DD2434 - Advanced Machine Learning

Regression

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

Royal Institute of Technology

November 19, 2014

What have you seen up till now?

- (In)-dependency structures
- Language of Graphical Models
- Mixture Models
- Sequential models

PREVIOUSLY

Whats the focus of this part of the course

My plan

- My view on Machine Learning
- Look at each part of a probabilistic model in detail
  - how do they interact
  - what do they provide
- Different models
  - parametric
  - non-parameteric
  - hierarchical
- You should translate what you have seen in Jens part to this
- Really simple data: *"as there is no free lunch lets avoid eating"*

Whats the focus of this part of the course

My plan

- My view on Machine Learning
- Look at each part of a probabilistic model in detail
  - how do they interact
  - what do they provide
- Different models
  - parametric
  - non-parameteric
  - hierarchical
- You should translate what you have seen in Jens part to this
- Really simple data: *"as there is no free lunch lets avoid eating"*

Whats the focus of this part of the course

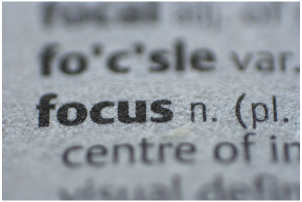
Theme

"How I can incorporate my knowledge/belief with observations such that when I see data it reduces my uncertainty according to the evidence provided in the observations."

Whats the focus of this part of the course

Structure

- 4 Lectures
- 2 Practical sessions
- 1 Assignment
  - Deadline December 3rd
  - Review December 4th and 5th



Assignment

- Three parts aligned with lectures
- Part 1 (Lecture 6 & 7)
  - Task: probabilistic regression
  - Aim: understand probabilistic objects
- Part 2 (Lecture 7 & 8)
  - 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 6 & 7)
  - Task: probabilistic regression
  - Aim: understand probabilistic objects
- Part 2 (Lecture 7 & 8)
  - 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 6 & 7)
  - Task: probabilistic regression
  - Aim: understand probabilistic objects
- Part 2 (Lecture 7 & 8)
  - 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 6 & 7)
  - Task: probabilistic regression
  - Aim: understand probabilistic objects
- Part 2 (Lecture 7 & 8)
  - 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

My view on Machine Learning

1

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] URL

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] URL

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] URL

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] URL

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] URL

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] URL

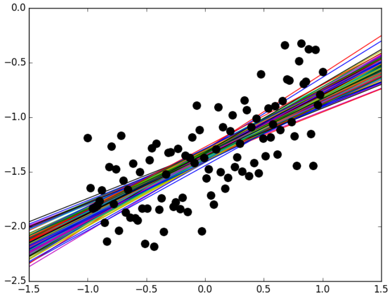
Introduction

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[2]

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)

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

[2]Murphy 2012, p. 2.2.7.

Probability Basics[2]

Variance

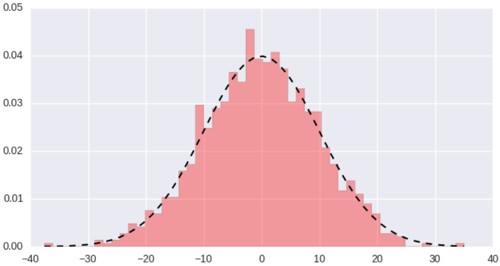
$$\sigma^2(\mathbf{x}) = \mathrm{var}[\mathbf{x}] = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])^2] \tag{3}$$

- Shows the "spread" of a distribution
- Sample variance

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

[2] Murphy 2012, p. 2.2.7.

Probability Basics[3]

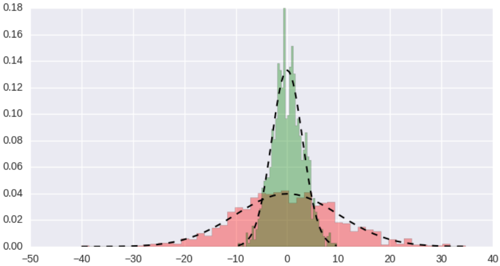


2

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

[3] Murphy 2012, p. 2.2.7.

Probability Basics[3]

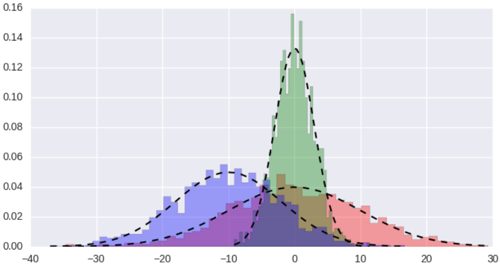


2

[2] Matplotlib3D, /Lecture1/probBasics.py

[3] Murphy 2012, p. 2.2.7.

Probability Basics[3]



2

[2] `Matplotlib3D, /Lecture1/probBasics.py`
[3] Murphy 2012, p. 2.2.7.

Probability Basics[2]

Covariance

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

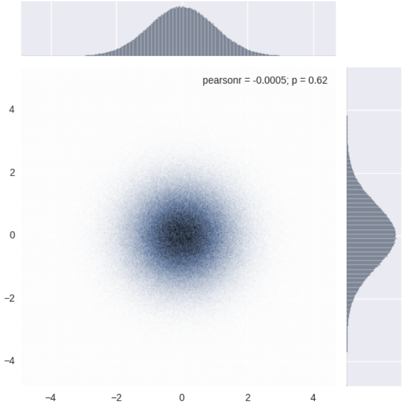
$$[\sigma(\mathbf{X}, \mathbf{Y})]_{ij} = \sigma(\mathbf{x}_i, \mathbf{y}_j) = k(\mathbf{x}_i, \mathbf{y}_j) \tag{6}$$

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

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

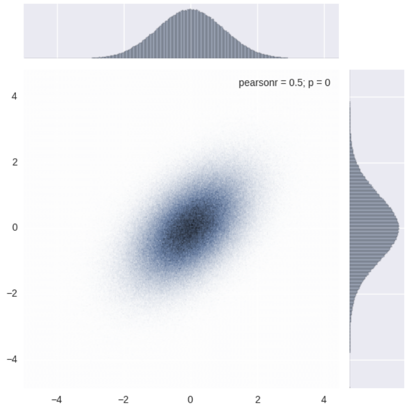
[2] Murphy 2012, p. 2.2.7.

Probability Basics[2]



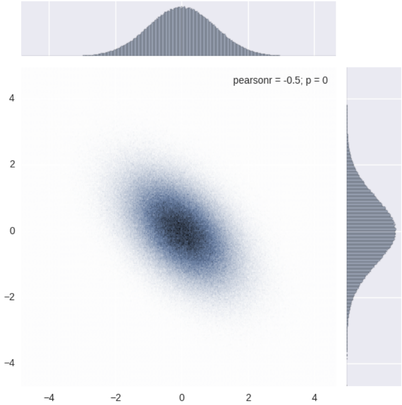
[2] Murphy 2012, p. 2.2.7.

Probability Basics[2]



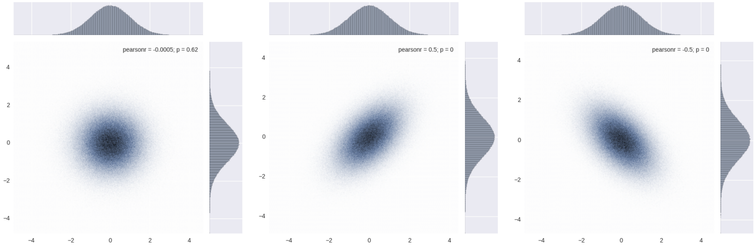
[2] Murphy 2012, p. 2.2.7.

Probability Basics[2]



[2] Murphy 2012, p. 2.2.7.

Probability Basics[3]



[2]

[2] `Matplotlib3D`, `/Lecture1/probBasics.py`
[3] Murphy 2012, p. 2.2.7.

Linear Regression[4]

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

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?

[4] Murphy 2012, Ch 7.

Linear Regression[4]

$$\mathbf{y}_i = \mathbf{W}\mathbf{x}_i + \epsilon \tag{9}$$

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?

[4] Murphy 2012, Ch 7.

Linear Regression[4]

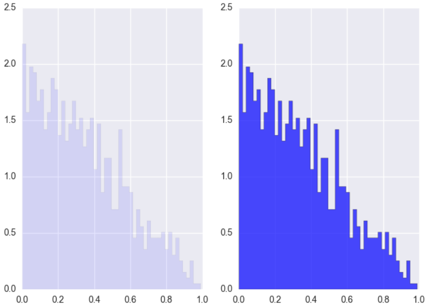
Why always Gaussians?

- 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] Murphy 2012, Sec. 2.6.3

[4] Murphy 2012, Ch 7.

Linear Regression[5]



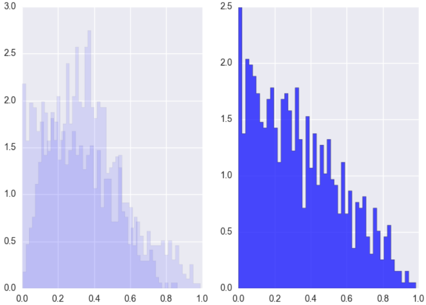
4

---

[4] `/Lecture1/centralLimit.py`

[5] Murphy 2012, Ch 7.

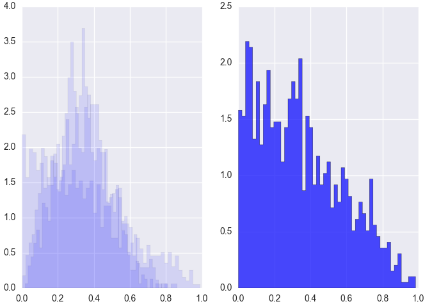
Linear Regression[5]



4

[4] `/Lecture1/centralLimit.py`
[5] Murphy 2012, Ch 7.

Linear Regression[5]

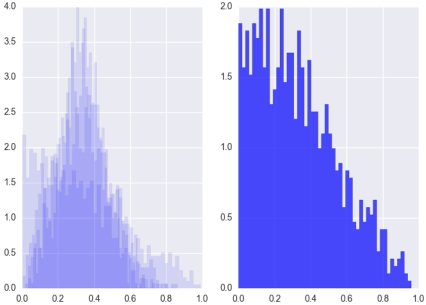


4

[4] /Lecture1/centralLimit.py

[5] Murphy 2012, Ch 7.

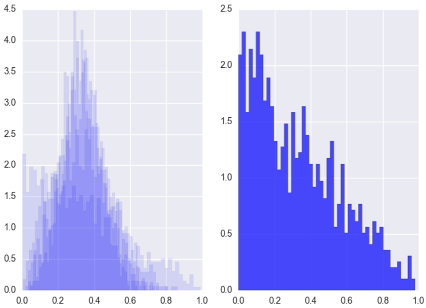
Linear Regression[5]



4

[4] `/Lecture1/centralLimit.py`
[5] Murphy 2012, Ch 7.

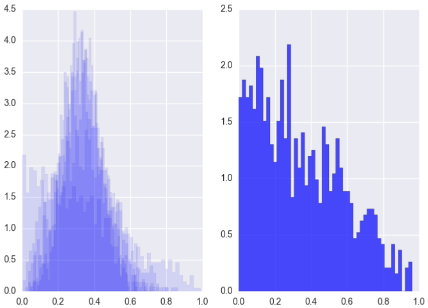
Linear Regression[5]



4

[4] `/Lecture1/centralLimit.py`
[5] Murphy 2012, Ch 7.

Linear Regression[5]

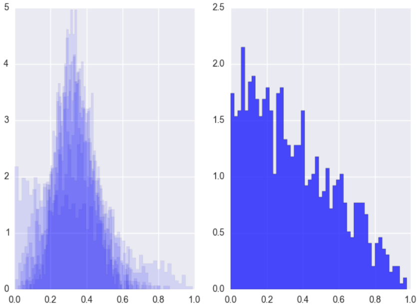


4

[4] `/Lecture1/centralLimit.py`

[5] Murphy 2012, Ch 7.

Linear Regression[5]



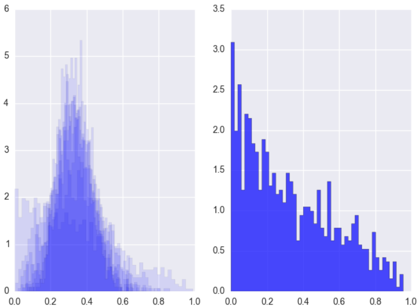
4

---

[4] /Lecture1/centralLimit.py

[5] Murphy 2012, Ch 7.

Linear Regression[5]

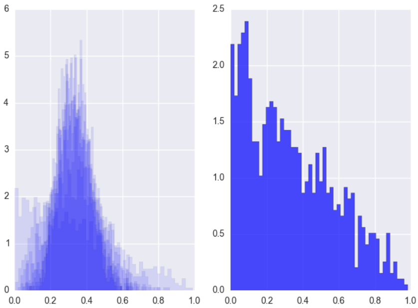


4

[4] `/Lecture1/centralLimit.py`

[5] Murphy 2012, Ch 7.

Linear Regression[5]

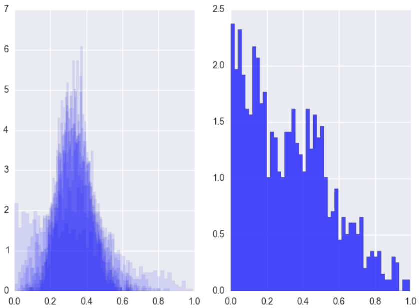


4

[4] `/Lecture1/centralLimit.py`

[5] Murphy 2012, Ch 7.

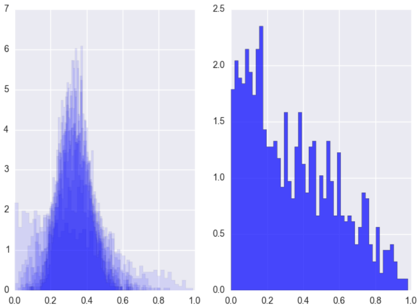
Linear Regression[5]



4

[4] `/Lecture1/centralLimit.py`
[5] Murphy 2012, Ch 7.

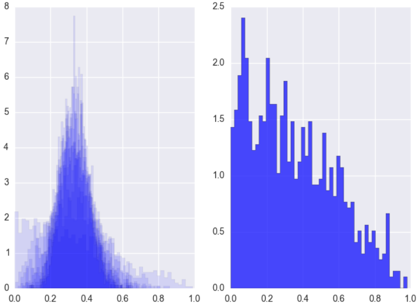
Linear Regression[5]



4

[4] `/Lecture1/centralLimit.py`
[5] Murphy 2012, Ch 7.

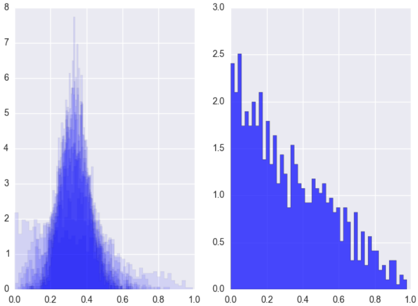
Linear Regression[5]



4

[4] `/Lecture1/centralLimit.py`
[5] Murphy 2012, Ch 7.

Linear Regression[5]

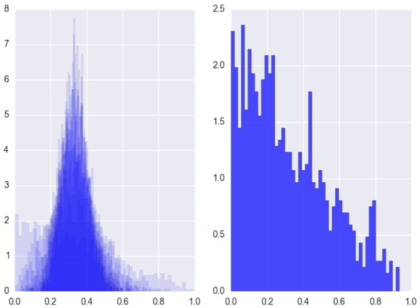


4

---

[4] `/Lecture1/centralLimit.py`
[5] Murphy 2012, Ch 7.

Linear Regression[5]

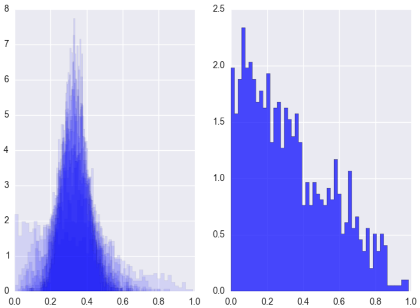


4

[4] `/Lecture1/centralLimit.py`

[5] Murphy 2012, Ch 7.

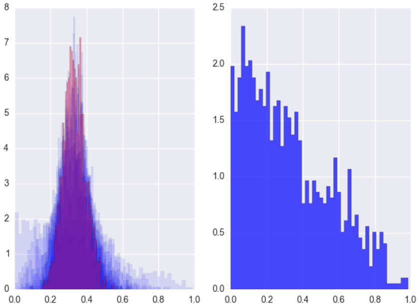
Linear Regression[5]



4

[4] `/Lecture1/centralLimit.py`
[5] Murphy 2012, Ch 7.

Linear Regression[5]



4

[4] `/Lecture1/centralLimit.py`
[5] Murphy 2012, Ch 7.

Linear Regression[5]

URL

4

[4] /Lecture1/centralLimit.py
[5] Murphy 2012, Ch 7.

Linear Regression[4]

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

Uncertainty in Model

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

[4] Murphy 2012, Ch 7.

Linear Regression[4]

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

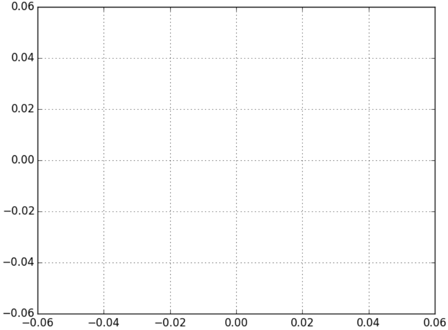
Belief about model before seeing data

- Prior
- What do I know about the regression parameters
- Swear word of the day: "Empirical Bayes"

[4] Murphy 2012, Ch 7.

Linear Regression[4]

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



Linear Regression[4]

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

Belief about model before seeing data

- Prior
- What do I know about the regression parameters

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

- Swear word of the day: "Empirical Bayes"

[4] Murphy 2012, Ch 7.

Linear Regression[4]

$$p(\mathbf{W}) \tag{15}$$

Belief about model before seeing data

- Prior
- What do I know about the regression parameters
- Swear word of the day: “Empirical Bayes”

[4]Murphy 2012, Ch 7.

Linear Regression[4]

$$p(\mathbf{y}_i|\mathbf{W},\mathbf{x}_i) \tag{16}$$

How well does my model predict the data

- Likelihood
- Think error function but also how different errors

$$p(\mathbf{y}_i|\mathbf{W},\mathbf{x}_i) = \mathcal{N}(\mathbf{y}_i|\mathbf{W}\mathbf{x}_i,\tau^2\mathbf{I}) \tag{17}$$

[4] Murphy 2012, Ch 7.

Linear Regression[4]

Structure

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

$$p(\mathbf{Y}|\mathbf{W},\mathbf{X}) = \prod_i^N p(\mathbf{y}_i|\mathbf{W},\mathbf{x}_i) \tag{18}$$

[4] Murphy 2012, Ch 7.

Linear Regression[4]

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
- How about a weighted combination?

[4] Murphy 2012, Ch 7.

Linear Regression[4]

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
- How about a weighted combination?

[4] Murphy 2012, Ch 7.

Linear Regression[4]

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
- How about a weighted combination?

[4] Murphy 2012, Ch 7.

Linear Regression[4]

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
- How about a weighted combination?

[4] Murphy 2012, Ch 7.

Linear Regression[4]

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
- How about a weighted combination?

[4] Murphy 2012, Ch 7.

Linear Regression[4]

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

$$p(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{W})p(\mathbf{W})\mathrm{d}\mathbf{W} \tag{20}$$

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 $\mathbf{W}$

[4] Murphy 2012, Ch 7.

Linear Regression[4]

$$p(\mathbf{W}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{W})p(\mathbf{W})}{p(\mathcal{D})} \tag{21}$$

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

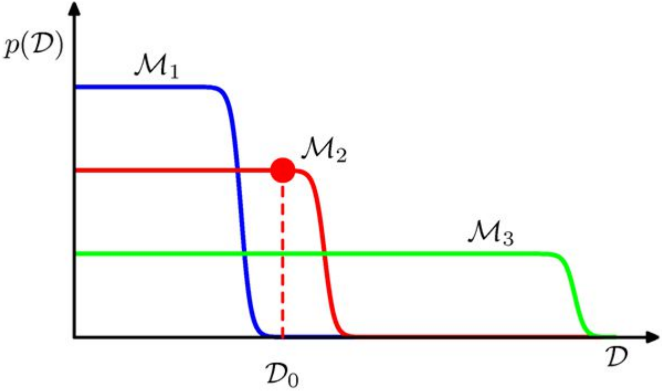
$$p(\mathbf{W}|\mathbf{Y},\mathbf{X}) \propto p(\mathbf{Y}|\mathbf{W},\mathbf{X})p(\mathbf{W}) \tag{23}$$

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 **W**

[4] Murphy 2012, Ch 7.

Linear Regression[5]



[4]

[4] Murphy 2012, p. 5.3.1
[5] Murphy 2012, Ch 7.

Linear Regression cont.

Toolbox

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

Linear Regression cont.

Toolbox

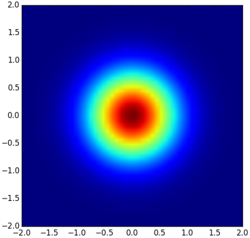
1. Formulate prediction error by likelihood
2. Formulate belief of model in prior
3. Choose model based on *evidence* $p_{\mathcal{M}}(\mathcal{D})$ (Assignment)

Linear Regression cont.

Toolbox

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

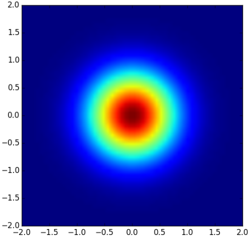
$p(\mathbf{W})$



[6]

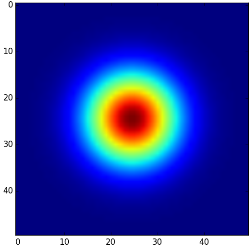
[6] Murphy 2012, p. 7.6.1

$p(\mathbf{W})$

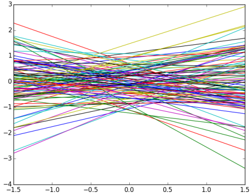


[6]

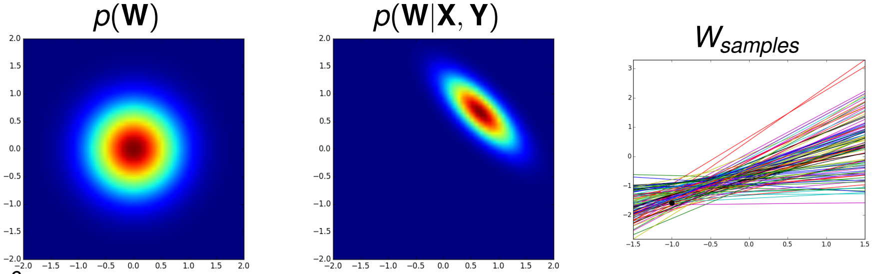
$p(\mathbf{W}|\mathbf{X}, \mathbf{Y})$



$W_{samples}$

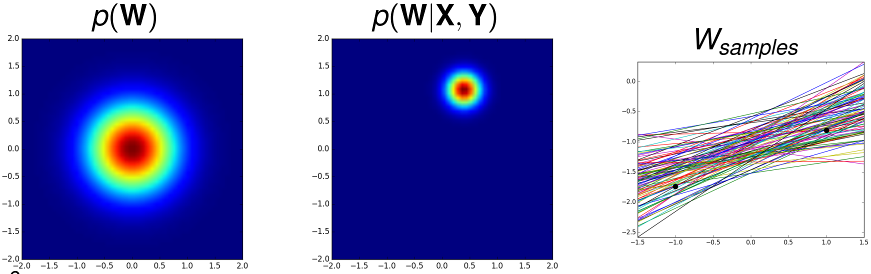


[6] Murphy 2012, p. 7.6.1



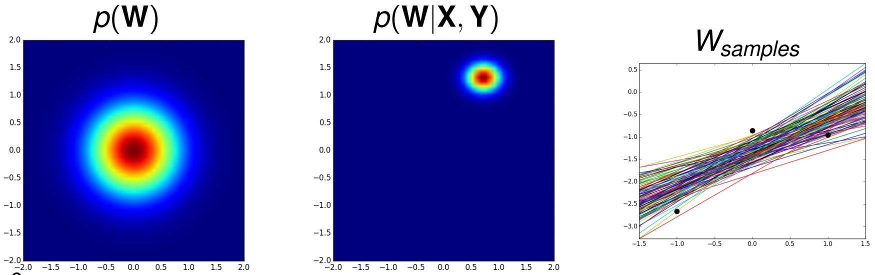
[6]

[6] Murphy 2012, p. 7.6.1



[6]

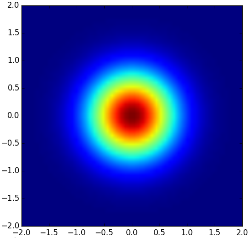
[6] Murphy 2012, p. 7.6.1



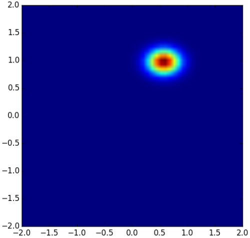
[6]

[6] Murphy 2012, p. 7.6.1

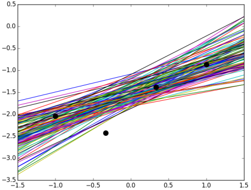
$p(\mathbf{W})$



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

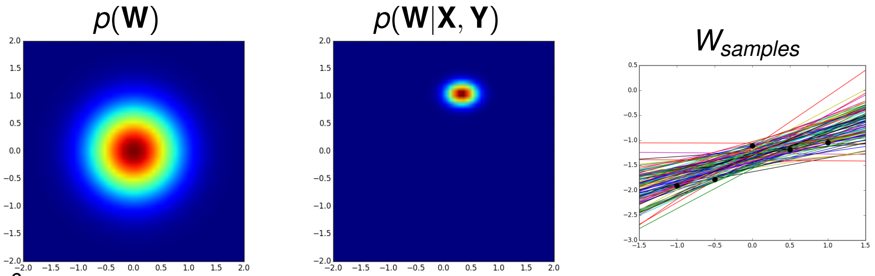


$W_{samples}$



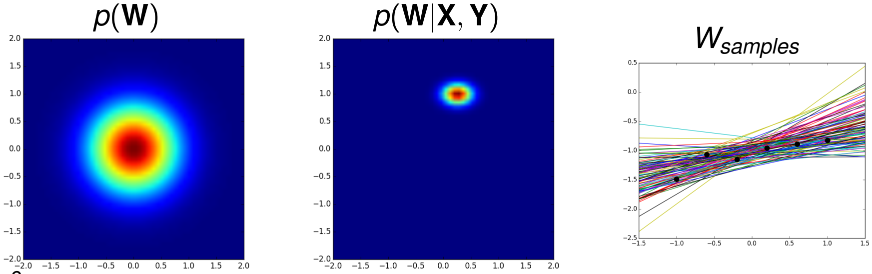
[6]

[6] Murphy 2012, p. 7.6.1



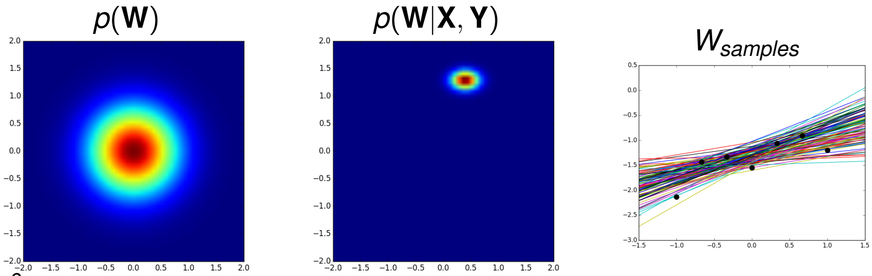
[6]

[6]Murphy 2012, p. 7.6.1



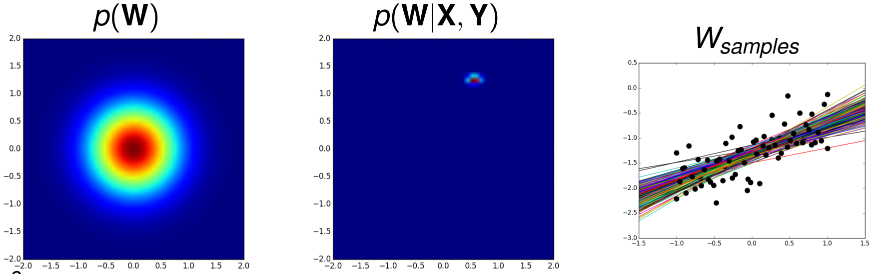
[6]

[6] Murphy 2012, p. 7.6.1



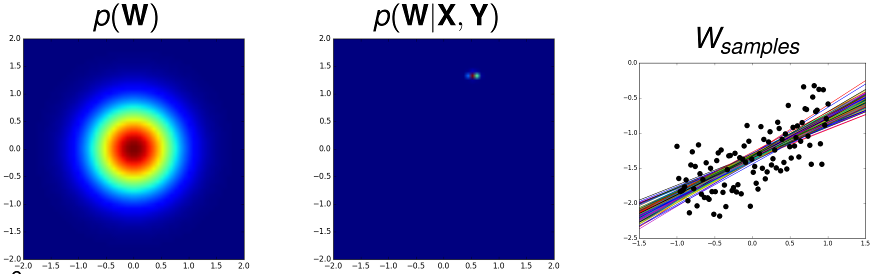
[6]

[6] Murphy 2012, p. 7.6.1



[6]

[6] Murphy 2012, p. 7.6.1



[6]

[6] Murphy 2012, p. 7.6.1

Assignment

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

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

$$p(\mathbf{W}) = \int p(\mathbf{W}|\theta)p(\theta)\mathrm{d}\theta \tag{24}$$

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

Marginalisation

$$p(\mathbf{W}) = \int p(\mathbf{W}|\theta)p(\theta)\mathrm{d}\theta \tag{25}$$

- 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

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

Conjugate Distributions

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

[7]

[7] Wikipedia

Choosing Distributions

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

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"

7

[7] Wikipedia

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 practical 4 I will show you approximate ways to compute an integral
- Hedvig will look at Dirichlet priors where you will see other combinations which are conjugate

Choosing Distributions

$$p(\mathbf{X}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})}{p(\mathbf{Y})} \tag{29}$$

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 practical 4 I will show you approximate ways to compute an integral
- Hedvig will look at Dirichlet priors where you will see other combinations which are conjugate

Choosing Distributions

$$p(\mathbf{X}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})}{p(\mathbf{Y})} \tag{30}$$

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 practical 4 I will show you approximate ways to compute an integral
- Hedvig will look at Dirichlet priors where you will see other combinations which are conjugate

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
- This is the beauty of being Bayesian

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
- This is the beauty of being Bayesian

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
- This is the beauty of being Bayesian

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
- This is the beauty of being Bayesian

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
- This is the beauty of being Bayesian

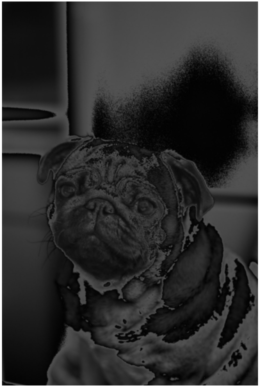
Reflection

- 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
- This is the beauty of being Bayesian

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
- This is the beauty of being Bayesian

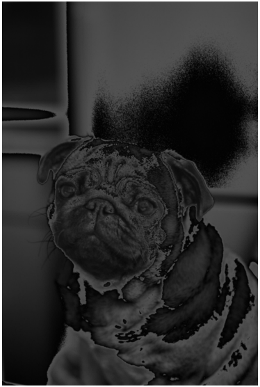
Example: Image restoration[7]



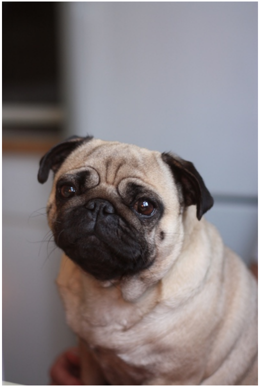
$\leftarrow$

[7] `Lecture1/imageExample.py`

Example: Image restoration[7]



$\leftarrow$



[7] `Lecture1/imageExample.py`

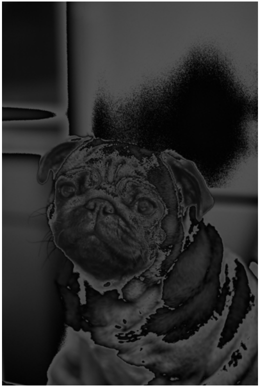
Example: Image restoration[7]

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

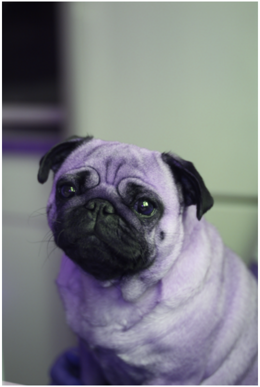
$$\mathbf{y}_i = \frac{1}{3}(\mathbf{x}_i^r + \mathbf{x}_i^g + \mathbf{x}_i^b) \tag{32}$$

[7] `Lecture1/imageExample.py`

Example: Image restoration[7]

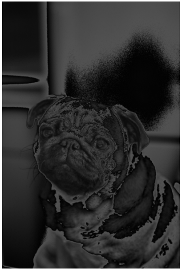


$\leftarrow$



[7] `Lecture1/imageExample.py`

Example: Image restoration[7]



$\leftarrow$

$$p(\mathbf{Y}|\mathbf{X},\theta) = \mathcal{N}(\mathbf{W}\mathbf{X}, \sigma^2\mathbf{I}) \tag{33}$$

$$\mathbf{y}_i = \frac{1}{3}(\mathbf{x}_i^r + \mathbf{x}_i^g + \mathbf{x}_i^b) \tag{34}$$

$$p(\mathbf{X}|\mathbf{Y},\theta) \propto p(\mathbf{Y}|\mathbf{X},\theta)p(\mathbf{X}) \tag{35}$$

[7] `Lecture1/imageExample.py`

Introduction

Regression

Kernel Methods

Dual Linear Regression[8]

$$p(\mathbf{W}|\mathbf{Y},\mathbf{X}) = \frac{p(\mathbf{Y}|\mathbf{W},\mathbf{X})p(\mathbf{W})}{p(\mathbf{Y})} \tag{36}$$

$$p(\mathbf{Y}|\mathbf{W},\mathbf{X}) = \prod_{i}^{N} p(\mathbf{y}_i|\mathbf{W},\mathbf{X}) = \prod_{i}^{N} \mathcal{N}(\mathbf{y}_i|\cdot,\sigma^2\mathbf{I}) \tag{37}$$

$$p(\mathbf{W}) = \mathcal{N}(\mathbf{0},\tau^2\mathbf{I}) \tag{38}$$

[8] Murphy 2012, p. 14.4.3.

Dual Linear Regression[8]

$$p(\mathbf{W}|\mathbf{Y},\mathbf{X}) = \frac{p(\mathbf{Y}|\mathbf{W},\mathbf{X})p(\mathbf{W})}{p(\mathbf{Y})} \tag{39}$$

$$p(\mathbf{Y}|\mathbf{W},\mathbf{X}) = \prod_i^N p(\mathbf{y}_i|\mathbf{W},\mathbf{X}) = \prod_i^N \mathcal{N}(\mathbf{y}_i|\cdot,\sigma^2\mathbf{I}) \tag{40}$$

$$p(\mathbf{W}) = \mathcal{N}(\mathbf{0},\tau^2\mathbf{I}) \tag{41}$$

$$p(\mathbf{W}|\mathbf{Y},\mathbf{X}) \propto p(\mathbf{Y}|\mathbf{W},\mathbf{X})p(\mathbf{W}) \tag{42}$$

[8] Murphy 2012, p. 14.4.3.

Dual Linear Regression[8]

- Lets look at a simple 1D problem

$$\mathbf{y} \in \mathbb{R}^{1\times N} \tag{43}$$
$$\mathbf{x} \in \mathbb{R}^{1\times N} \tag{44}$$

[8] Murphy 2012, p. 14.4.3.

Dual Linear Regression[8]

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

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

Objective

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

[8] Murphy 2012, p. 14.4.3.

Dual Linear Regression[8]

$$p(\mathbf{w}|\mathbf{y},\mathbf{X}) \propto \prod_i^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - y_i)^{\mathrm{T}}(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - y_i)} \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{1}{2\tau^2}(\mathbf{w}^{\mathrm{T}}\mathbf{w})} \tag{47}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{y})^{\mathrm{T}}(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{y})} \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{1}{2\tau^2}(\mathbf{w}^{\mathrm{T}}\mathbf{w})} \tag{48}$$

Objective

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

[8] Murphy 2012, p. 14.4.3.

Dual Linear Regression[8]

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

Objective

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

[8] Murphy 2012, p. 14.4.3.

Dual Linear Regression[8]

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

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

Optimisation

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

[8]Murphy 2012, p. 14.4.3.

Dual Linear Regression[8]

$$J(\mathbf{w}) = \frac{1}{2}(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{y})^{\mathrm{T}}(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{y}) + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \tag{52}$$

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

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

Optimisation

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

[8] Murphy 2012, p. 14.4.3.

Dual Linear Regression[8]

$$J(\mathbf{w}) = \frac{1}{2}(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{y})^{\mathrm{T}}(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{y}) + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \tag{55}$$

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

$$\mathbf{w} = -\frac{1}{\lambda}\mathbf{X}^{\mathrm{T}}(\mathbf{w}^{\mathrm{T}}\mathbf{X} - \mathbf{y}) = \mathbf{X}^{\mathrm{T}}\mathbf{a} = \sum_{n}^{N} \alpha_n \mathbf{x}_n \tag{57}$$

Optimisation

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

[8] Murphy 2012, p. 14.4.3.

Dual Linear Regression[8]

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

$$\mathbf{w} = \mathbf{X}^{\mathrm{T}}\mathbf{a} \tag{59}$$

Formulate Dual

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

[8] Murphy 2012, p. 14.4.3.

Dual Linear Regression[8]

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

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

[8] Murphy 2012, p. 14.4.3.

Dual Linear Regression[8]

$$[\mathbf{K}]_{ij} = \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j \tag{63}$$

$$J(\mathbf{a}) = \frac{1}{2}\mathbf{a}^{\mathrm{T}}\mathbf{K}\mathbf{K}\mathbf{a} - \mathbf{a}\mathbf{K}\mathbf{y} + \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{y} + \frac{\lambda}{2}\mathbf{a}^{\mathrm{T}}\mathbf{K}\mathbf{a} \tag{64}$$

$$\alpha_i = -\frac{1}{\lambda}(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - y_i) \tag{65}$$

$$\mathbf{w} = \sum_i^N \alpha_i \mathbf{x}_i \tag{66}$$

$$\Rightarrow \mathbf{a} = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y} \tag{67}$$

[8] Murphy 2012, p. 14.4.3.

Dual Linear Regression[8]

$$[\mathbf{K}]_{ij} = \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j \tag{68}$$

$$J(\mathbf{a}) = \frac{1}{2}\mathbf{a}^{\mathrm{T}}\mathbf{K}\mathbf{K}\mathbf{a} - \mathbf{a}\mathbf{K}\mathbf{y} + \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{y} + \frac{\lambda}{2}\mathbf{a}^{\mathrm{T}}\mathbf{K}\mathbf{a} \tag{69}$$

$$\mathbf{a} = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y} \tag{70}$$

$$\mathbf{y}(\mathbf{x}_*) = \mathbf{w}^{\mathrm{T}}\mathbf{x}_* = \mathbf{a}^{\mathrm{T}}\mathbf{X}\mathbf{x}_* = \mathbf{a}^{\mathrm{T}}k(\mathbf{X}, \mathbf{x}_*) = \tag{71}$$

$$= ((\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y})^{\mathrm{T}}k(\mathbf{X}, \mathbf{x}_*) = k(\mathbf{x}_*, \mathbf{X})(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y} \tag{72}$$

[8] Murphy 2012, p. 14.4.3.

Dual Linear Regression[8]

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 $\mathbf{W}$

[8] Murphy 2012, p. 14.4.3.

Dual Linear Regression[8]

Linear Regression

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

[8] Murphy 2012, p. 14.4.3.

Dual Linear Regression[8]

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

[8] Murphy 2012, p. 14.4.3.

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?

$$\phi : \mathbf{x}_i \rightarrow \mathbf{f}_i \tag{73}$$

- *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?

$$\phi : \mathbf{x}_i \rightarrow \mathbf{f}_i \tag{74}$$

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

Kernels

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^{\mathrm{T}}\phi(\mathbf{x}_j) = ||\phi(\mathbf{x}_i)||||\phi(\mathbf{x}_j)||\cos(\theta) \tag{75}$$

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

$$
\begin{aligned}
\mathbf{x} &\in \mathbb{R}^2 && (76)\\
(\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j)^2 &= (x_{i1}x_{j1} + x_{i2}x_{j2})^2 = && (77)\\
&= x_{i1}^2 x_{j1}^2 + 2x_{i1}x_{j1}x_{i2}x_{j2} + x_{i2}^2 x_{j2}^2 = && (78)\\
&= (x_{i1}^2, \sqrt{2}x_{i1}x_{i2}, x_{i2}^2)(x_{j1}^2, \sqrt{2}x_{j1}x_{j2}, x_{j2}^2)^{\mathrm{T}} = && (79)\\
&= \phi(\mathbf{x}_i)^{\mathrm{T}}\phi(\mathbf{x}_j) && (80)
\end{aligned}
$$

Kernels

$$
\begin{aligned}
\mathbf{x} &\in \mathbb{R}^2 && (81)\\
(\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j)^2 &= (x_{i1}x_{j1} + x_{i2}x_{j2})^2 = && (82)\\
&= x_{i1}^2 x_{j1}^2 + 2x_{i1}x_{j1}x_{i2}x_{j2} + x_{i2}^2 x_{j2}^2 = && (83)\\
&= (x_{i1}^2, \sqrt{2}x_{i1}x_{i2}, x_{i2}^2)(x_{j1}^2, \sqrt{2}x_{j1}x_{j2}, x_{j2}^2)^{\mathrm{T}} = && (84)\\
&= \phi(\mathbf{x}_i)^{\mathrm{T}}\phi(\mathbf{x}_j) && (85)
\end{aligned}
$$

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

9

[9] Murphy 2012, p. 14.2.3

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 25th 8-10 E2
- Continue with Kernels
  - relation to co-variance
- Non-parametric Regression
  - Gaussian Processes
- Start Assignment

Next Time

Lecture 2

- November 25th 8-10 E2
- Continue with Kernels
  - relation to co-variance
- Non-parametric Regression
  - Gaussian Processes
- Start Assignment

Next Time

Practical Session 1

- November 21st, 15-17 in Q31
- My best friend the Gaussian
  - Multiplication
  - Marginalisation
  - Recap: Matrix derivatives
- Things that you will need for Assignment 1

e.o.f.

References I

Kevin P Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN: 0262018020, 9780262018029.