University of BRISTOL

Machine Learning

Linear Regression

Carl Henrik Ek - carlhenrik.ek@bristol.ac.uk

October 3, 2017

http://www.carlhenrik.com

Introduction

So Far

- Lecture 1 What is machine Learning
  - assumptions are the fundation of learning
  - probabilities are the language of assumptions

So Far

- Lecture 1 What is machine Learning
  - assumptions are the fundation of learning
  - probabilities are the language of assumptions
- Lecture 2 Probabilities
  - what are the rules of probability
  - distributions are the parametrised form of a probability

So Far

- Lecture 1 What is machine Learning
  - assumptions are the fundation of learning
  - probabilities are the language of assumptions
- Lecture 2 Probabilities
  - what are the rules of probability
  - distributions are the parametrised form of a probability
- Lecture 3 Distributions
  - discrete and continous distributions
  - conjugate distributions

Likelihood or Prior

$$\underbrace{p(X|Y)}_{\text{posterior}} = \underbrace{P(Y|X)}_{\text{likelihood}} \cdot \underbrace{p(X)}_{\text{prior}} \cdot \frac{1}{\underbrace{p(Y)}_{\text{evidence}}}$$

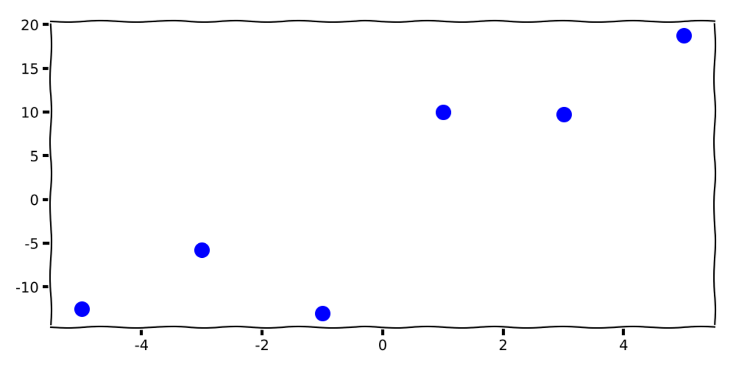
$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Today

Lets build our first model

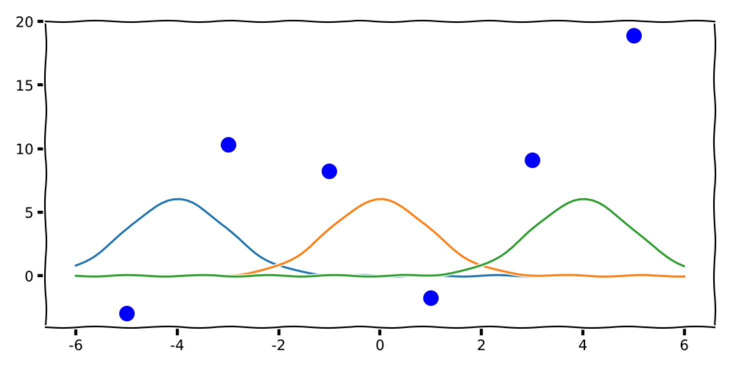
Linear Regression

Linear Regression [1] Ch 3.1



- Linear function in both parameters and data

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \ldots w_D x_D = \mathbf{w}^{\mathrm{T}} \mathbf{x} + w_0 = \{D = 1\} w_0 + w_1 * x$$



- Linear function only in parameters

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}) = \{\phi_0(\mathbf{x}) = 1\} = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

- We can choose many types of basis functions $\phi(\mathbf{x})$


YES!
FINALLY!
memegenerator.net

Linear Regression

- Model

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}) + \epsilon$$
$$\epsilon \sim \mathcal{N}(0, I)$$

Linear Regression

- Model

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}) + \epsilon$$
$$\epsilon \sim \mathcal{N}(0, I)$$

- Likelihood

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}\left(t|\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}), \beta^{-1}\right)$$

Linear Regression

- Model

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}) + \epsilon$$
$$\epsilon \sim \mathcal{N}(0, I)$$

- Likelihood

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}\left(t|\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}), \beta^{-1}\right)$$

- Independence

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t|\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}), \beta^{-1}\right)$$

Maximum Likelihood [1] Ch 3.1.1

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t|\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n), \beta^{-1}\right)$$

Maximum Likelihood [1] Ch 3.1.1

$$
\begin{aligned}
p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) &= \prod_{n=1}^{N} \mathcal{N}\left(t|\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n), \beta^{-1}\right) \\
&= \prod_{n=1}^{N} \frac{1}{(2\pi\beta^{-1})^{\frac{1}{2}}} e^{-\frac{1}{2}\beta(t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n))^2}
\end{aligned}
$$

Maximum Likelihood [1] Ch 3.1.1

$$
\begin{aligned}
p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) &= \prod_{n=1}^{N} \mathcal{N}\left(t|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right) \\
&= \prod_{n=1}^{N} \frac{1}{(2\pi\beta^{-1})^{\frac{1}{2}}} e^{-\frac{1}{2}\beta(t_n-\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n))^2} \\
&= (\frac{\beta}{2\pi})^{\frac{N}{2}} e^{-\frac{\beta}{2}\sum_{n=1}^{N}(t_n-\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n))^2}
\end{aligned}
$$

Maximum Likelihood [1] Ch 3.1.1

$$
\begin{aligned}
p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) &= \prod_{n=1}^{N} \mathcal{N}\left(t|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right) \\
&= \prod_{n=1}^{N} \frac{1}{(2\pi\beta^{-1})^{\frac{1}{2}}} e^{-\frac{1}{2}\beta(t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n))^2} \\
&= (\frac{\beta}{2\pi})^{\frac{N}{2}} e^{-\frac{\beta}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n))^2} \\
\log p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) &= \frac{N}{2}(\log(\beta) - \log(2\pi)) - \beta\frac{1}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n))^2
\end{aligned}
$$

Maximum Likelihood

$$\log p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \frac{N}{2}(\underbrace{\log(\beta)}_{\text{A}} - \underbrace{\log(2\pi)}_{\text{B}}) - \underbrace{\beta\frac{1}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n))^2}_{\text{C}}$$

A noise precision

B constant

C error

Maximum Likelihood

- Take derivative

$$\nabla \log\, p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \beta \sum_{n=1}^{N} (t_n - \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}_n)) \phi(\mathbf{x}_n)^{\mathrm{T}}$$

Maximum Likelihood

- Take derivative

$$\nabla \log\, p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \beta \sum_{n=1}^{N} (t_n - \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}_n)) \phi(\mathbf{x}_n)^{\mathrm{T}}$$

- Stationary point

$$0 = \sum_{n=1}^{N} t_n \phi(\mathbf{x}_n)^{\mathrm{T}} - \mathbf{w}^{\mathrm{T}} \left(\sum_{n=1}^{N} \phi(\mathbf{x}_n) \phi(\mathbf{x}_n)^{\mathrm{T}} \right)$$

Maximum Likelihood

- Take derivative

$$\nabla \log\, p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \beta \sum_{n=1}^{N} (t_n - \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}_n)) \phi(\mathbf{x}_n)^{\mathrm{T}}$$

- Stationary point

$$0 = \sum_{n=1}^{N} t_n \phi(\mathbf{x}_n)^{\mathrm{T}} - \mathbf{w}^{\mathrm{T}} \left(\sum_{n=1}^{N} \phi(\mathbf{x}_n) \phi(\mathbf{x}_n)^{\mathrm{T}} \right)$$

- Solve for parameters $\mathbf{w}$

$$\mathbf{w}_{\mathrm{ML}} = (\phi(\mathbf{X})^{\mathrm{T}} \phi(\mathbf{X}))^{-1} \phi(\mathbf{X})^{\mathrm{T}} \mathbf{t}$$

Maximum Likelihood

- Take derivative

$$\nabla \log p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \beta \sum_{n=1}^{N} (t_n - \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}_n)) \phi(\mathbf{x}_n)^{\mathrm{T}}$$

- Stationary point

$$0 = \sum_{n=1}^{N} t_n \phi(\mathbf{x}_n)^{\mathrm{T}} - \mathbf{w}^{\mathrm{T}} \left(\sum_{n=1}^{N} \phi(\mathbf{x}_n) \phi(\mathbf{x}_n)^{\mathrm{T}} \right)$$

- Solve for parameters $\mathbf{w}$

$$\mathbf{w}_{\mathrm{ML}} = (\phi(\mathbf{X})^{\mathrm{T}} \phi(\mathbf{X}))^{-1} \phi(\mathbf{X})^{\mathrm{T}} \mathbf{t}$$

- and precision

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N} \sum_{n=1}^{N} (t_n - \mathbf{w}_{\mathrm{ML}}^{\mathrm{T}} \phi(\mathbf{x}_n))^2$$

Maximum Likelihood

$$\mathbf{w}_{\mathsf{ML}} = \underbrace{(\phi(\mathbf{X})^{\mathrm{T}}\phi(\mathbf{X}))^{-1}\phi(\mathbf{X})^{\mathrm{T}}}_{\phi(\mathbf{X})^{+}}\mathbf{t}$$

- Moore-Penrose inverse (`np.linalg.pinv` in numpy)

Bayesian Linear Regression [1] Ch 3.3

- Likelihood is Gaussian in $\mathbf{w}$

Bayesian Linear Regression [1] Ch 3.3

- Likelihood is Gaussian in $\mathbf{w}$
- Conjugate Prior

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$$

Bayesian Linear Regression [1] Ch 3.3

- Likelihood is Gaussian in $\mathbf{w}$
- Conjugate Prior

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$$

- Posterior

$$p(w|\mathbf{t} = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

Bayesian Linear Regression [1] Ch 3.3

- Likelihood is Gaussian in $\mathbf{w}$
- Conjugate Prior

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$$

- Posterior

$$p(w|\mathbf{t} = \mathcal{N}\left(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N\right)$$

- *Gaussian identities!*

Posterior

- Posterior is Gaussian

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

Posterior

- Posterior is Gaussian

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

- Identification

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}) \propto p(\mathbf{t}|\mathbf{X}, \mathbf{w})p(\mathbf{w})$$

Posterior

- Posterior is Gaussian

$$p(\mathbf{w}|\mathbf{t},\mathbf{X}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N,\mathbf{S}_N)$$

- Identification

$$p(\mathbf{w}|\mathbf{t},\mathbf{X}) \propto p(\mathbf{t}|\mathbf{X},\mathbf{w})p(\mathbf{w})$$

- Posterior

$$\mathbf{m}_N = \left(\mathbf{S}_0^{-1} + \beta\phi(\mathbf{X})^{\mathrm{T}}\phi(\mathbf{X})\right)^{-1}\left(S_0^{-1}\mathbf{m}_0 + \beta\phi(\mathbf{X})^{\mathrm{T}}\mathbf{t}\right)$$
$$\mathbf{S}_N = \left(\mathbf{S}_0^{-1} + \beta\phi(\mathbf{X})^{\mathrm{T}}\phi(\mathbf{X})\right)^{-1}$$

Posterior

- Assumption Zero mean isotropic Gaussian

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

Posterior

- Assumption Zero mean isotropic Gaussian

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

- Posterior

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}) = \mathcal{N}(\mathbf{w}|\beta \left(\alpha\mathbf{I} + \beta\phi(\mathbf{X})^{\mathrm{T}}\phi(\mathbf{X})\right)^{-1} \phi(\mathbf{X})^{\mathrm{T}}\mathbf{t}, \\ \left(\alpha\mathbf{I} + \beta\phi(\mathbf{X})^{\mathrm{T}}\phi(\mathbf{X})\right)^{-1})$$

Posterior

- Assumption Zero mean isotropic Gaussian

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

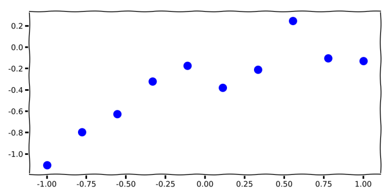
- Posterior

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}) = \mathcal{N}(\mathbf{w}|\beta \left(\alpha\mathbf{I} + \beta\phi(\mathbf{X})^{\mathrm{T}}\phi(\mathbf{X})\right)^{-1} \phi(\mathbf{X})^{\mathrm{T}}\mathbf{t}, \\ \left(\alpha\mathbf{I} + \beta\phi(\mathbf{X})^{\mathrm{T}}\phi(\mathbf{X})\right)^{-1})$$

- ML

$$\mathbf{w}_{\mathrm{ML}} = (\phi(\mathbf{X})^{\mathrm{T}}\phi(\mathbf{X}))^{-1}\phi(\mathbf{X})^{\mathrm{T}}\mathbf{t}$$

Linear Regression Example [1] Figure 3.7



- Model

$$y(x, \mathbf{w}) = w_0 + w_1 x$$

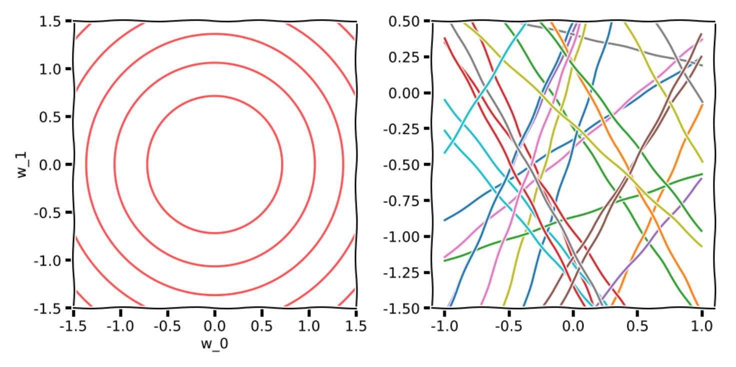
- Data

$$f(x, \mathbf{a}) = a_0 + a_1 x, \ \{a_0, a_1\} = \{-0.3, 0.5\}$$
$$t = f(x, \mathbf{a}) + \epsilon, \ \epsilon \sim \mathcal{N}(0, 0.2^2)$$

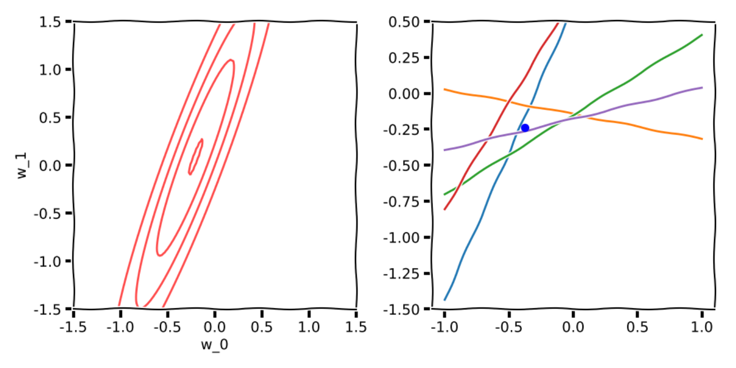
- Prior

$$p(\mathbf{w}) = \mathcal{N}(\boldsymbol{\alpha} | \mathbf{0}, 2.0 \cdot \mathbf{I})$$

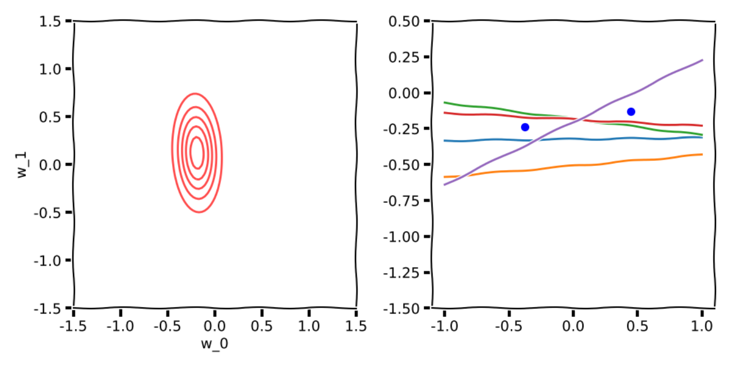
Linear Regression Example



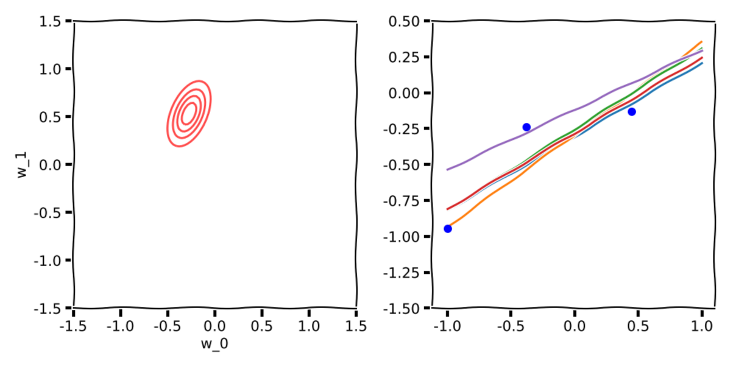
Linear Regression Example

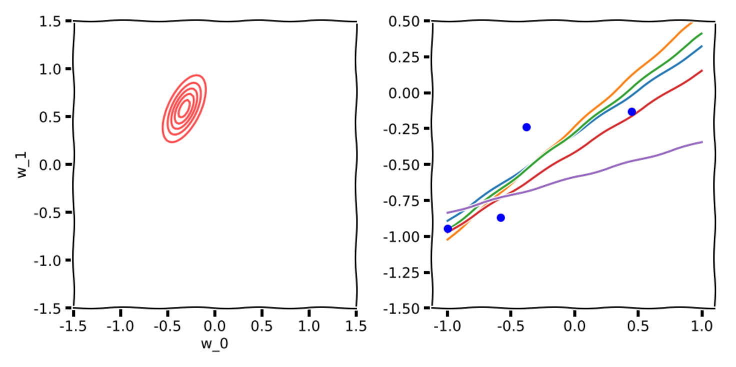


Linear Regression Example



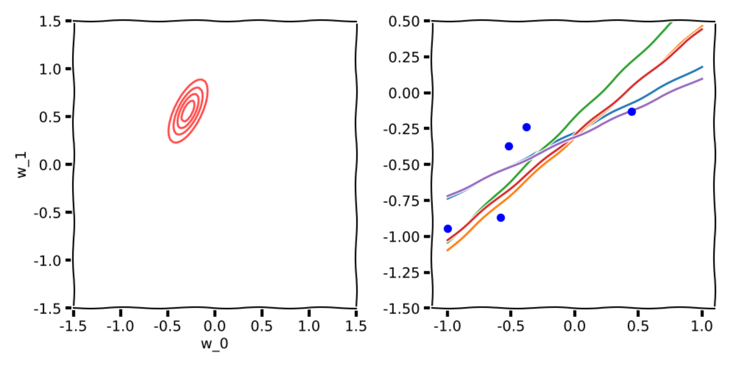
Linear Regression Example



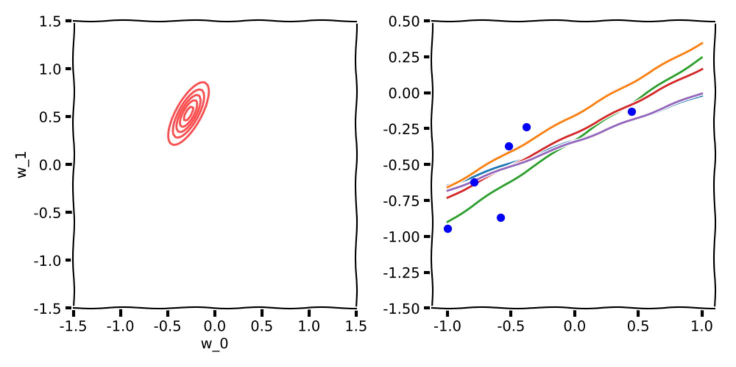


1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
w_1
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
w_0
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.00
-1.25
-1.50
-1.0
-0.5
0.0
0.5
1.0

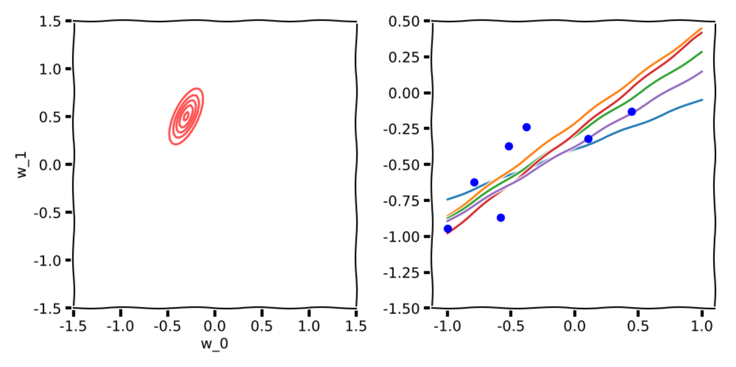
Linear Regression Example



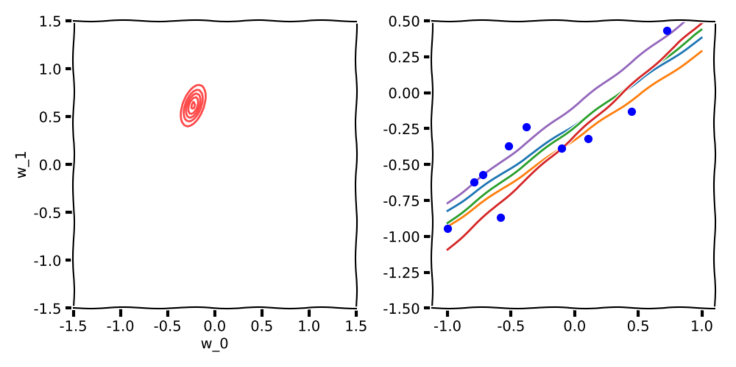
Linear Regression Example



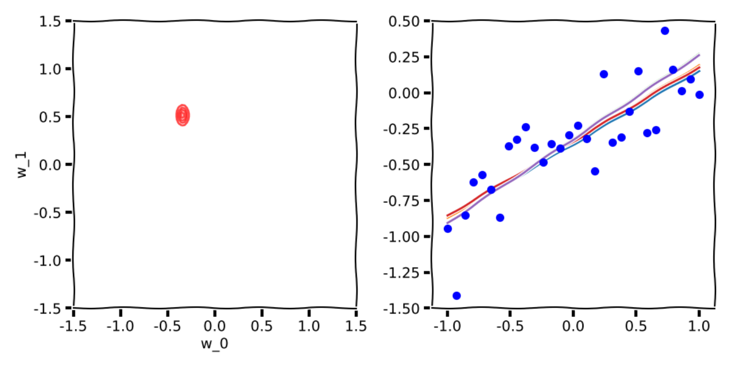
Linear Regression Example



Linear Regression Example



Linear Regression Example



Linear Regression

- Don't underestimate what we just did

Linear Regression

- Don't underestimate what we just did
- We saw data, which we knew where it came from

Linear Regression

- Don't underestimate what we just did
- We saw data, which we knew where it came from
- We made an assumption

- Don't underestimate what we just did
- We saw data, which we knew where it came from
- We made an assumption
- We recovered the system

Linear Regression

- Don't underestimate what we just did
- We saw data, which we knew where it came from
- We made an assumption
- We recovered the system
- We generated knowledge from data!!!

Linear Regression

- Don't underestimate what we just did
- We saw data, which we knew where it came from
- We made an assumption
- We recovered the system
- We generated knowledge from data!!!
- Understand [1] 3.3 it might be the most important thing in the unit

Statistics or Machine Learning

"The difference between statistics and machine learning is that the former cares about parameters while the latter cares about prediction"
– Prof. Neil D. Lawrence

Prediction

$$p(t_*|\mathbf{t}, \mathbf{x}_*, \mathbf{X}, \alpha, \beta) = \int p(t_*|\mathbf{x}_*, \mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \alpha, \beta) \mathrm{d}\mathbf{w}$$

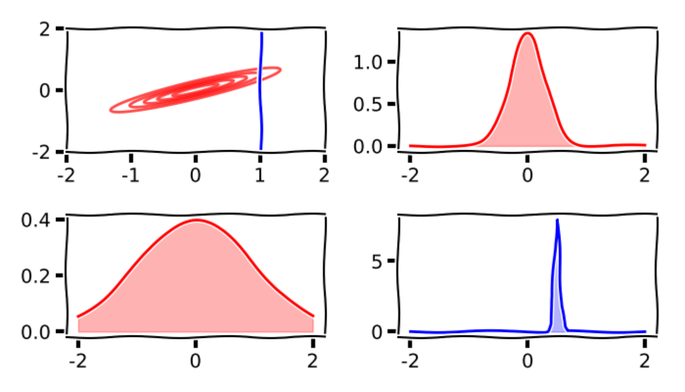
- we do not really care about w we care about new prediction t_* at location $\mathbf{x}_*$
- look at the marginal distribution, i.e. when we average out the weight
- integrate a Gaussian over a Gaussian $\Rightarrow$ Gaussian identities

Prediction

$$p(t_*|\mathbf{t}, \mathbf{x}_*, \mathbf{X}, \alpha, \beta) = \int p(t_*|\mathbf{x}_*, \mathbf{w}, \beta)p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \alpha, \beta)\mathrm{d}\mathbf{w}$$

- we do not really care about w we care about new prediction t_* at location $\mathbf{x}_*$
- look at the marginal distribution, i.e. when we average out the weight
- integrate a Gaussian over a Gaussian $\Rightarrow$ Gaussian identities
- *They are really important so look at them once in detail!!*

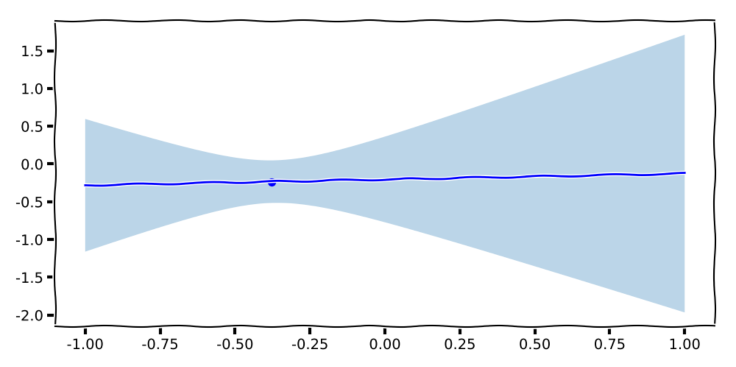
Prediction



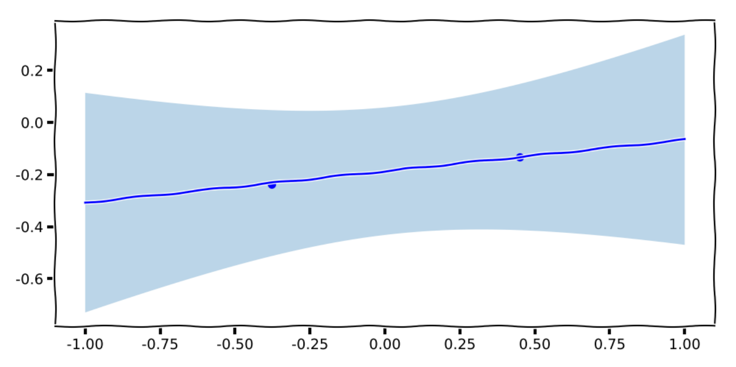
$$p(t_*|\mathbf{t}, \mathbf{x}_*, \mathbf{X}, \alpha, \beta) = \int p(t_*|\mathbf{x}_*, \mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \alpha, \beta) \mathrm{d}\mathbf{w}$$

$$nnnnnnnnnnn\mathcal{N}\left(t_*|\mathbf{m}_N^{\mathrm{T}}\phi(\mathbf{x}_*), \frac{1}{\beta} + \phi(\mathbf{x}_*)^{\mathrm{T}}\mathbf{S}_N\phi(\mathbf{x}_*)\right)$$

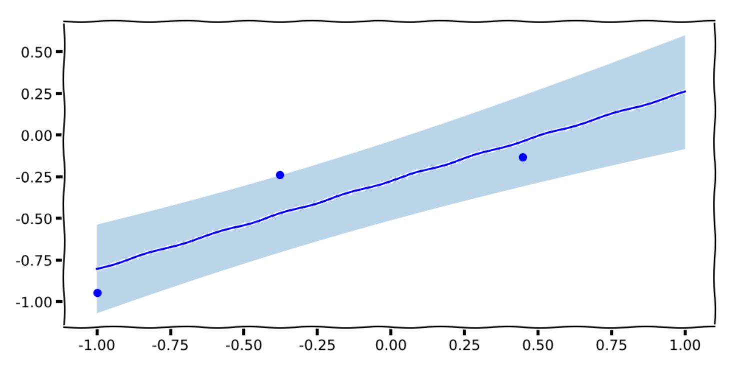
Predictive Posterior



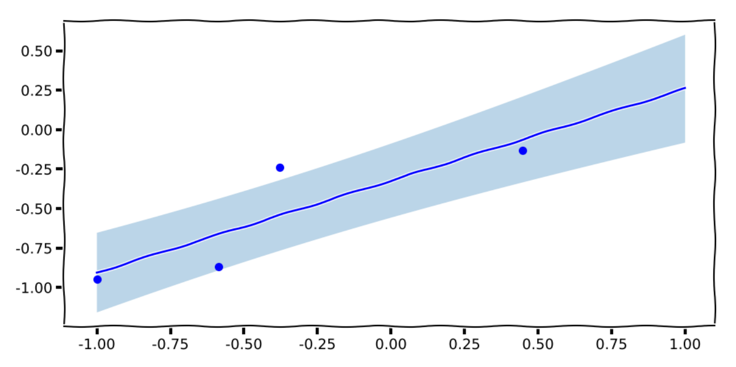
Predictive Posterior



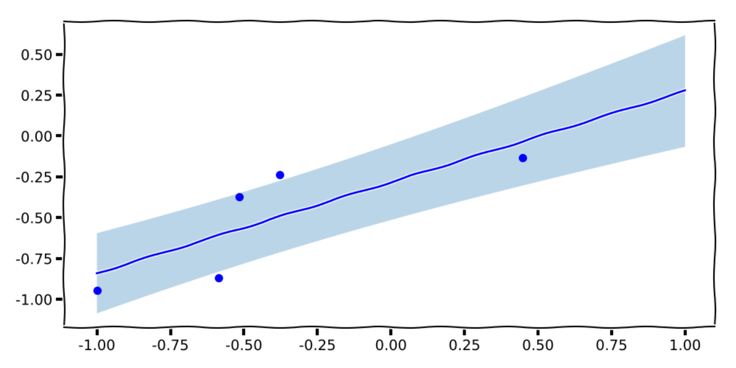
Predictive Posterior



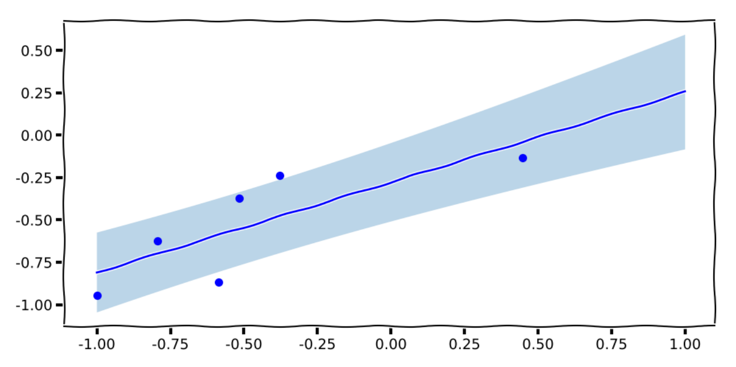
Predictive Posterior



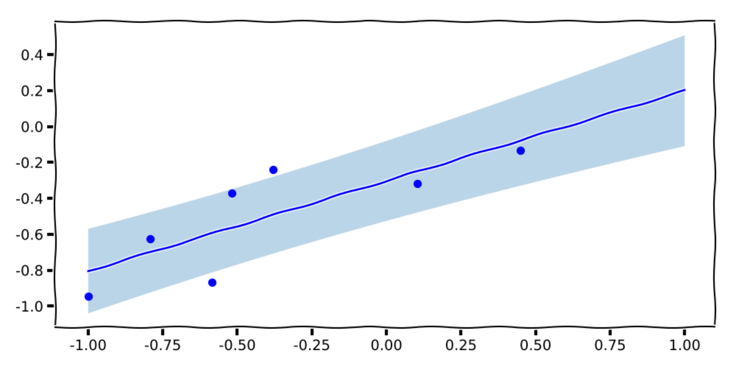
Predictive Posterior



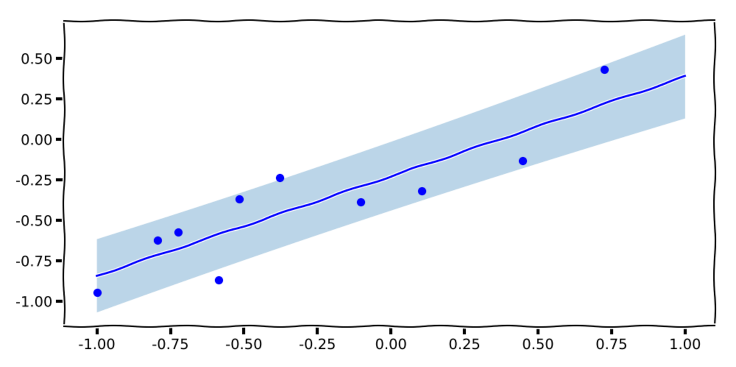
Predictive Posterior



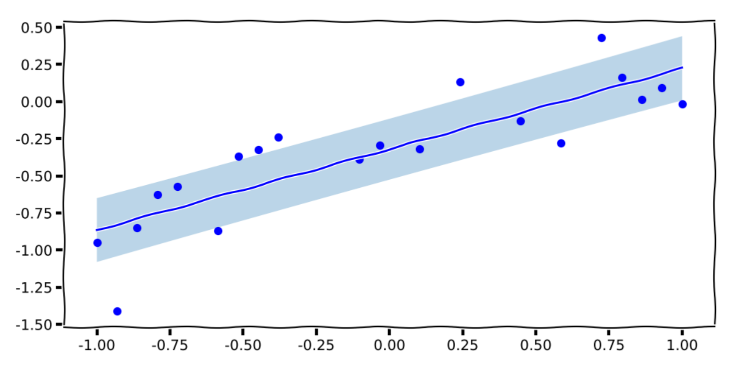
Predictive Posterior



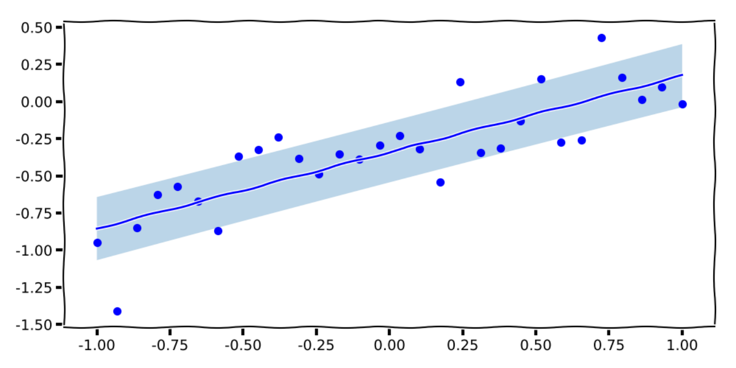
Predictive Posterior



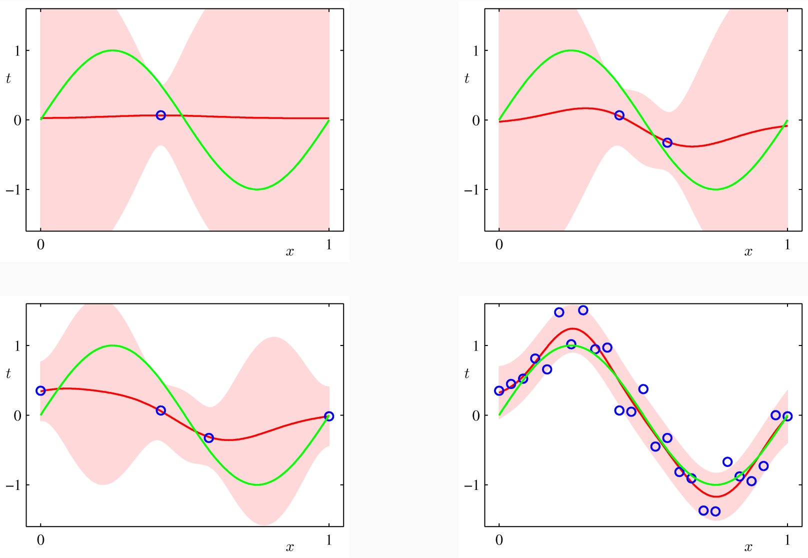
Predictive Posterior



Predictive Posterior



Predictive Posterior [1] Figure 3.8



Which Parametrisation

- Should I use a line, polynomial, quadratic basis function?
- Likelihood won't help me
- How do we proceed?

Being Bayesian

$$p(\mathcal{M}_i|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M}_i)p(\mathcal{M}_i)}{p(\mathcal{D})}$$

- Treat the model as uncertain itself, i.e make assumptions
- Same as with parameters, just learn it from data

Being Bayesian

$$p(\mathcal{M}_i|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M}_i)p(\mathcal{M}_i)}{p(\mathcal{D})}$$

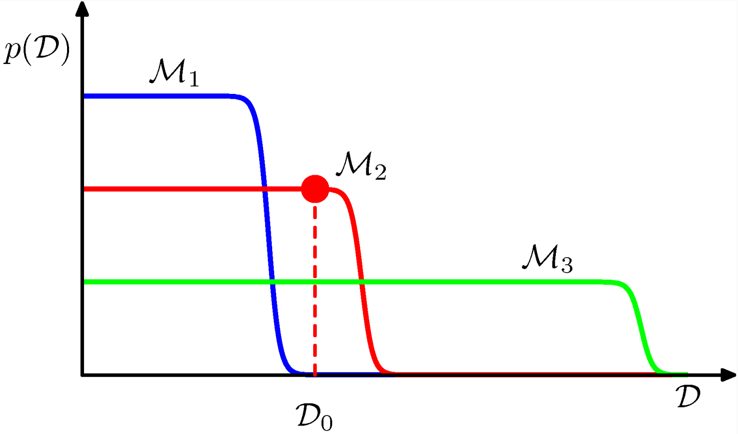
- Treat the model as uncertain itself, i.e make assumptions
- Same as with parameters, just learn it from data
- often totally intractable to compute $p(\mathcal{D}|\mathcal{M}_i)$

Being Bayesian

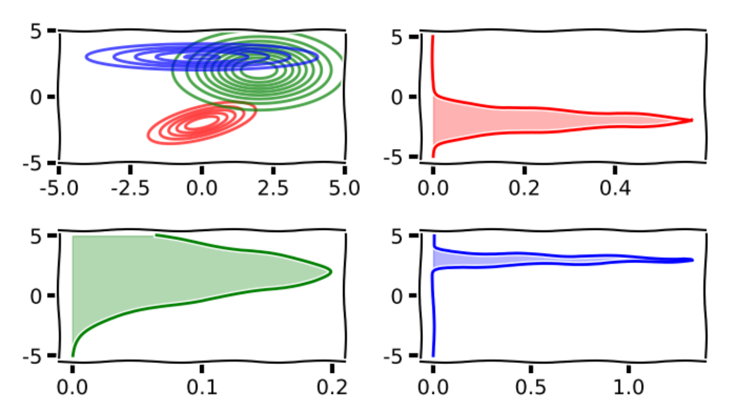
$$p(\mathcal{M}_i|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M}_i)p(\mathcal{M}_i)}{p(\mathcal{D})}$$

- Treat the model as uncertain itself, i.e make assumptions
- Same as with parameters, just learn it from data
- often totally intractable to compute $p(\mathcal{D}|\mathcal{M}_i)$
- marginalise all parameters from the model

Marginal Likelihood [1] Figure 3.13



Marginal Distribution



Summary

So Far

Lecture 1 What is machine Learning

- assumptions are the fundation of learning
- probabilities are the language of assumptions

Lecture 2 Probabilities

- what are the rules of probability
- distributions are the parametrised form of a probability

Lecture 3 Distributions

- discrete and continous distributions
- conjugate distributions

Today Models

- how to apply our assumptions to data
- how to learn for real

Summary[1]

274 000$

[1] http://www.paysa.com

Part II

- Linear models can only take us that far
    - Monday - Non-linear models
- Fixed model complexity
    - Tuesday - Non-parametric models

Question 1-6   12

References

Christopher M. Bishop.
Pattern Recognition and Machine Learning (Information Science and Statistics).
Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.