|\mathbf{X})$
  - It is a normalization constant—ensures $\int p(\mathbf{w}|\mathbf{X},\mathbf{y})\, d\mathbf{w} = 1$.
- **Posterior density**: $p(\mathbf{w}|\mathbf{X},\mathbf{y})$

Bayesian Linear Regression: Posterior Distribution

- Bayes rule:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{w})p(\mathbf{y}|\mathbf{X}, \mathbf{w})}{\int p(\mathbf{w})p(\mathbf{y}|\mathbf{X}, \mathbf{w})d\mathbf{w}} = \frac{p(\mathbf{w})p(\mathbf{y}|\mathbf{X}, \mathbf{w})}{p(\mathbf{y}|\mathbf{X})}$$

- **Prior density**: $p(\mathbf{w})$
  - Anything we know about parameters *before* we see any data
- **Conditional Likelihood** : $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$
  - Measure of "fitness"
- **Marginal likelihood**: $p(\mathbf{y}|\mathbf{X})$
  - It is a normalization constant—ensures $\int p(\mathbf{w}|\mathbf{X}, \mathbf{y})\, d\mathbf{w} = 1$.
- **Posterior density**: $p(\mathbf{w}|\mathbf{X}, \mathbf{y})$
  - Distribution over parameters *after* observing data

Bayesian Linear Regression: Posterior Distribution

- Recall Gaussian prior over weights $p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \mathbf{S})$

Bayesian Linear Regression: Posterior Distribution

- Recall Gaussian prior over weights $p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \mathbf{S})$
- Also Gaussian likelihood assumption $p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \sigma^2) = \mathcal{N}(\mathbf{\Phi}\mathbf{w}, \sigma^2\mathbf{I})$

Bayesian Linear Regression: Posterior Distribution

- Recall Gaussian prior over weights $p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \mathbf{S})$
- Also Gaussian likelihood assumption $p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \sigma^2) = \mathcal{N}(\Phi\mathbf{w}, \sigma^2\mathbf{I})$
- Posterior **must be** Gaussian (Proof in the Appendix)

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \sigma^2) = \mathcal{N}(\boldsymbol{\mu}, \Sigma)$$

Bayesian Linear Regression: Posterior Distribution

- Recall Gaussian prior over weights $p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \mathbf{S})$
- Also Gaussian likelihood assumption $p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \sigma^2) = \mathcal{N}(\mathbf{\Phi}\mathbf{w}, \sigma^2\mathbf{I})$
- Posterior **must be** Gaussian (Proof in the Appendix)

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \sigma^2) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$$

- Covariance: $\mathbf{\Sigma} = \left(\frac{1}{\sigma^2}\mathbf{\Phi}^\top\mathbf{\Phi} + \mathbf{S}^{-1}\right)^{-1}$, Mean: $\boldsymbol{\mu} = \frac{1}{\sigma^2}\mathbf{\Sigma}\mathbf{\Phi}^\top\mathbf{y}$

Bayesian Linear Regression: Posterior Distribution

- Recall Gaussian prior over weights $p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \mathbf{S})$
- Also Gaussian likelihood assumption $p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \sigma^2) = \mathcal{N}(\mathbf{\Phi}\mathbf{w}, \sigma^2\mathbf{I})$
- Posterior **must be** Gaussian (Proof in the Appendix)

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \sigma^2) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$$

- Covariance: $\mathbf{\Sigma} = \left(\frac{1}{\sigma^2}\mathbf{\Phi}^\top\mathbf{\Phi} + \mathbf{S}^{-1}\right)^{-1}$, Mean: $\boldsymbol{\mu} = \frac{1}{\sigma^2}\mathbf{\Sigma}\mathbf{\Phi}^\top\mathbf{y}$
- Mean of posterior is equal to its mode

Bayesian Linear Regression: Posterior Distribution

- Recall Gaussian prior over weights $p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \mathbf{S})$
- Also Gaussian likelihood assumption $p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \sigma^2) = \mathcal{N}(\Phi\mathbf{w}, \sigma^2\mathbf{I})$
- Posterior **must be** Gaussian (Proof in the Appendix)

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \sigma^2) = \mathcal{N}(\boldsymbol{\mu}, \Sigma)$$

- Covariance: $\Sigma = \left(\frac{1}{\sigma^2}\Phi^\top\Phi + \mathbf{S}^{-1}\right)^{-1}$, Mean: $\boldsymbol{\mu} = \frac{1}{\sigma^2}\Sigma\Phi^\top\mathbf{y}$
- Mean of posterior is equal to its mode
- Maximum a posteriori (MAP) :

$$\hat{\mathbf{w}}_{\text{MAP}} = \arg\max_{\mathbf{w}} \log p(\mathbf{w} \,|\, \mathbf{X}, \mathbf{y}, \sigma^2) = \arg\max_{\mathbf{w}} \left[\log p(\mathbf{w}) + \log p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \sigma^2)\right]$$

Bayesian Linear Regression: Predictive Distribution

We are interested in making predictions at a new test point $\mathbf{x}_*$

- We obtain the predictive distribution by *averaging* over all possible parameter values

$$p(y_* \mid \mathbf{X}, \mathbf{y}, \mathbf{x}_*, \sigma^2) = \int p(y_* \mid \mathbf{w}, \mathbf{x}_*, \sigma^2) p(\mathbf{w} \mid \mathbf{X}, \mathbf{y}, \sigma^2) d\mathbf{w}$$
$$= \mathcal{N}(\mu_*, \sigma_*^2)$$

Bayesian Linear Regression: Predictive Distribution

We are interested in making predictions at a new test point $\mathbf{x}_*$

- We obtain the predictive distribution by *averaging* over all possible parameter values

$$p(y_* \,|\, \mathbf{X}, \mathbf{y}, \mathbf{x}_*, \sigma^2) = \int p(y_* \,|\, \mathbf{w}, \mathbf{x}_*, \sigma^2) p(\mathbf{w} \,|\, \mathbf{X}, \mathbf{y}, \sigma^2) d\mathbf{w}$$
$$= \mathcal{N}(\mu_*, \sigma_*^2)$$

- Predictive mean: $\mu_* = \boldsymbol{\varphi}(\mathbf{x}_*)^\top \boldsymbol{\mu} = \sigma^{-2} \boldsymbol{\varphi}(\mathbf{x}_*)^\top \Sigma \Phi^\top \mathbf{y}$

Bayesian Linear Regression: Predictive Distribution

We are interested in making predictions at a new test point $\mathbf{x}_*$

- We obtain the predictive distribution by *averaging* over all possible parameter values

$$p(y_* \,|\, \mathbf{X}, \mathbf{y}, \mathbf{x}_*, \sigma^2) = \int p(y_* \,|\, \mathbf{w}, \mathbf{x}_*, \sigma^2) p(\mathbf{w} \,|\, \mathbf{X}, \mathbf{y}, \sigma^2) d\mathbf{w}$$
$$= \mathcal{N}(\mu_*, \sigma_*^2)$$

- Predictive mean: $\mu_* = \boldsymbol{\varphi}(\mathbf{x}_*)^\top \boldsymbol{\mu} = \sigma^{-2} \boldsymbol{\varphi}(\mathbf{x}_*)^\top \Sigma \Phi^\top \mathbf{y}$
  - Linear predictor

Bayesian Linear Regression: Predictive Distribution

We are interested in making predictions at a new test point $\mathbf{x}_*$

- We obtain the predictive distribution by *averaging* over all possible parameter values

$$
\begin{aligned}
p(y_* \,|\, \mathbf{X}, \mathbf{y}, \mathbf{x}_*, \sigma^2) &= \int p(y_* \,|\, \mathbf{w}, \mathbf{x}_*, \sigma^2) p(\mathbf{w} \,|\, \mathbf{X}, \mathbf{y}, \sigma^2) d\mathbf{w} \\
&= \mathcal{N}(\mu_*, \sigma_*^2)
\end{aligned}
$$

- Predictive mean: $\mu_* = \boldsymbol{\varphi}(\mathbf{x}_*)^\top \boldsymbol{\mu} = \sigma^{-2} \boldsymbol{\varphi}(\mathbf{x}_*)^\top \Sigma \Phi^\top \mathbf{y}$
  - ▶ Linear predictor
- Predictive variance: $\sigma_*^2 = \sigma^2 + \boldsymbol{\varphi}(\mathbf{x}_*)^\top \Sigma \boldsymbol{\varphi}(\mathbf{x}_*)$

Bayesian Linear Regression: Predictive Distribution

We are interested in making predictions at a new test point $\mathbf{x}_*$

- We obtain the predictive distribution by *averaging* over all possible parameter values

$$p(y_* \mid \mathbf{X}, \mathbf{y}, \mathbf{x}_*, \sigma^2) = \int p(y_* \mid \mathbf{w}, \mathbf{x}_*, \sigma^2) p(\mathbf{w} \mid \mathbf{X}, \mathbf{y}, \sigma^2) d\mathbf{w}$$
$$= \mathcal{N}(\mu_*, \sigma_*^2)$$

- Predictive mean: $\mu_* = \boldsymbol{\varphi}(\mathbf{x}_*)^\top \boldsymbol{\mu} = \sigma^{-2} \boldsymbol{\varphi}(\mathbf{x}_*)^\top \boldsymbol{\Sigma} \boldsymbol{\Phi}^\top \mathbf{y}$
  - Linear predictor
- Predictive variance: $\sigma_*^2 = \sigma^2 + \boldsymbol{\varphi}(\mathbf{x}_*)^\top \boldsymbol{\Sigma} \boldsymbol{\varphi}(\mathbf{x}_*)$
- Note computation of D-dimensional inverse $\boldsymbol{\Sigma}$

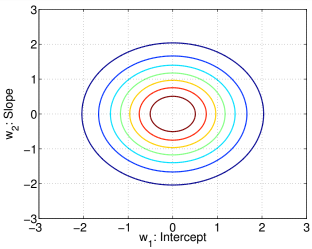
Bayesian Linear Regression: Point Prediction

To make a point prediction we need to consider the expected loss (or risk):

$$y_{\text{opt}} = \arg\min_{y_{\text{pred}}} \int \text{Loss}(y_*, y_{\text{pred}}) p(y_* \mid \mathbf{X}, \mathbf{y}, \mathbf{x}_*, \sigma^2) dy_*$$

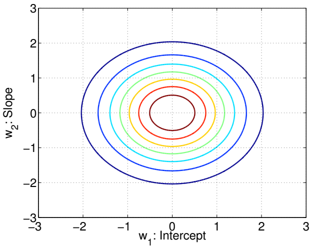
- e.g., square loss: $\text{Loss}(y_*, y_{\text{pred}}) = (y_* - y_{\text{pred}})^2$
- Predictions at the mean of the distribution
- c.f. empirical risk minimization (ERM)

Bayesian Linear Regression Example

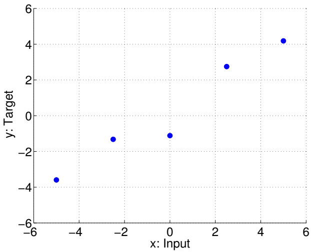


Prior Weights

Bayesian Linear Regression Example

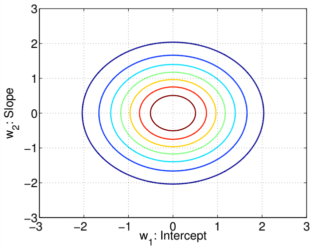


Prior Weights

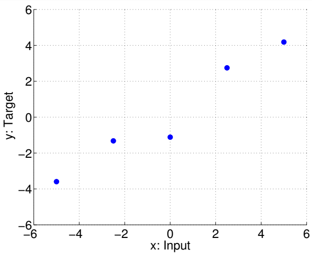


Observed Data

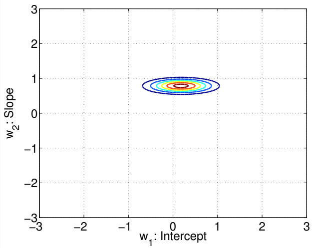
Bayesian Linear Regression Example



Prior Weights

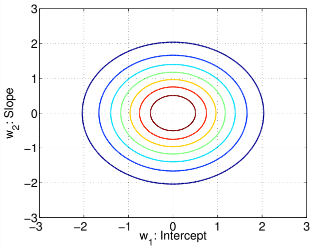


Observed Data

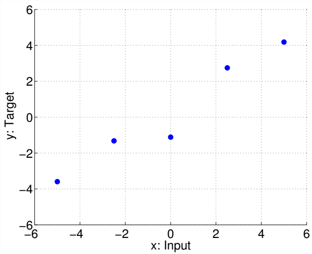


Likelihood

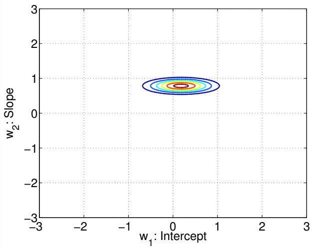
Bayesian Linear Regression Example



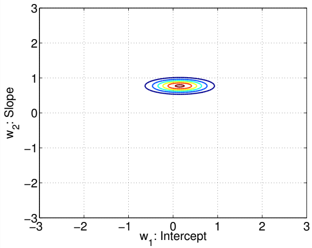
Prior Weights



Observed Data

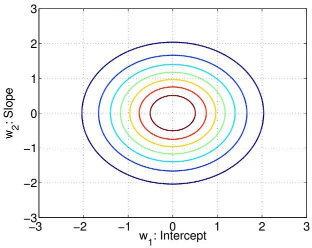


Likelihood

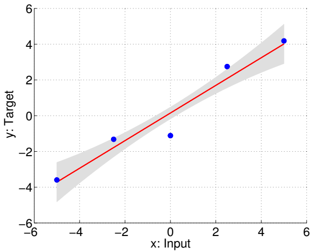


Posterior Weights

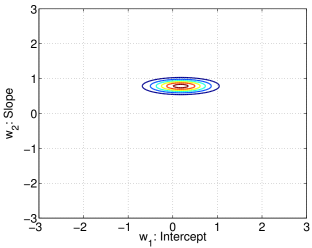
Bayesian Linear Regression Example



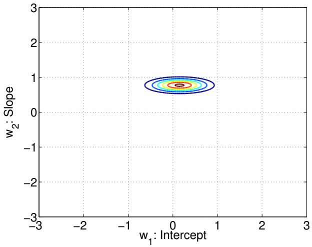
Prior Weights



Predictive Distribution



Likelihood



Posterior Weights

www.kahoot.it

Conclusions

- Importance of quantification of uncertainty in machine learning
- Probability theory is key
- Joint distributions, marginals, conditionals
- Bayesian inference: Prior, likelihood, posterior
- Bayesian linear (in-the-parameters) regression
  - Full predictive distribution in closed-form
  - Fixed set of basis functions
  - Cubic cost on these features' dimensionality

Appendix

Bayesian Linear Regression - Finding posterior parameters

- Ignoring normalizing constants, the posterior is:

$$
\begin{aligned}
p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \sigma^2) &\propto \exp\left\{-\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{w} - \boldsymbol{\mu})\right\} \\
&= \exp\left\{-\frac{1}{2}(\mathbf{w}^\top \boldsymbol{\Sigma}^{-1}\mathbf{w} - 2\mathbf{w}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right\} \\
&\propto \exp\left\{-\frac{1}{2}(\mathbf{w}^\top \boldsymbol{\Sigma}^{-1}\mathbf{w} - 2\mathbf{w}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right\}
\end{aligned}
$$

Bayesian Linear Regression - Finding posterior parameters

- Ignoring non-$\mathbf{w}$ terms, the prior multiplied by the likelihood is:

$$
\begin{aligned}
& p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \sigma^2) \\
\propto \quad & \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{\Phi}\mathbf{w})^\top(\mathbf{y} - \mathbf{\Phi}\mathbf{w})\right\} \exp\left\{-\frac{1}{2}\mathbf{w}^\top \mathbf{S}^{-1}\mathbf{w}\right\} \\
\propto \quad & \exp\left\{-\frac{1}{2}\left(\mathbf{w}^\top\left[\frac{1}{\sigma^2}\mathbf{\Phi}^\top\mathbf{\Phi} + \mathbf{S}^{-1}\right]\mathbf{w} - \frac{2}{\sigma^2}\mathbf{w}^\top\mathbf{\Phi}^\top\mathbf{y}\right)\right\}
\end{aligned}
$$

- Posterior (from previous slide):

$$
\propto \exp\left\{-\frac{1}{2}(\mathbf{w}^\top\mathbf{\Sigma}^{-1}\mathbf{w} - 2\mathbf{w}^\top\mathbf{\Sigma}^{-1}\boldsymbol{\mu})\right\}
$$

Bayesian Linear Regression - Finding posterior parameters

- Equate individual terms on each side.
- Covariance:

$$
\begin{aligned}
\mathbf{w}^\top \boldsymbol{\Sigma}^{-1} \mathbf{w} &= \mathbf{w}^\top \left[\frac{1}{\sigma^2} \boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \mathbf{S}^{-1} \right] \mathbf{w} \\
\boldsymbol{\Sigma} &= \left(\frac{1}{\sigma^2} \boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \mathbf{S}^{-1} \right)^{-1}
\end{aligned}
$$

- Mean:

$$
\begin{aligned}
2\mathbf{w}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} &= \frac{2}{\sigma^2} \mathbf{w}^\top \boldsymbol{\Phi}^\top \mathbf{y} \\
\boldsymbol{\mu} &= \frac{1}{\sigma^2} \boldsymbol{\Sigma} \boldsymbol{\Phi}^\top \mathbf{y}
\end{aligned}
$$