DD2434 - Advanced Machine Learning

Modelling

Carl Henrik Ek
{chek}@csc.kth.se

Royal Institute of Technology

November 4, 2015

Who do I think you are?

- Mathematically competent
  - linear algebra
  - multivariate calculus
- Ok programmers
- Able to extend knowledge beyond lectures
- Motivated and willing to learn
- Not expecting cookbook recipies

Who do I hope that you will become?

- Understand the importance of uncertainty
- See ML as a science not a collection of methods
- Capable to place methods in context
- Have the background to learn by yourselves
- Appreciating the difficulties and challenges to ML

Whats the focus of this part of the course

My plan

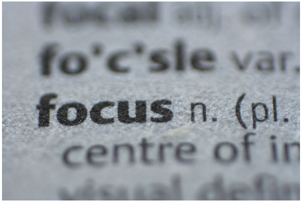
- My view on Machine Learning

1. Look at each part of a probabilistic model in detail
   - how do they interact
   - what do they provide
2. Different models
   - parametric
   - non-parameteric
3. Inference

- Really simple data

Block Structure

- 4 Lectures
- 2 Practical sessions
- 1 Assignment
  - Deadline November 19th
- 1 Scheduled Help session



Assignment

- Three parts aligned with lectures
- Part 1 (Lecture 2 & 3)
  - Task: probabilistic regression
  - Aim: understand probabilistic objects
- Part 2 (Lecture 3 & 4)
  - Task: probabilistic representation learning
  - Aim: understand probabilistic methodology
- Part 3 (Self study)
  - Task: probabilistic model selection
  - Aim: show that you can extend your knowledge from 1 and 2

Assignment

- Three parts aligned with lectures
- Part 1 (Lecture 2 & 3)
  - Task: probabilistic regression
  - Aim: understand probabilistic objects
- Part 2 (Lecture 3 & 4)
  - Task: probabilistic representation learning
  - Aim: understand probabilistic methodology
- Part 3 (Self study)
  - Task: probabilistic model selection
  - Aim: show that you can extend your knowledge from 1 and 2

Assignment

- Three parts aligned with lectures
- Part 1 (Lecture 2 & 3)
  - Task: probabilistic regression
  - Aim: understand probabilistic objects
- Part 2 (Lecture 3 & 4)
  - Task: probabilistic representation learning
  - Aim: understand probabilistic methodology
- Part 3 (Self study)
  - Task: probabilistic model selection
  - Aim: show that you can extend your knowledge from 1 and 2

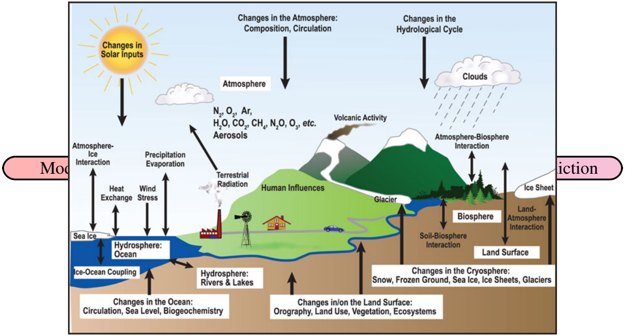
Assignment

- Three parts aligned with lectures
- Part 1 (Lecture 2 & 3)
  - Task: probabilistic regression
  - Aim: understand probabilistic objects
- Part 2 (Lecture 3 & 4)
  - Task: probabilistic representation learning
  - Aim: understand probabilistic methodology
- Part 3 (Self study)
  - Task: probabilistic model selection
  - Aim: show that you can extend your knowledge from 1 and 2

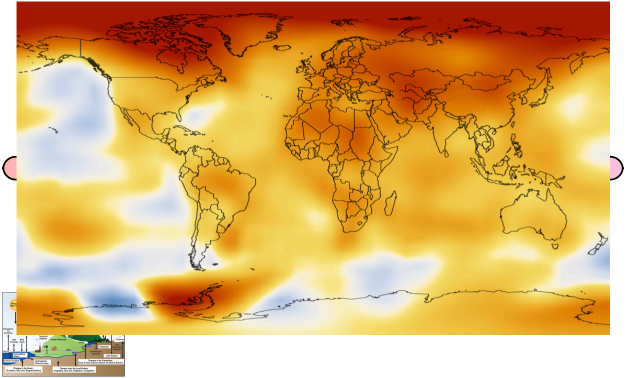
“Science”

Model Data Algorithm Prediction

"Science"


Mod
iction
Changes in Solar Inputs
Changes in the Atmosphere: Composition, Circulation
Changes in the Hydrological Cycle
Clouds
Atmosphere
N₂, O₂, Ar, H₂O, CO₂, CH₄, N₂O, O₃, etc. Aerosols
Volcanic Activity
Atmosphere-Biosphere Interaction
Atmosphere-Ice Interaction
Precipitation Evaporation
Terrestrial Radiation
Heat Exchange
Wind Stress
Human Influences
Glacier
Ice Sheet
Biosphere
Land-Atmosphere Interaction
Soil-Biosphere Interaction
Land Surface
Sea Ice
Hydrosphere: Ocean
Ice-Ocean Coupling
Hydrosphere: Rivers & Lakes
Changes in the Cryosphere: Snow, Frozen Ground, Sea Ice, Ice Sheets, Glaciers
Changes in the Ocean: Circulation, Sea Level, Biogeochemistry
Changes in/on the Land Surface: Orography, Land Use, Vegetation, Ecosystems

“Science”



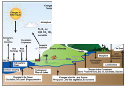
"Science"

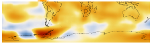
M

$$\begin{array}{r} 34 \\ 2\overline{)68} \\ -6 \\ \hline 08 \\ 8 \end{array}$$

$4 \times 2 = 8$

tion

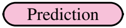


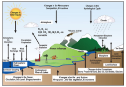


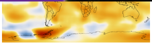
“Science”

Model

Prediction

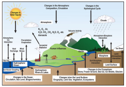


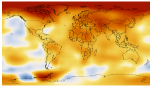


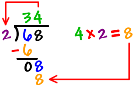


“Science”

Model Data Algorithm Prediction







My view on Machine Learning

My view on Machine Learning

An intelligence which at a given instant knew all the forces acting in nature and the position of every object in the universe

[1]

---

[1] *A philosophical essay on probabilities*, Laplace

My view on Machine Learning

An intelligence which at a given instant knew all the forces acting in nature and the position of every object in the universe - if endowed with a brain sufficiently vast to make all necessary calculations

[1]

[1] *A philosophical essay on probabilities*, Laplace

My view on Machine Learning

An intelligence which at a given instant knew all the forces acting in nature and the position of every object in the universe - if endowed with a brain sufficiently vast to make all necessary calculations - could describe with a single formula the motions of the largest astronomical bodies and those of the smallest atoms.

[1]

[1] *A philosophical essay on probabilities*, Laplace

My view on Machine Learning

An intelligence which at a given instant knew all the forces acting in nature and the position of every object in the universe - if endowed with a brain sufficiently vast to make all necessary calculations - could describe with a single formula the motions of the largest astronomical bodies and those of the smallest atoms. ***To such an intelligence, nothing would be uncertain; the future, like the past, would be an open book.***

[1]

[1] *A philosophical essay on probabilities*, Laplace

My view on Machine Learning

The theory of probabilities is at bottom nothing but common sense reduced to calculus; it enables us to appreciate with exactness that which accurate minds feel with a sort of instinct for which of times they are unable to account.

[1]

[1] *A philosophical essay on probabilities*, Laplace

My view on Machine Learning

It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge.

[1]

[1] *A philosophical essay on probabilities*, Laplace

My view on Machine Learning

It was our use of probability theory as logic that has enabled us to do so easily what was impossible for those who thought of probability as a physical phenomenon associated with "randomness". Quite the opposite; we have thought of probability distributions as carriers of information.

[1]

[1] *Probability theory: The logic of science*, Jaynes

My view on Machine Learning

Theme

- *Acknowledge that we do not know everything, i.e. my knowledge is uncertain.*
- *Methodology to propagate my uncertainty through all levels of reasoning.*
- *Incorporate my uncertain knowledge with observations such that when I see data it reduces my uncertainty according to the evidence provided in the observations.*

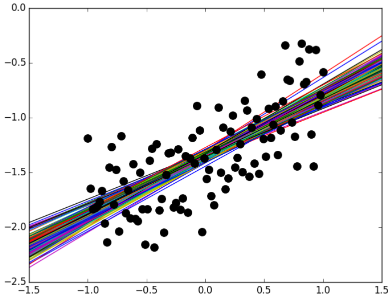
Introduction

Regression

Kernel Methods

Regression

- Two variates
  - Input data $\mathbf{x}_i \in \mathbb{R}^q$
  - Output data $\mathbf{y}_i \in \mathbb{R}^D$
- Relationship: $f : \mathbf{X} \rightarrow \mathbf{Y}$



Regression

Uncertainty

- We are uncertain in our data
- This means we cannot trust
  - our observations
  - the mapping that we learn
  - the predictions that we make under the mapping
- **This part of the course is about making this principled!**

Regression

Uncertainty

- We are uncertain in our data
- This means we cannot trust
  - our observations
  - the mapping that we learn
  - the predictions that we make under the mapping
- **This part of the course is about making this principled!**

Outline

- Re-cap of Probability basics
- Re-cap Central Limit Theorem
- Probabilistic formulation
- Dual Formulation

Probability Basics[1]

Expected Value

$$\mathbb{E}[\mathbf{x}] = \mu(\mathbf{x}) = \int \mathbf{x}p(\mathbf{x})\mathrm{d}\mathbf{x} \tag{1}$$

- Shows the “center of gravity” of a distribution
- Sampled expected value (mean)

$$\overline{\mathbf{x}} = \frac{1}{N}\sum_{i}^{N}\mathbf{x}_i \tag{2}$$

[1] Bishop 2006, p. 1.2.2.

Probability Basics[1]

Variance

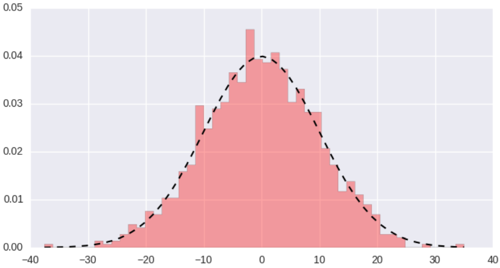
$$\sigma^2(\mathbf{x}) = \mathrm{var}[\mathbf{x}] = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])^2] \tag{3}$$

- Shows the “spread” of a distribution
- Sample variance

$$\overline{\sigma^2(\mathbf{x})} = \frac{1}{N-1}\sum_{i}^{N}(\mathbf{x}_i - \mu(\mathbf{x}_i))^2 \tag{4}$$

[1] Bishop 2006, p. 1.2.2.

Probability Basics[2]

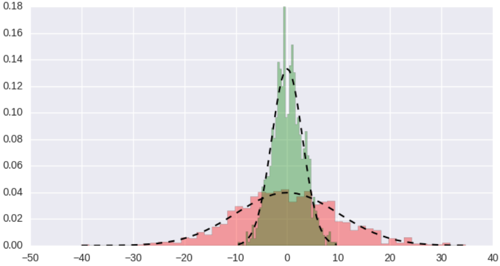


1

[1] `Matplotlib3D, /Lecture1/probBasics.py`

[2] Bishop 2006, p. 1.2.2.

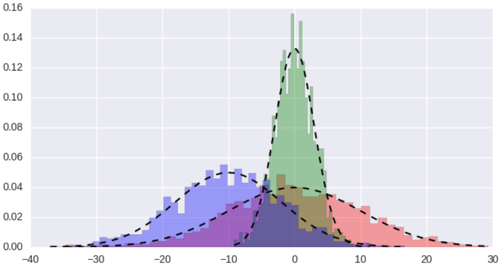
Probability Basics[2]



1

[1] `Matplotlib3D, /Lecture1/probBasics.py`
[2] Bishop 2006, p. 1.2.2.

Probability Basics[2]



1

[1] `Matplotlib3D, /Lecture1/probBasics.py`
[2] Bishop 2006, p. 1.2.2.

Probability Basics[1]

Covariance

$$\sigma(\mathbf{x}, \mathbf{y}) = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])] \tag{5}$$

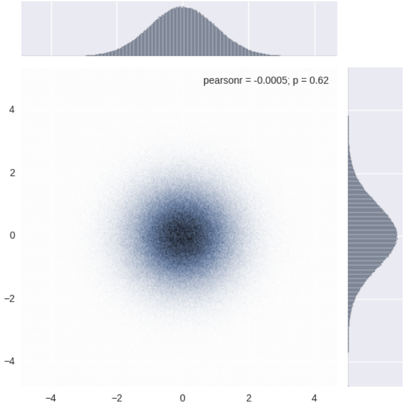
$$[\sigma(\mathbf{X}, \mathbf{Y})]_{ij} = \sigma(\mathbf{x}_i, \mathbf{y}_j) = k(\mathbf{x}_i, \mathbf{y}_j) \tag{6}$$

- Shows how the “spread” of how to variables vary *together*
- Sample co-variance

$$\overline{\sigma(\mathbf{x}, \mathbf{y})} = \frac{1}{N-1} \sum_{i}^{N} (\mathbf{x}_i - \mu(\mathbf{x}_i))(\mathbf{y}_i - \mu(\mathbf{y})) \tag{7}$$

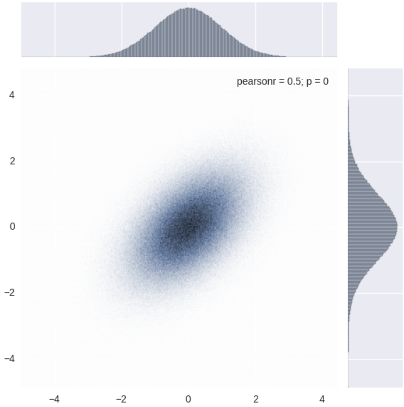
[1] Bishop 2006, p. 1.2.2.

Probability Basics[1]



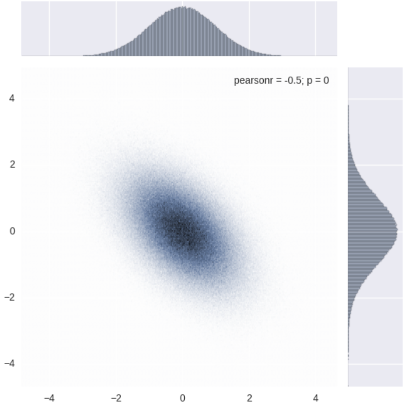
[1] Bishop 2006, p. 1.2.2.

Probability Basics[1]



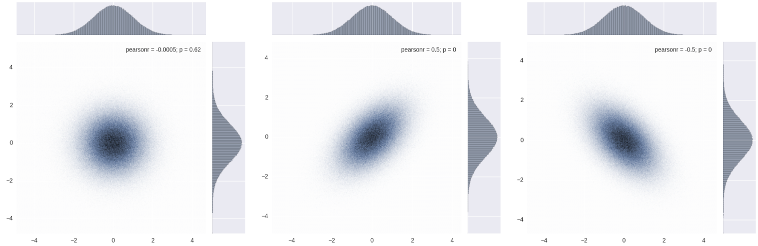
[1] Bishop 2006, p. 1.2.2.

Probability Basics[1]



[1] Bishop 2006, p. 1.2.2.

Probability Basics[2]



[1]

[1] Matplotlib3D, /Lecture1/probBasics.py

[2] Bishop 2006, p. 1.2.2.

Linear Regression[3]

$$\mathbf{y}_i = \mathbf{W}\mathbf{x}_i \tag{8}$$

Uncertainty

- Lets assume the relationship is linear
- Uncertainty in outputs $\mathbf{y}_i$
  - Addative noise $\mathbf{y}_i = \mathbf{W}\mathbf{x}_i + \epsilon$
  - What form does the noise have $\epsilon \propto$
  - What do we know about the generating process?

[3] Bishop 2006, p. 3.3.1.

Linear Regression[3]

$$\mathbf{y}_i = \mathbf{W}\mathbf{x}_i + \epsilon \quad (9)$$

Uncertainty

- Lets assume the relationship is linear
- Uncertainty in outputs $\mathbf{y}_i$
  - Addative noise $\mathbf{y}_i = \mathbf{W}\mathbf{x}_i + \epsilon$
  - What form does the noise have $\epsilon \propto$
  - What do we know about the generating process?

[3] Bishop 2006, p. 3.3.1.

Linear Regression[3]

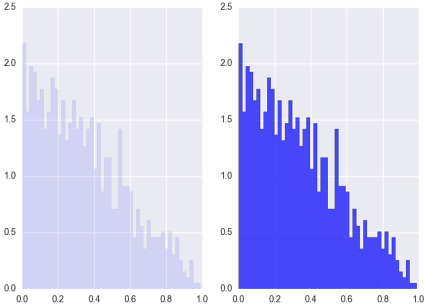
What distribution?

- Central Limit Theorem[a]
- The central limit theorem states that the distribution of the sum (or average) of a large number of independent, identically distributed variables will be approximately normal, regardless of the underlying distribution.

[a] Bishop 2006, p. 78

[3] Bishop 2006, p. 3.3.1.

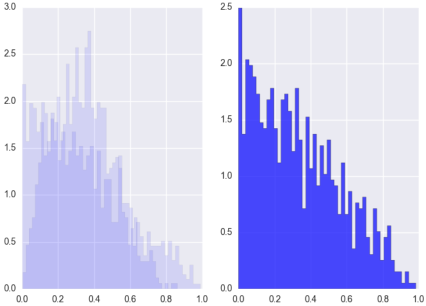
Linear Regression[4]



3

[3] `/Lecture1/centralLimit.py`
[4] Bishop 2006, p. 3.3.1.

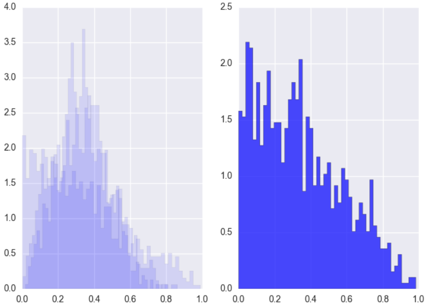
Linear Regression[4]



3

[3] /Lecture1/centralLimit.py
[4] Bishop 2006, p. 3.3.1.

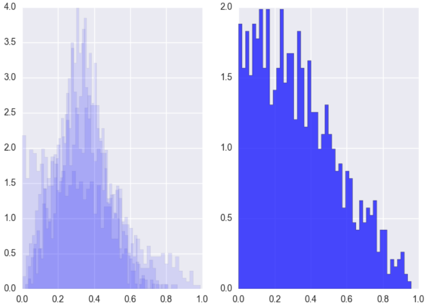
Linear Regression[4]



3

[3] /Lecture1/centralLimit.py
[4] Bishop 2006, p. 3.3.1.

Linear Regression[4]

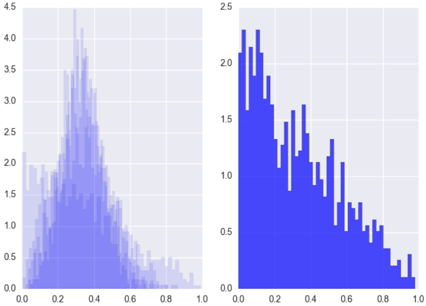


3

[3] /Lecture1/centralLimit.py

[4] Bishop 2006, p. 3.3.1.

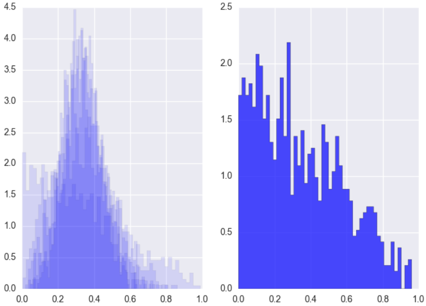
Linear Regression[4]



3

[3] `/Lecture1/centralLimit.py`
[4] Bishop 2006, p. 3.3.1.

Linear Regression[4]

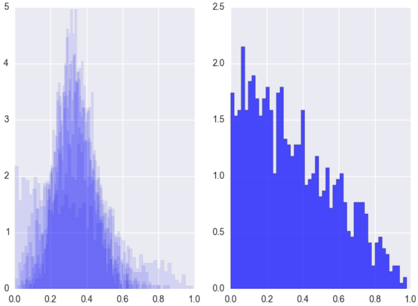


3

[3] `/Lecture1/centralLimit.py`

[4] Bishop 2006, p. 3.3.1.

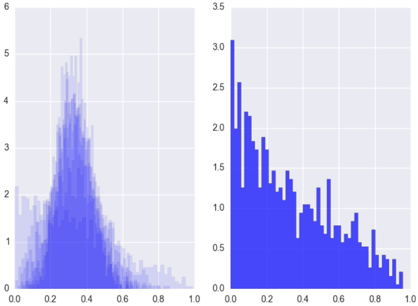
Linear Regression[4]



3

[3] `/Lecture1/centralLimit.py`
[4] Bishop 2006, p. 3.3.1.

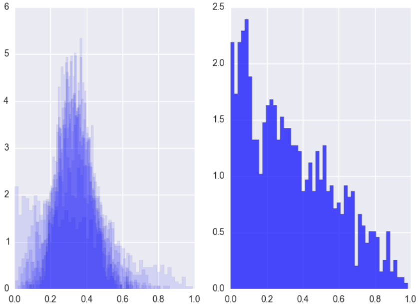
Linear Regression[4]



3

[3] `/Lecture1/centralLimit.py`
[4] Bishop 2006, p. 3.3.1.

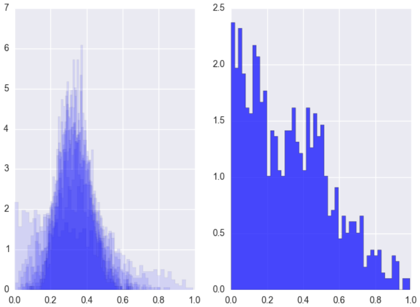
Linear Regression[4]



3

[3] `/Lecture1/centralLimit.py`
[4] Bishop 2006, p. 3.3.1.

Linear Regression[4]

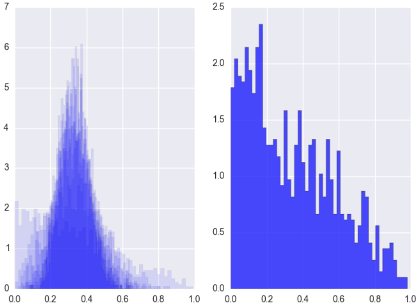


3

[3] `/Lecture1/centralLimit.py`

[4] Bishop 2006, p. 3.3.1.

Linear Regression[4]

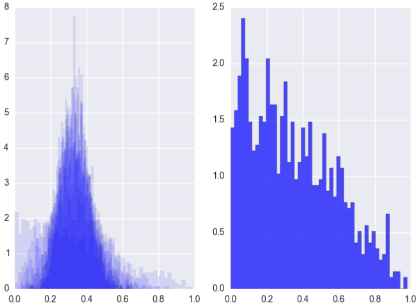


3

[3] `/Lecture1/centralLimit.py`

[4] Bishop 2006, p. 3.3.1.

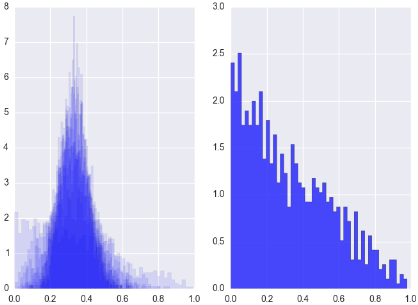
Linear Regression[4]



3

[3] `/Lecture1/centralLimit.py`
[4] Bishop 2006, p. 3.3.1.

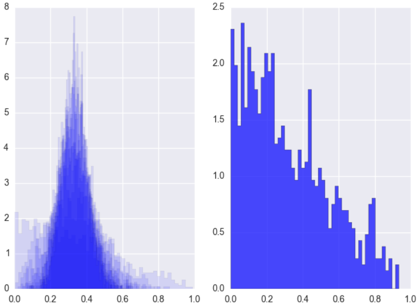
Linear Regression[4]



3

[3] `/Lecture1/centralLimit.py`
[4] Bishop 2006, p. 3.3.1.

Linear Regression[4]

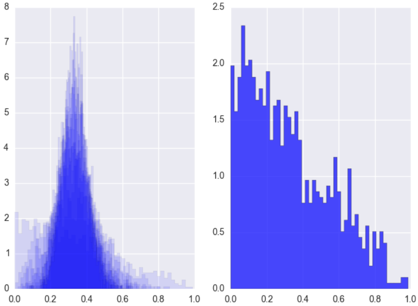


3

[3] /Lecture1/centralLimit.py

[4] Bishop 2006, p. 3.3.1.

Linear Regression[4]

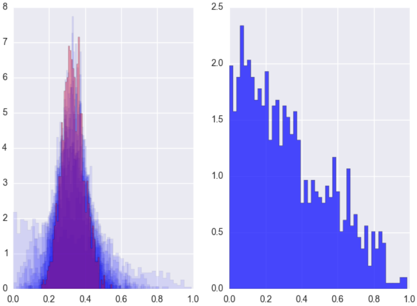


3

[3] `/Lecture1/centralLimit.py`

[4] Bishop 2006, p. 3.3.1.

Linear Regression[4]



3

[3] `/Lecture1/centralLimit.py`

[4] Bishop 2006, p. 3.3.1.

Linear Regression[3]

$$p(\mathbf{W}|\mathbf{Y}, \mathbf{X}) \tag{10}$$

Uncertainty in Model

- Posterior
  - conditional distribution
  - *after* the relevant information has been taken into account
- What is relevant
  - our belief
  - the observations

[3] Bishop 2006, p. 3.3.1.

Linear Regression[3]

$$p(\mathbf{W}) \tag{11}$$

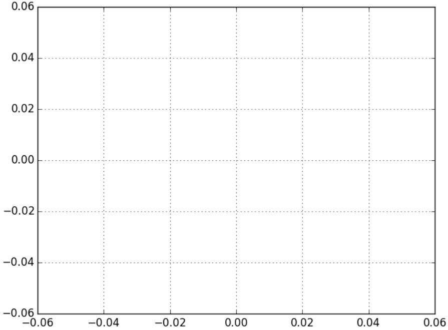
Belief about model before seeing data

- Prior
- What do I know about the regression parameters

[3] Bishop 2006, p. 3.3.1.

Linear Regression[3]

$$p(\mathbf{W}) \tag{12}$$



Linear Regression[3]

$$p(\mathbf{W}) \tag{13}$$

Belief about model before seeing data

- Prior
- What do I know about the regression parameters

$$p(\mathbf{W}) = \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \tag{14}$$

[3] Bishop 2006, p. 3.3.1.

Linear Regression[3]

$$p(\mathbf{y}_i|\mathbf{W},\mathbf{x}_i) \tag{15}$$

How well does my model predict the data

- Likelihood
- Think error function but also how different errors

$$p(\mathbf{y}_i|\mathbf{W},\mathbf{x}_i) = \mathcal{N}(\mathbf{y}_i|\mathbf{W}\mathbf{x}_i,\tau^2\mathbf{I}) \tag{16}$$

[3] Bishop 2006, p. 3.3.1.

Linear Regression[3]

Structure

- Do the variables co-vary?
- Are there (in-)dependency structures that I can exploit?

$$p(\mathbf{Y}|\mathbf{W},\mathbf{X}) = \prod_{i}^{N} p(\mathbf{y}_i|\mathbf{W},\mathbf{x}_i) \tag{17}$$

[3] Bishop 2006, p. 3.3.1.

Linear Regression[3]

How do we put everything together?

- Want to reach the posterior
  - distribution after *all* relevant information have been taken into account
- Prediction should reflect my beliefs in the model **and** the information in the observations
- We have a gigantic number of possible solutions that are allowed by our data and belief

[3] Bishop 2006, p. 3.3.1.

Linear Regression[3]

How do we put everything together?

- Want to reach the posterior
  - distribution after *all* relevant information have been taken into account
- Prediction should reflect my beliefs in the model **and** the information in the observations
- We have a gigantic number of possible solutions that are allowed by our data and belief

[3] Bishop 2006, p. 3.3.1.

Linear Regression[3]

How do we put everything together?

- Want to reach the posterior
  - distribution after *all* relevant information have been taken into account
- Prediction should reflect my beliefs in the model **and** the information in the observations
- We have a gigantic number of possible solutions that are allowed by our data and belief

[3] Bishop 2006, p. 3.3.1.

Linear Regression[3]

How do we put everything together?

- Want to reach the posterior
  - distribution after *all* relevant information have been taken into account
- Prediction should reflect my beliefs in the model **and** the information in the observations
- We have a gigantic number of possible solutions that are allowed by our data and belief

[3] Bishop 2006, p. 3.3.1.

Linear Regression[3]

How do we put everything together?

- Want to reach the posterior
  - distribution after *all* relevant information have been taken into account
- Prediction should reflect my beliefs in the model **and** the information in the observations
- We have a gigantic number of possible solutions that are allowed by our data and belief

[3] Bishop 2006, p. 3.3.1.

Linear Regression[3]

$$p(\mathbf{W}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{W})p(\mathbf{W})}{p(\mathcal{D})} \tag{18}$$

Evidence

- The denominator shows where the model spreads it probability mass over the data-space (evidence of the model)
- The denominator does not change with $\mathbf{W}$

[3] Bishop 2006, p. 3.3.1.

Linear Regression[3]

$$p(\mathbf{W}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{W})p(\mathbf{W})}{p(\mathcal{D})} \tag{19}$$

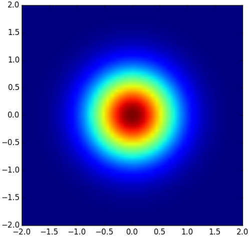
$$p(\mathbf{W}|\mathbf{Y},\mathbf{X}) \propto p(\mathbf{Y}|\mathbf{W},\mathbf{X})p(\mathbf{W}) \tag{20}$$

Evidence

- The denominator shows where the model spreads it probability mass over the data-space (evidence of the model)
- The denominator does not change with $\mathbf{W}$

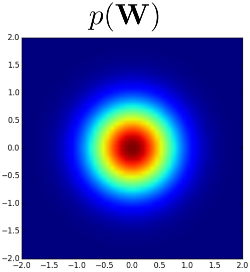
[3] Bishop 2006, p. 3.3.1.

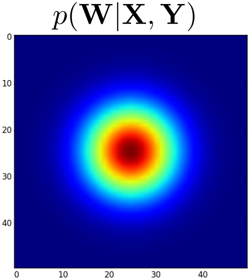
$p(\mathbf{W})$

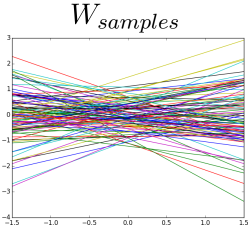


[4]

[4] Bishop 2006, p. 155

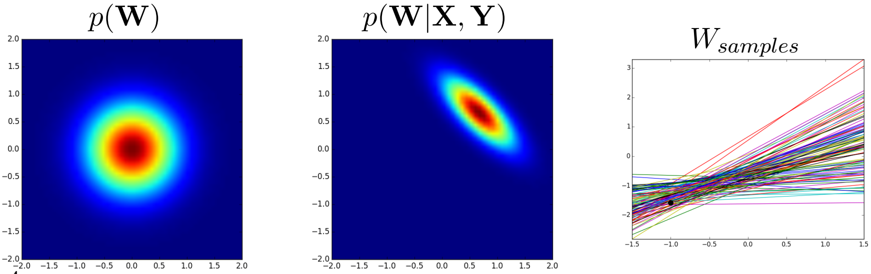






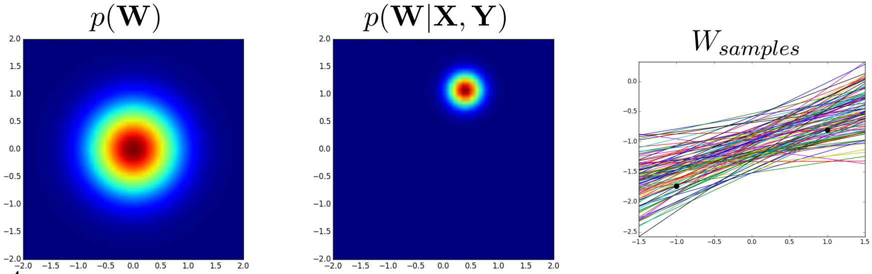
[4]

[4] Bishop 2006, p. 155



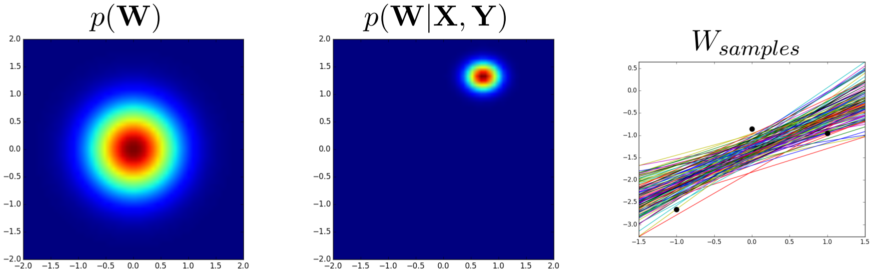
[4]

[4] Bishop 2006, p. 155



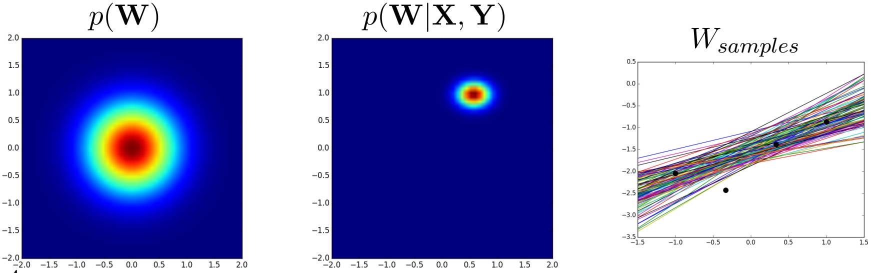
[4]

[4] Bishop 2006, p. 155



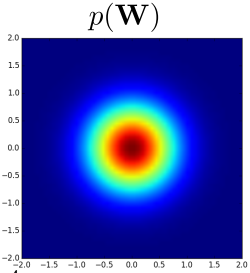
[4]

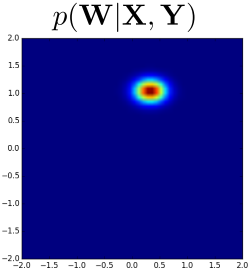
[4] Bishop 2006, p. 155

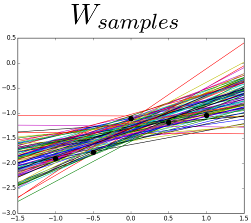


[4]

[4] Bishop 2006, p. 155

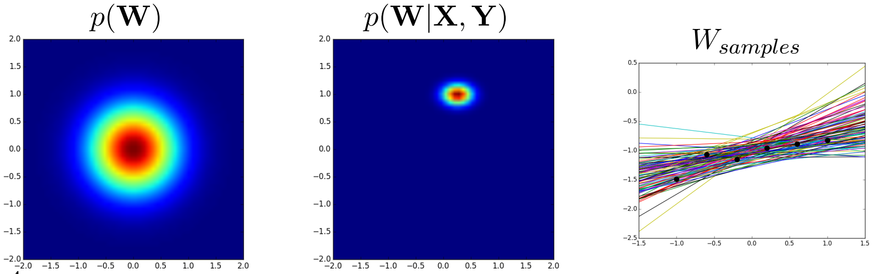






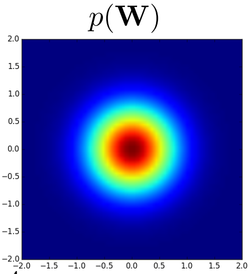
[4]

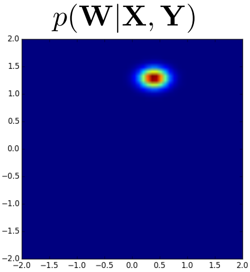
[4] Bishop 2006, p. 155

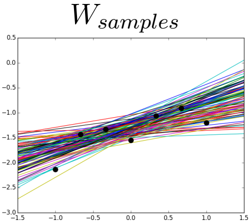


[4]

[4] Bishop 2006, p. 155



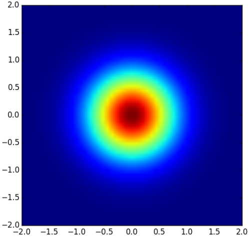




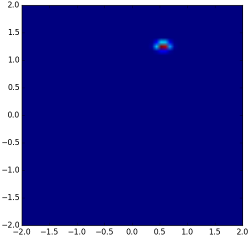
[4]

[4] Bishop 2006, p. 155

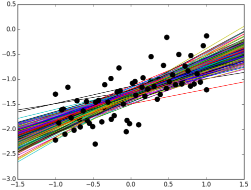
$p(\mathbf{W})$



$p(\mathbf{W}|\mathbf{X},\mathbf{Y})$

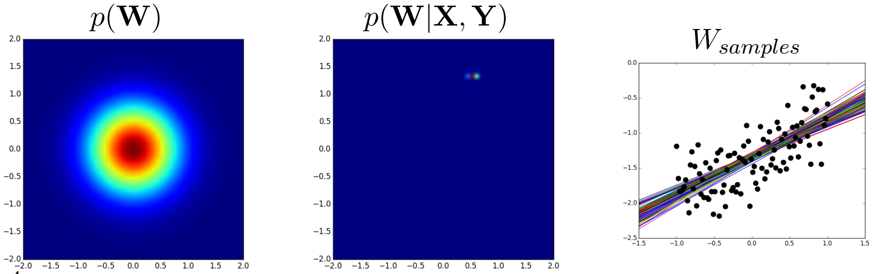


$W_{samples}$



[4]

[4] Bishop 2006, p. 155



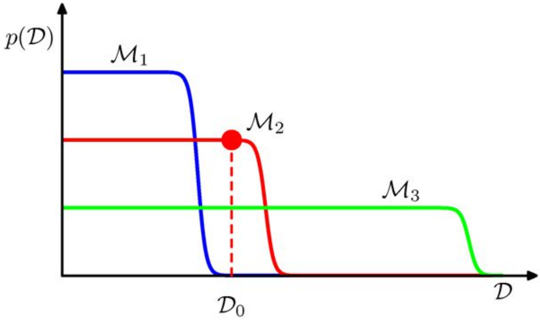
4

[4] Bishop 2006, p. 155

Assignment

You should now be able to do the linear part of Task 2.1 and Task 2.2 of the assignment.

Evidence



[5]

$$p(\mathcal{D}) \tag{21}$$

[5] Bishop 2006, 3.4 p. 163-164

Toolbox

1. Formulate prediction error by likelihood
2. Formulate belief of model in prior
3. Marginalise irrelevant variables
4. Choose model based on *evidence* $p_{\mathcal{M}}(\mathcal{D})$ (Assignment)

Toolbox

1. Formulate prediction error by likelihood
2. Formulate belief of model in prior
3. Marginalise irrelevant variables
4. Choose model based on *evidence* $p_{\mathcal{M}}(\mathcal{D})$ (Assignment)

Toolbox

1. Formulate prediction error by likelihood
2. Formulate belief of model in prior
3. Marginalise irrelevant variables
4. Choose model based on *evidence* $p_{\mathcal{M}}(\mathcal{D})$ (Assignment)

Toolbox

1. Formulate prediction error by likelihood
2. Formulate belief of model in prior
3. Marginalise irrelevant variables
4. Choose model based on *evidence* $p_{\mathcal{M}}(\mathcal{D})$ (Assignment)

Marginalisation

$$p(\mathbf{W}) = \int p(\mathbf{W}|\theta)p(\theta)\mathrm{d}\theta \tag{22}$$

- Average according to belief and how well the model fits the observations
- “Pushes” belief through model

Marginalisation

$$p(\mathbf{W}) = \int p(\mathbf{W}|\theta)p(\theta)\mathrm{d}\theta \tag{23}$$

- Average according to belief and how well the model fits the observations
- "Pushes" belief through model

Marginalisation

Nature laughs at the difficulties of integration

Choosing Distributions

$$p(\mathbf{X}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})}{p(\mathbf{Y})} \tag{24}$$

Conjugate Distributions

- The posterior and the prior are in the same *family*
- Relationship with all **three** terms

[6]

[6] `Wikipedia`

Choosing Distributions

$$p(\mathbf{X}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})}{p(\mathbf{Y})} \tag{25}$$

Conjugate Distributions

- The posterior and the prior are in the same *family*
- Relationship with all **three** terms

Carls intuition

"combining belief in parameters through model should not change the family of the distribution over the parameters"

6

[6] `Wikipedia`

Choosing Distributions

$$p(\mathbf{X}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})}{p(\mathbf{Y})} \tag{26}$$

Remainder of this part

- In this part of the course we will only look at Gaussians
- Gaussians are self-conjugate
  - Gaussian likelihood + Gaussian prior $\Rightarrow$ Gaussian posterior
- On lecture 5 I will show you approximate ways to compute an integral
- Hedvig will look at non-gaussian priors and likelihoods in her part.

Choosing Distributions

$$p(\mathbf{X}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})}{p(\mathbf{Y})} \tag{27}$$

Remainder of this part

- In this part of the course we will only look at Gaussians
- Gaussians are self-conjugate
  - Gaussian likelihood + Gaussian prior ⇒ Gaussian posterior
- On lecture 5 I will show you approximate ways to compute an integral
- Hedvig will look at non-gaussian priors and likelihoods in her part.

Choosing Distributions

$$p(\mathbf{X}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})}{p(\mathbf{Y})} \tag{28}$$

Remainder of this part

- In this part of the course we will only look at Gaussians
- Gaussians are self-conjugate
  - Gaussian likelihood + Gaussian prior $\Rightarrow$ Gaussian posterior
- On lecture 5 I will show you approximate ways to compute an integral
- Hedvig will look at non-gaussian priors and likelihoods in her part.

Reflection

- That was ALL of Machine Learning
- Everything else is just details
  - how to choose model
  - what is the right prior
  - how to integrate
- You will have to approximate and use heuristics but always relate to this

Reflection

- That was ALL of Machine Learning
- Everything else is just details
  - how to choose model
  - what is the right prior
  - how to integrate
- You will have to approximate and use heuristics but always relate to this

Reflection

- That was ALL of Machine Learning
- Everything else is just details
  - how to choose model
  - what is the right prior
  - how to integrate
- You will have to approximate and use heuristics but always relate to this

Reflection

- That was ALL of Machine Learning
- Everything else is just details
  - how to choose model
  - what is the right prior
  - how to integrate
- You will have to approximate and use heuristics but always relate to this

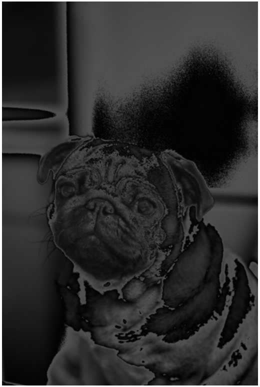
Reflection

- That was ALL of Machine Learning
- Everything else is just details
  - how to choose model
  - what is the right prior
  - how to integrate
- You will have to approximate and use heuristics but always relate to this

Reflection

- That was ALL of Machine Learning
- Everything else is just details
  - how to choose model
  - what is the right prior
  - how to integrate
- You will have to approximate and use heuristics but always relate to this

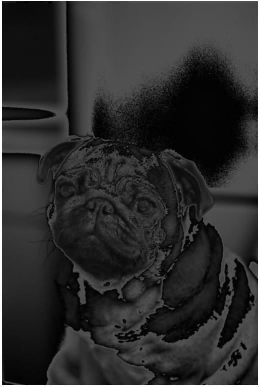
Example: Image restoration[6]



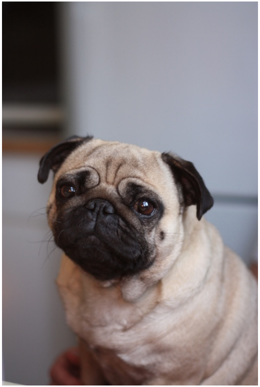
$\leftarrow$

[6] `Lecture1/imageExample.py`

Example: Image restoration[6]



$\leftarrow$



[6] `Lecture1/imageExample.py`

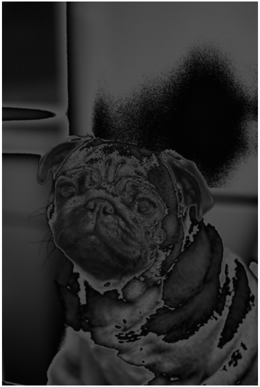
Example: Image restoration[6]

$$p(\mathbf{Y}|\mathbf{X}, \theta) = \mathcal{N}(f(\mathbf{X}), \sigma^2\mathbf{I}) \tag{29}$$

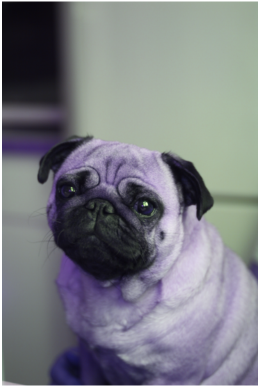
$$\mathbf{y}_i = \frac{1}{3}(\mathbf{x}_i^r + \mathbf{x}_i^g + \mathbf{x}_i^b) \tag{30}$$

[6] `Lecture1/imageExample.py`

Example: Image restoration[6]

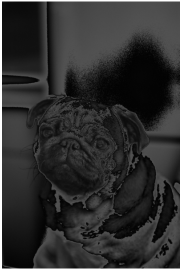


$\leftarrow$



[6] `Lecture1/imageExample.py`

Example: Image restoration[6]



$\leftarrow$

$$p(\mathbf{Y}|\mathbf{X},\theta) = \mathcal{N}(f(\mathbf{X}),\sigma^2\mathbf{I}) \tag{31}$$

$$\mathbf{y}_i = \frac{1}{3}(\mathbf{x}_i^r + \mathbf{x}_i^g + \mathbf{x}_i^b) \tag{32}$$

$$p(\mathbf{X}|\mathbf{Y},\theta) \propto p(\mathbf{Y}|\mathbf{X},\theta)p(\mathbf{X}) \tag{33}$$

[6] `Lecture1/imageExample.py`

Introduction

Regression

Kernel Methods

Dual Linear Regression[7]

$$p(\mathbf{W}|\mathbf{Y},\mathbf{X}) = \frac{p(\mathbf{Y}|\mathbf{W},\mathbf{X})p(\mathbf{W})}{p(\mathbf{Y})} \tag{34}$$

$$p(\mathbf{Y}|\mathbf{W},\mathbf{X}) = \prod_{i}^{N} p(\mathbf{y}_i|\mathbf{W},\mathbf{X}) = \prod_{i}^{N} \mathcal{N}(\mathbf{y}_i|\cdot,\sigma^2\mathbf{I}) \tag{35}$$

$$p(\mathbf{W}) = \mathcal{N}(\mathbf{0},\tau^2\mathbf{I}) \tag{36}$$

[7] Bishop 2006, p. 6.1.

Dual Linear Regression[7]

$$p(\mathbf{W}|\mathbf{Y},\mathbf{X}) = \frac{p(\mathbf{Y}|\mathbf{W},\mathbf{X})p(\mathbf{W})}{p(\mathbf{Y})} \tag{37}$$

$$p(\mathbf{Y}|\mathbf{W},\mathbf{X}) = \prod_{i}^{N} p(\mathbf{y}_i|\mathbf{W},\mathbf{X}) = \prod_{i}^{N} \mathcal{N}(\mathbf{y}_i|\cdot,\sigma^2\mathbf{I}) \tag{38}$$

$$p(\mathbf{W}) = \mathcal{N}(\mathbf{0},\tau^2\mathbf{I}) \tag{39}$$

$$p(\mathbf{W}|\mathbf{Y},\mathbf{X}) \propto p(\mathbf{Y}|\mathbf{W},\mathbf{X})p(\mathbf{W}) \tag{40}$$

[7] Bishop 2006, p. 6.1.

Dual Linear Regression[7]

- Lets look at a simple 1D problem

$$\mathbf{y} \in \mathbb{R}^{1\times N} \tag{41}$$

$$\mathbf{x} \in \mathbb{R}^{1\times N} \tag{42}$$

[7] Bishop 2006, p. 6.1.

Dual Linear Regression[7]

$$p(\mathbf{w}|\mathbf{y},\mathbf{X}) \propto \prod_{i}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - y_i)^{\mathrm{T}}(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - y_i)} \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{1}{2\tau^2}(\mathbf{w}^{\mathrm{T}}\mathbf{w})} \tag{43}$$

$$= \frac{1}{(\sqrt{2\pi\sigma^2})^N} e^{-\frac{1}{2\sigma^2}(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{y})^{\mathrm{T}}(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{y})} \frac{1}{(\sqrt{2\pi\tau^2})^N} e^{-\frac{1}{2\tau^2}(\mathbf{w}^{\mathrm{T}}\mathbf{w})} \tag{44}$$

Objective

- Want to find the parameters that maximises the above
- Logarithm is monotonic
- Minimise negative logarithm of $p(\mathbf{w}|\mathbf{y},\mathbf{X})$

Dual Linear Regression[7]

$$p(\mathbf{w}|\mathbf{y},\mathbf{X}) \propto \prod_i^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - y_i)^{\mathrm{T}}(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - y_i)} \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{1}{2\tau^2}(\mathbf{w}^{\mathrm{T}}\mathbf{w})} \tag{45}$$

$$= \frac{1}{(\sqrt{2\pi\sigma^2})^N} e^{-\frac{1}{2\sigma^2}(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{y})^{\mathrm{T}}(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{y})} \frac{1}{(\sqrt{2\pi\tau^2})^N} e^{-\frac{1}{2\tau^2}(\mathbf{w}^{\mathrm{T}}\mathbf{w})} \tag{46}$$

Objective

- Want to find the parameters that maximises the above
- Logarithm is monotonic
- Minimise negative logarithm of $p(\mathbf{w}|\mathbf{y},\mathbf{X})$

Dual Linear Regression[7]

$$J(\mathbf{w}) = \frac{1}{2}(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{y})^{\mathrm{T}}(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{y}) + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \tag{47}$$

Objective

- Want to find the parameters that maximises the above
- Logarithm is monotonic
- Minimise negative logarithm of $p(\mathbf{w}|\mathbf{y}, \mathbf{X})$

[7] Bishop 2006, p. 6.1.

Dual Linear Regression[7]

$$J(\mathbf{w}) = \frac{1}{2}(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{y})^{\mathrm{T}}(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{y}) + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \tag{48}$$

$$\frac{\delta}{\delta \mathbf{w}} J(\mathbf{w}) = \frac{1}{2} 2\mathbf{X}^{\mathrm{T}}(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{y}) + \frac{\lambda}{2} 2\mathbf{w} \tag{49}$$

Optimisation

- Lets make a point-estimate
- Pick $\mathbf{w}$ that minimises $J(\mathbf{w})$

[7] Bishop 2006, p. 6.1.

Dual Linear Regression[7]

$$J(\mathbf{w}) = \frac{1}{2}(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{y})^{\mathrm{T}}(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{y}) + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \tag{50}$$

$$\frac{\delta}{\delta\mathbf{w}}J(\mathbf{w}) = \frac{1}{2}2\mathbf{X}^{\mathrm{T}}(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{y}) + \frac{\lambda}{2}2\mathbf{w} \tag{51}$$

$$\mathbf{w} = -\frac{1}{\lambda}\mathbf{X}^{\mathrm{T}}(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{y}) = \tag{52}$$

Optimisation

- Lets make a point-estimate
- Pick $\mathbf{w}$ that minimises $J(\mathbf{w})$

[7] Bishop 2006, p. 6.1.

Dual Linear Regression[7]

$$J(\mathbf{w}) = \frac{1}{2}(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{y})^{\mathrm{T}}(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{y}) + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \tag{53}$$

$$\frac{\delta}{\delta \mathbf{w}} J(\mathbf{w}) = \frac{1}{2} 2\mathbf{X}^{\mathrm{T}}(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{y}) + \frac{\lambda}{2} 2\mathbf{w} \tag{54}$$

$$\mathbf{w} = -\frac{1}{\lambda}\mathbf{X}^{\mathrm{T}}(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{y}) = \mathbf{X}^{\mathrm{T}}\mathbf{a} = \sum_{n}^{N} \alpha_n \mathbf{x}_n \tag{55}$$

Optimisation

- Lets make a point-estimate
- Pick $\mathbf{w}$ that minimises $J(\mathbf{w})$

[7] Bishop 2006, p. 6.1.

Dual Linear Regression[7]

$$J(\mathbf{w}) = \frac{1}{2}(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{y})^{\mathrm{T}}(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{y}) + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \tag{56}$$

$$\mathbf{w} = \mathbf{X}^{\mathrm{T}}\mathbf{a} \tag{57}$$

Formulate Dual

$$J(\mathbf{a}) = \frac{1}{2}\mathbf{a}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{a} - \mathbf{a}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{y} + \frac{\lambda}{2}\mathbf{a}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{a} \tag{58}$$

[7] Bishop 2006, p. 6.1.

Dual Linear Regression[7]

$$[\mathbf{K}]_{ij} = \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j \tag{59}$$

$$J(\mathbf{a}) = \frac{1}{2}\mathbf{a}^{\mathrm{T}}\mathbf{K}\mathbf{K}\mathbf{a} - \mathbf{a}\mathbf{K}\mathbf{y} + \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{y} + \frac{\lambda}{2}\mathbf{a}^{\mathrm{T}}\mathbf{K}\mathbf{a} \tag{60}$$

[7] Bishop 2006, p. 6.1.

Dual Linear Regression[7]

$$[\mathbf{K}]_{ij} = \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j \tag{61}$$

$$J(\mathbf{a}) = \frac{1}{2}\mathbf{a}^{\mathrm{T}}\mathbf{K}\mathbf{K}\mathbf{a} - \mathbf{a}\mathbf{K}\mathbf{y} + \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{y} + \frac{\lambda}{2}\mathbf{a}^{\mathrm{T}}\mathbf{K}\mathbf{a} \tag{62}$$

$$\alpha_i = -\frac{1}{\lambda}(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - y_i) \tag{63}$$

$$\mathbf{w} = \sum_i^N \alpha_i \mathbf{x}_i = \mathbf{X}^{\mathrm{T}}\mathbf{a} \tag{64}$$

$$\Rightarrow \mathbf{a} = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y} \tag{65}$$

[7] Bishop 2006, p. 6.1.

Dual Linear Regression[7]

$$[\mathbf{K}]_{ij} = \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j \tag{66}$$

$$J(\mathbf{a}) = \frac{1}{2}\mathbf{a}^{\mathrm{T}}\mathbf{K}\mathbf{K}\mathbf{a} - \mathbf{a}\mathbf{K}\mathbf{y} + \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{y} + \frac{\lambda}{2}\mathbf{a}^{\mathrm{T}}\mathbf{K}\mathbf{a} \tag{67}$$

$$\mathbf{a} = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y} \tag{68}$$

$$\mathbf{y}(\mathbf{x}_*) = \mathbf{w}^{\mathrm{T}}\mathbf{x}_* = \mathbf{a}^{\mathrm{T}}\mathbf{X}\mathbf{x}_* = \mathbf{a}^{\mathrm{T}}k(\mathbf{X}, \mathbf{x}_*) = \tag{69}$$

$$= ((\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y})^{\mathrm{T}}k(\mathbf{X}, \mathbf{x}_*) = k(\mathbf{x}_*, \mathbf{X})(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y} \tag{70}$$

[7] Bishop 2006, p. 6.1.

Dual Linear Regression[7]

Linear Regression

1. See data $(\mathbf{x}_i, y)_i^N$
2. Encode relationship in parameter $\mathbf{W}$
3. Throw training away data
4. Make predictions using $\mathbf{W}$

[7] Bishop 2006, p. 6.1.

Dual Linear Regression[7]

Linear Regression

1. See data $(\mathbf{x}_i, y)_i^N$
2. Encode relationship in parameter $\mathbf{W}$
3. Throw training away data
4. Make predictions using $\mathbf{W}$

Dual

- Do **NOT** throw away data
- Make predictions using relationship to training data
- Model complexity depends on data (i.e. it adapts)
- Non parametric regression

[7]Bishop 2006, p 6.1

Dual Linear Regression[7]

Linear Regression

1. See data $(\mathbf{x}_i, y)_i^N$
2. Encode relationship in parameter $\mathbf{W}$
3. Throw training away data
4. Make predictions using $\mathbf{W}$

Dual

- Do **NOT** throw away data
- Make predictions using relationship to training data
- Model complexity depends on data (i.e. it adapts)
- Non parametric regression

[7] Bishop 2006, p 6.1

Kernels

- Dual linear regression allows us to write everything in terms of inner products
  - we do not *need* representation $\mathbf{x}_i$
- What if we map data prior to regression?

$$\phi : \mathbf{x}_i \rightarrow \mathbf{f}_i$$
$$\mathbf{y}(\mathbf{x}_*) = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_*) = \mathbf{a}^{\mathrm{T}}\phi(\mathbf{X})\phi(\mathbf{x}_*) = k(\mathbf{x}_*, \mathbf{X})(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y}$$
$$k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^{\mathrm{T}}\phi(\mathbf{x}') \tag{71}$$

- *In dual case we do not need to know $\phi(\cdot)$ only $\phi(\cdot)^T\phi(\cdot)$*

Kernels

- Dual linear regression allows us to write everything in terms of inner products
  - we do not *need* representation $\mathbf{x}_i$
- What if we map data prior to regression?

$$\phi : \mathbf{x}_i \rightarrow \mathbf{f}_i$$
$$\mathbf{y}(\mathbf{x}_*) = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_*) = \mathbf{a}^{\mathrm{T}}\phi(\mathbf{X})\phi(\mathbf{x}_*) = k(\mathbf{x}_*, \mathbf{X})(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y}$$
$$k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^{\mathrm{T}}\phi(\mathbf{x}') \tag{72}$$

- *In dual case we do not need to know $\phi(\cdot)$ only $\phi(\cdot)^T\phi(\cdot)$*

Kernels

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^{\mathrm{T}}\phi(\mathbf{x}_j) = ||\phi(\mathbf{x}_i)||||\phi(\mathbf{x}_j)||\cos(\theta) \tag{73}$$

Kernel Functions

- A function that describes an inner product
- Sub-class of functions
  - think triangle in-equality
- If we have $k(\cdot,\cdot)$ we *never* have to know the mapping

Kernels

$$
\begin{aligned}
\mathbf{x} &\in \mathbb{R}^2 && (74)\\
(\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j)^2 &= (x_{i1}x_{j1} + x_{i2}x_{j2})^2 = && (75)\\
&= x_{i1}^2 x_{j1}^2 + 2x_{i1}x_{j1}x_{i2}x_{j2} + x_{i2}^2 x_{j2}^2 = && (76)\\
&= (x_{i1}^2, \sqrt{2}x_{i1}x_{i2}, x_{i2}^2)(x_{j1}^2, \sqrt{2}x_{j1}x_{j2}, x_{j2}^2)^{\mathrm{T}} = && (77)\\
&= \phi(\mathbf{x}_i)^{\mathrm{T}}\phi(\mathbf{x}_j) && (78)
\end{aligned}
$$

[8]

[8] Bishop 2006, p. 6.2

Kernels

$$\mathbf{x} \in \mathbb{R}^2 \tag{79}$$

$$(\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j)^2 = (x_{i1}x_{j1} + x_{i2}x_{j2})^2 = \tag{80}$$

$$= x_{i1}^2 x_{j1}^2 + 2x_{i1}x_{j1}x_{i2}x_{j2} + x_{i2}^2 x_{j2}^2 = \tag{81}$$

$$= (x_{i1}^2, \sqrt{2}x_{i1}x_{i2}, x_{i2}^2)(x_{j1}^2, \sqrt{2}x_{j1}x_{j2}, x_{j2}^2)^{\mathrm{T}} = \tag{82}$$

$$= \phi(\mathbf{x}_i)^{\mathrm{T}}\phi(\mathbf{x}_j) \tag{83}$$

So $k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j)^2$ is a kernel of the mapping $\phi(\mathbf{x}) = ((\mathbf{e}_1^{\mathrm{T}}\mathbf{x})^2, \sqrt{2}\mathbf{e}_1^{\mathrm{T}}\mathbf{x}\mathbf{e}_2^{\mathrm{T}}\mathbf{x}, (\mathbf{e}_2^{\mathrm{T}}\mathbf{x})^2)$

8

[8] Bishop 2006, p. 6.2

The benefits of Kernels

- Kernels allows for *implicit* feature mappings
  - We do **NOT** need to know the feature space
  - The space can have infinite dimensionality
  - The mapping can be non-linear but the problem is still linear!
  - Allows for putting weird things like, strings (DNA) in a vector space
  - More next lecture, these things are very powerful

The benefits of Kernels

- Kernels allows for *implicit* feature mappings
  - We do **NOT** need to know the feature space
  - The space can have infinite dimensionality
  - The mapping can be non-linear but the problem is still linear!
  - Allows for putting weird things like, strings (DNA) in a vector space
  - More next lecture, these things are very powerful

The benefits of Kernels

- Kernels allows for *implicit* feature mappings
  - We do **NOT** need to know the feature space
  - The space can have infinite dimensionality
  - The mapping can be non-linear but the problem is still linear!
  - Allows for putting weird things like, strings (DNA) in a vector space
  - More next lecture, these things are very powerful

The benefits of Kernels

- Kernels allows for *implicit* feature mappings
  - We do **NOT** need to know the feature space
  - The space can have infinite dimensionality
  - The mapping can be non-linear but the problem is still linear!
  - Allows for putting weird things like, strings (DNA) in a vector space
  - More next lecture, these things are very powerful

The benefits of Kernels

- Kernels allows for *implicit* feature mappings
  - We do **NOT** need to know the feature space
  - The space can have infinite dimensionality
  - The mapping can be non-linear but the problem is still linear!
  - Allows for putting weird things like, strings (DNA) in a vector space
  - More next lecture, these things are very powerful

Next Time

Lecture 2

- November 5th 13-15 M2
- Continue with Kernels
  - relation to co-variance
- Non-parametric Regression
  - Gaussian Processes
- Start Assignment

Next Time

Lecture 2

- November 5th 13-15 M2
- Continue with Kernels
  - relation to co-variance
- Non-parametric Regression
  - Gaussian Processes
- Start Assignment

e.o.f.

References I

- Pierre Simon Laplace. *A philosophical essay on probabilities*. 1902.
- E T Jaynes. *Probability theory: The logic of science*. Ed. by G Larry Bretthorst. Cambridge university press, June 2003.
- Christopher M Bishop. *Pattern recognition and machine learning*. 2006.