From Likelihood to Bayesianism

CS4061 / CS5014 Machine Learning

Paul Henderson
University of Glasgow

paul.henderson@glasgow.ac.uk

Based on previous material by
Simon Rogers & Ke Yuan

Discrete RVs

- random events with outcomes that we can count
- defined by probabilities of different events taking place
  e.g. probability of random variable X taking value x:

$$P(X = x)$$

- example: fair 6-sided die:

$$P(Y = y) = \frac{1}{6} \quad \text{for } y = 1, \ldots, 6$$

Discrete RVs

- random events with outcomes that we can count
- defined by probabilities of different events taking place
  e.g. probability of random variable X taking value x:

$$P(X = x)$$

- example: fair 6-sided die:

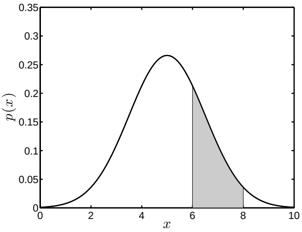
$$P(Y = y) = \frac{1}{6} \quad \text{for } y = 1, \ldots, 6$$

- probabilities are constrained:

$$0 \leq P(Y = y) \leq 1, \quad \sum_y P(Y = y) = 1$$

Continuous RVs

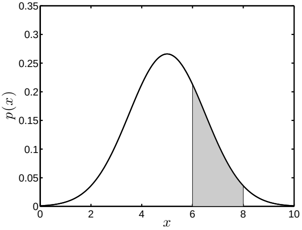
- random events with outcomes that we cannot count
- define a density function $p(x)$:



- $p(x)$ tells us how likely different values are
- these are **not** probabilities!

Continuous RVs

- random events with outcomes that we cannot count
- define a density function $p(x)$:



- $p(x)$ tells us how likely different values are
- these are **not** probabilities!

- probabilities of ranges given by area under the curve:

$$P(6 \leq X \leq 8) = \int_{x=6}^{x=8} p(x)\; dx$$

- densities are constrained:

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

Modelling uncertainties

- observations (data) are noisy $\Rightarrow$ treat them as random variables

Modelling uncertainties

- observations (data) are noisy $\Rightarrow$ treat them as random variables
- model predicts **random variable** T_n

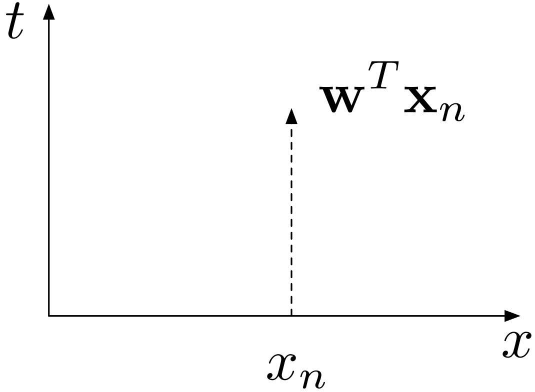
Modelling uncertainties

- observations (data) are noisy $\Rightarrow$ treat them as random variables
- model predicts **random variable** T_n

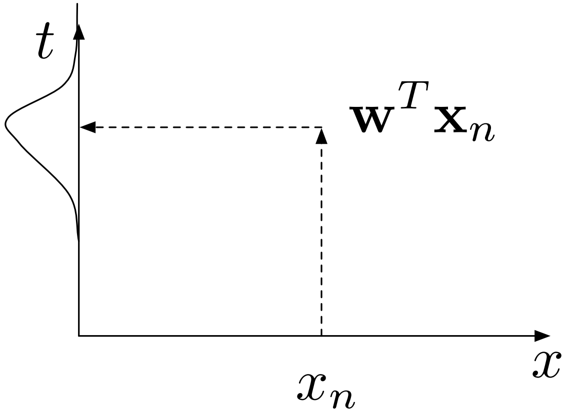
$$T_n = \mathbf{w}^\mathsf{T}\mathbf{x}_n + \epsilon_n$$

- ϵ_n is the **noise**, $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$
- equivalently:

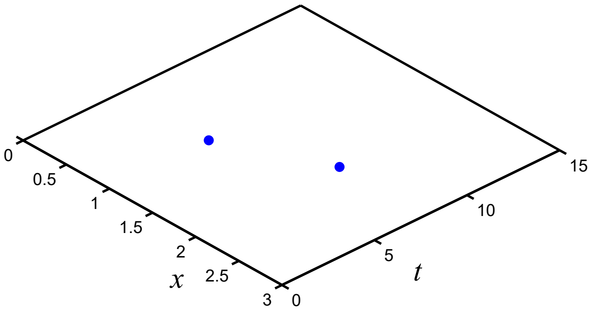
$$T_n \sim \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2)$$



t
$\mathbf{w}^T\mathbf{x}_n$
x_n
x

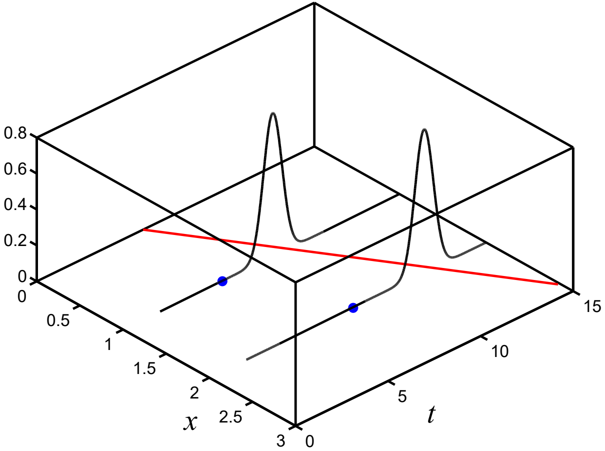

t
$\mathbf{w}^T\mathbf{x}_n$
x
x_n

2023 PPTX here has a ‘model 1’ slide with points and red line, but without 3D-ness and gaussians!



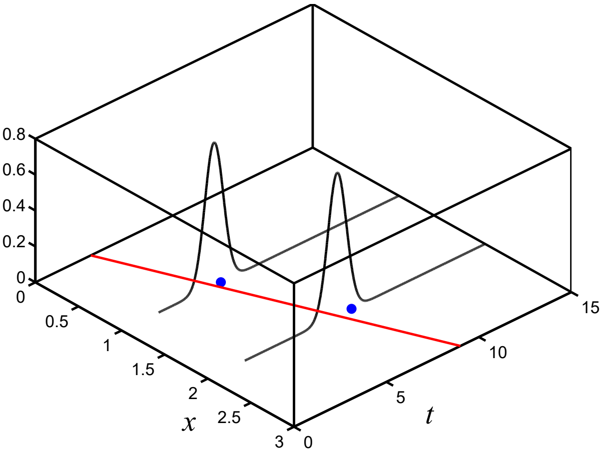
0
0.5
1
1.5
2
2.5
3
x
0
5
10
15
t

$$p(T_n = t_n \mid \mathbf{w}, \mathbf{x}_n, \sigma^2)$$



Model 1: low likelihood

$$p(T_n = t_n \mid \mathbf{w}, \mathbf{x}_n, \sigma^2)$$



Model 2: high likelihood

Likelihood

- T_n is a Gaussian random variable

$$T_n \sim \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2)$$

- it has probability density

$$p(T_n = t \mid \mathbf{w}, \mathbf{x}_n, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(t - \mathbf{w}^\mathsf{T}\mathbf{x}_n)^2\right\}$$

Likelihood

- T_n is a Gaussian random variable

$$T_n \sim \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2)$$

- it has probability density

$$p(T_n = t \mid \mathbf{w}, \mathbf{x}_n, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(t - \mathbf{w}^\mathsf{T}\mathbf{x}_n)^2\right\}$$

- t_n is our (non-random!) observation
- density of T_n at point t_n is called **likelihood** of t_n
  - i.e. $p(T_n = t_n \mid \mathbf{w}, \mathbf{x}_n, \sigma^2)$

Likelihood optimisation

- For each input-response pair, we have a Gaussian likelihood

$$p(T_n = t_n \,|\, \mathbf{w}, \mathbf{x}_n, \sigma^2)$$

- To combine them all, we want the joint likelihood:

$$p(T_1 = t_1, \ldots, T_N = t_N \,|\, \mathbf{w}, \sigma^2, \mathbf{x}_1, \ldots, \mathbf{x}_N)$$

Likelihood optimisation

- For each input-response pair, we have a Gaussian likelihood

$$p(T_n = t_n \mid \mathbf{w}, \mathbf{x}_n, \sigma^2)$$

- To combine them all, we want the joint likelihood:

$$p(T_1 = t_1, \ldots, T_N = t_N \mid \mathbf{w}, \sigma^2, \mathbf{x}_1, \ldots, \mathbf{x}_N)$$

- Assume that the t_n are independent:

$$p(T_1 = t_1, \ldots, T_N = t_N \mid \mathbf{w}, \sigma^2, \mathbf{x}_1, \ldots, \mathbf{x}_N) = \prod_{n=1}^{N} p(T_n = t_n | \mathbf{w}, \mathbf{x}_n, \sigma^2)$$

Likelihood optimisation

Find the parameters that maximise the joint likelihood:

$$\underset{\mathbf{w},\sigma^2}{\operatorname{argmax}} \prod_{n=1}^{N} p(T_n = t_n | \mathbf{w}, \mathbf{x}_n, \sigma^2)$$

Likelihood optimisation

Find the parameters that maximise the joint likelihood:

$$\underset{\mathbf{w},\sigma^2}{\text{argmax}} \prod_{n=1}^{N} p(T_n = t_n | \mathbf{w}, \mathbf{x}_n, \sigma^2)$$

Easier: optimise the log-likelihood:

- if we increase z, $\log(z)$ increases
- if we decrease z, $\log(z)$ decreases
- so, at a maximum of z, $\log(z)$ will also be at a maximum

Likelihood optimisation

Find the parameters that maximise the joint likelihood:

$$\underset{\mathbf{w},\sigma^2}{\operatorname{argmax}} \prod_{n=1}^{N} p(T_n = t_n | \mathbf{w}, \mathbf{x}_n, \sigma^2)$$

Easier: optimise the log-likelihood:

- if we increase z, $\log(z)$ increases
- if we decrease z, $\log(z)$ decreases
- so, at a maximum of z, $\log(z)$ will also be at a maximum

$$\underset{\mathbf{w},\sigma^2}{\operatorname{argmax}} \log \prod_{n=1}^{N} p(T_n = t_n | \mathbf{w}, \mathbf{x}_n, \sigma^2)$$

Some re-arranging...

$$
\begin{aligned}
p(T_n = t_n | \mathbf{w}, \mathbf{x}_n, \sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2}(t_n - \mathbf{w}^\mathsf{T}\mathbf{x}_n)^2 \right\} \\
\log L &= \log \prod_{n=1}^{N} p(t_n | \mathbf{w}, \mathbf{x}_n, \sigma^2)
\end{aligned}
$$

Some re-arranging...

$$
\begin{aligned}
p(T_n = t_n|\mathbf{w}, \mathbf{x}_n, \sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2}(t_n - \mathbf{w}^\mathsf{T}\mathbf{x}_n)^2 \right\} \\
\log L &= \log \prod_{n=1}^{N} p(t_n|\mathbf{w}, \mathbf{x}_n, \sigma^2) \\
&= \sum_{n=1}^{N} \log p(t_n|\mathbf{w}, \mathbf{x}_n, \sigma^2) \\
&= \sum_{n=1}^{N} \log \left(\frac{1}{\sigma\sqrt{2\pi}} \right) - \sum_{n=1}^{N} \frac{1}{2\sigma^2}(t_n - \mathbf{w}^\mathsf{T}\mathbf{x}_n)^2 \\
&= -N \log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (t_n - \mathbf{w}^\mathsf{T}\mathbf{x}_n)^2
\end{aligned}
$$

Some re-arranging...

$$
\begin{aligned}
p(T_n = t_n|\mathbf{w}, \mathbf{x}_n, \sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(t_n - \mathbf{w}^\mathsf{T}\mathbf{x}_n)^2\right\} \\
\log L &= \log \prod_{n=1}^{N} p(t_n|\mathbf{w}, \mathbf{x}_n, \sigma^2) \\
&= -N\log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\mathsf{T}\mathbf{x}_n)^2
\end{aligned}
$$

Some re-arranging...

$$
\begin{aligned}
p(T_n = t_n | \mathbf{w}, \mathbf{x}_n, \sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2}(t_n - \mathbf{w}^\mathsf{T}\mathbf{x}_n)^2 \right\} \\
\log L &= \log \prod_{n=1}^{N} p(t_n | \mathbf{w}, \mathbf{x}_n, \sigma^2) \\
&= -N \log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (t_n - \mathbf{w}^\mathsf{T}\mathbf{x}_n)^2
\end{aligned}
$$

- Looks familiar! To continue (good exercise):

$$
\frac{\partial \log L}{\partial \mathbf{w}} = 0, \quad \frac{\partial \log L}{\partial \sigma^2} = 0
$$

Optimum parameters

- optimum $\widehat{\mathbf{w}}$ is still

$$\widehat{\mathbf{w}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t}$$

Optimum parameters

- optimum $\widehat{\mathbf{w}}$ is still
$$\widehat{\mathbf{w}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$$

- optimum $\widehat{\sigma^2}$ is
$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})^\mathsf{T}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})$$

Optimum parameters

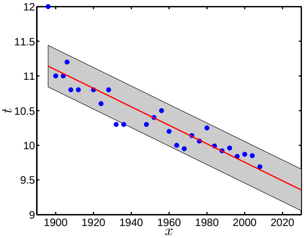
- optimum $\widehat{\mathbf{w}}$ is still

$$\widehat{\mathbf{w}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$$

- optimum $\widehat{\sigma^2}$ is

$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})^\mathsf{T}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})$$

- e.g. Olympic 100m data (again!)



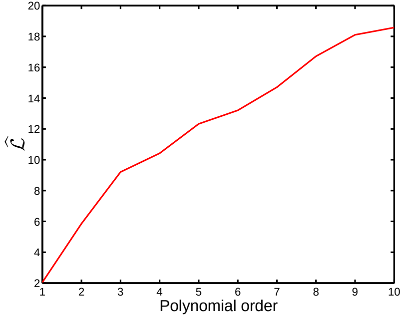
$$\widehat{\mathbf{w}} = \begin{bmatrix} 36.416 \\ -0.0133 \end{bmatrix}, \ \widehat{\sigma^2} = 0.0503$$

Can we use likelihood to choose models?

- We've already seen that training loss is no good for model choice
- Described cross-validation as an alternative
- Can we use the likelihood L or $\log L$?

Can we use likelihood to choose models?

- We've already seen that training loss is no good for model choice
- Described cross-validation as an alternative
- Can we use the likelihood L or $\log L$?

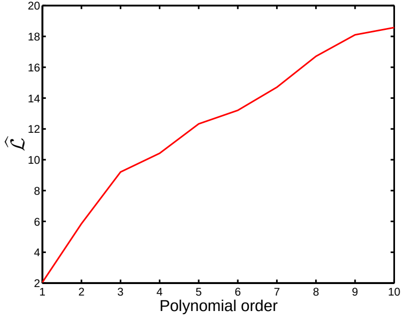


Data from 3rd order polynomial

- **No!**

Can we use likelihood to choose models?

- We've already seen that training loss is no good for model choice
- Described cross-validation as an alternative
- Can we use the likelihood L or $\log L$?

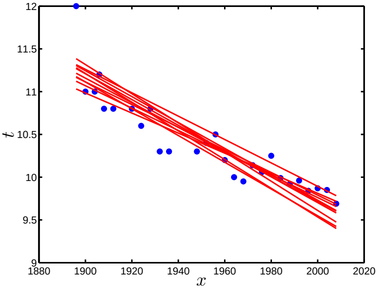


Data from 3rd order polynomial

- **No!**
  - More complex models can always get closer to the data
  - Results in lower $\widehat{\sigma^2}$ and higher likelihood

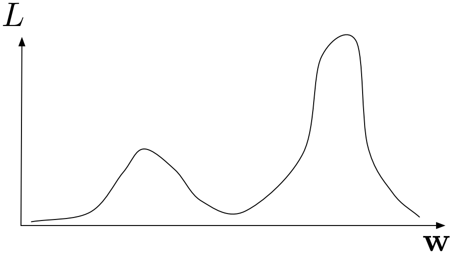
More uncertainty

- are different values of **w** (almost as) consistent with the data?
- is a different noise level σ^2 (almost as) consistent with the data?



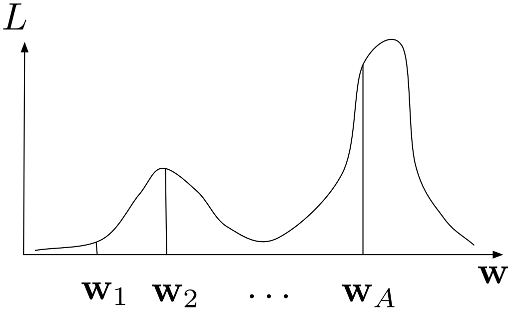
- **Bayesian** methods let us reason about different possible models

Problems with a point estimate



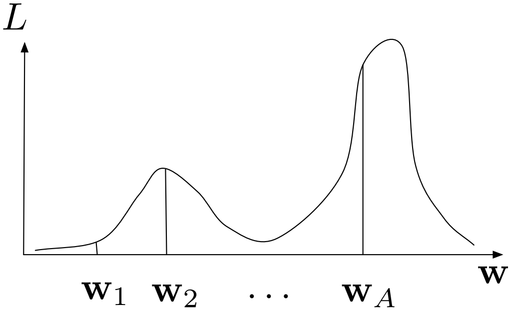
- might be more than one ‘best’ value
- might not be a single representative value
- different values might give very different predictions
  - ...but similar training loss

Averaging



- Prediction at $\mathbf{x}_{\text{new}}$ is some function of $\mathbf{w}$. Say $f(\mathbf{x}_{\text{new}}, \mathbf{w})$
- Choose A different values $\mathbf{w}_1, \ldots, \mathbf{w}_A$

Averaging



- Prediction at $\mathbf{x}_{\text{new}}$ is some function of $\mathbf{w}$. Say $f(\mathbf{x}_{\text{new}}, \mathbf{w})$
- Choose A different values $\mathbf{w}_1, \ldots, \mathbf{w}_A$
- Compute $\sum_{a=1}^{A} q_a f(\mathbf{x}_{\text{new}}, \mathbf{w}_a)$
- q_a is proportional to L (subject to $\sum_a q_a = 1$)
- Increasing A seems like a good idea...

Example

- Olympic 100m data
- Want to predict winning time at London 2012: $x_{\text{new}} = 2012$
- Choose two 'good' values of $\mathbf{w}$
  - $\mathbf{w}_1$ predicts $t_{\text{new}} = 9.5\ s$
  - $\mathbf{w}_2$ predicts $t_{\text{new}} = 9.2\ s$
- According to likelihood, $\mathbf{w}_2$ is twice as likely as $\mathbf{w}_1$
  - $q_1 + q_2 = 1$, $q_2 = 2q_1$
  - ...so $q_1 = 1/3$, $q_2 = 2/3$
- Average prediction is $(1/3) \times 9.5 + (2/3) \times 9.2 = 9.3\ s$

Averaging

- **Idea #1:** since $\mathbf{w}$ is uncertain, make it a random variable!
  - ...with density $p(\mathbf{w}|\text{stuff})$

Averaging

- **Idea #1:** since $\mathbf{w}$ is uncertain, make it a random variable!
  - ...with density $p(\mathbf{w}|\text{stuff})$
- **Idea #2:** Average over *every* value of $\mathbf{w}$!
- We do this with the following **expectation**:

$$\mathbf{E}_{p(\mathbf{w}|\text{stuff})}\left\{f(\mathbf{x}_{\text{new}}, \mathbf{w})\right\} = \int f(\mathbf{x}_{\text{new}}, \mathbf{w}) p(\mathbf{w}|\text{stuff})\, d\mathbf{w}$$

- An average of predictions from each possible $\mathbf{w}$ weighted by how likely that $\mathbf{w}$ value is

Averaging

- **Idea #1:** since $\mathbf{w}$ is uncertain, make it a random variable!
  - ...with density $p(\mathbf{w}|\text{stuff})$

- **Idea #2:** Average over *every* value of $\mathbf{w}$!
- We do this with the following **expectation**:

$$\mathbf{E}_{p(\mathbf{w}|\text{stuff})}\left\{f(\mathbf{x}_{\text{new}},\mathbf{w})\right\} = \int f(\mathbf{x}_{\text{new}},\mathbf{w})p(\mathbf{w}|\text{stuff})\;d\mathbf{w}$$

- An average of predictions from each possible $\mathbf{w}$ weighted by how likely that $\mathbf{w}$ value is

- But: what is 'stuff'? How do we compute $p(\mathbf{w}|\text{stuff})$?

Bayes' rule

- 'Stuff' should include data: $\mathbf{X}, \mathbf{t}$: $p(\mathbf{w}|\mathbf{X}, \mathbf{t})$
  - i.e. what we know about $\mathbf{w}$ after observing some data.
- We've seen something like this before: $p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2)$ – the likelihood
  - We'll ignore σ^2 for now

Bayes' rule

- 'Stuff' should include data: $\mathbf{X}, \mathbf{t}$: $p(\mathbf{w}|\mathbf{X}, \mathbf{t})$
  - i.e. what we know about $\mathbf{w}$ after observing some data.
- We've seen something like this before: $p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2)$ – the likelihood
  - We'll ignore σ^2 for now
- Can we use $p(\mathbf{t}|\mathbf{X}, \mathbf{w})$ to find $p(\mathbf{w}|\mathbf{X}, \mathbf{t})$?

Bayes' rule

- 'Stuff' should include data: $\mathbf{X}, \mathbf{t}$: $p(\mathbf{w}|\mathbf{X}, \mathbf{t})$
  - i.e. what we know about $\mathbf{w}$ after observing some data.
- We've seen something like this before: $p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2)$ – the likelihood
  - We'll ignore σ^2 for now
- Can we use $p(\mathbf{t}|\mathbf{X}, \mathbf{w})$ to find $p(\mathbf{w}|\mathbf{X}, \mathbf{t})$?
- **Bayes' rule:**

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

Bayes' rule

- 'Stuff' should include data: $\mathbf{X}, \mathbf{t}$: $p(\mathbf{w}|\mathbf{X}, \mathbf{t})$
  - i.e. what we know about $\mathbf{w}$ after observing some data.
- We've seen something like this before: $p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2)$ – the likelihood
  - We'll ignore σ^2 for now
- Can we use $p(\mathbf{t}|\mathbf{X}, \mathbf{w})$ to find $p(\mathbf{w}|\mathbf{X}, \mathbf{t})$?
- **Bayes' rule:**

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

- Comes from:

$$\begin{aligned} p(\mathbf{w}|\mathbf{X}, \mathbf{t})p(\mathbf{t}|\mathbf{X}) &= p(\mathbf{t}|\mathbf{w}, \mathbf{X})p(\mathbf{w}) \\ p(\mathbf{w}, \mathbf{t}|\mathbf{X}) &= p(\mathbf{w}, \mathbf{t}|\mathbf{X}) \end{aligned}$$

Bayes' rule

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

- **Posterior density**: $p(\mathbf{w}|\mathbf{X}, \mathbf{t})$
  - this is what we want – how likely a given model $\mathbf{w}$ is

Bayes' rule

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

- **Posterior density**: $p(\mathbf{w}|\mathbf{X},\mathbf{t})$
  - this is what we want – how likely a given model $\mathbf{w}$ is
- **Likelihood** : $p(\mathbf{t}|\mathbf{X},\mathbf{w})$
  - we've used this before – how likely the data is for a given model

Bayes' rule

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

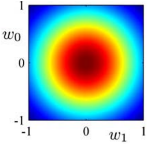
- **Posterior density**: $p(\mathbf{w}|\mathbf{X}, \mathbf{t})$
  - this is what we want – how likely a given model $\mathbf{w}$ is
- **Likelihood** : $p(\mathbf{t}|\mathbf{X}, \mathbf{w})$
  - we've used this before – how likely the data is for a given model
- **Prior density**: $p(\mathbf{w})$
  - this is new: represents our assumptions about 'sensible' model parameters

Bayes' rule

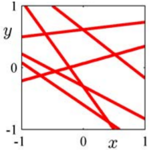
$$p(\mathbf{w}|\mathbf{X}, \mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{t}|\mathbf{X})}$$

- **Posterior density**: $p(\mathbf{w}|\mathbf{X}, \mathbf{t})$
  - this is what we want – how likely a given model $\mathbf{w}$ is
- **Likelihood** : $p(\mathbf{t}|\mathbf{X}, \mathbf{w})$
  - we've used this before – how likely the data is for a given model
- **Prior density**: $p(\mathbf{w})$
  - this is new: represents our assumptions about 'sensible' model parameters
- **Marginal likelihood**: $p(\mathbf{t}|\mathbf{X})$
  - this is is a normalisation constant ensuring $\int p(\mathbf{w}|\mathbf{X}, \mathbf{t})\, d\mathbf{w} = 1$

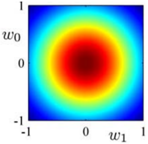
prior



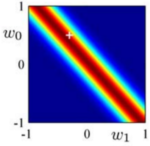
data & model



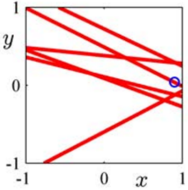
prior



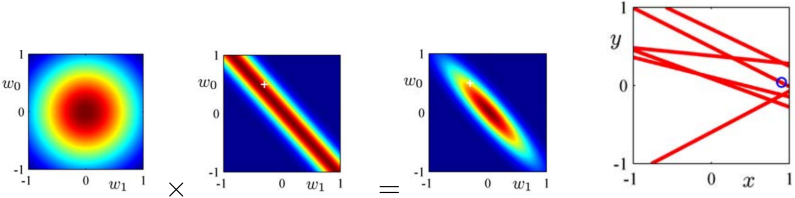
likelihood

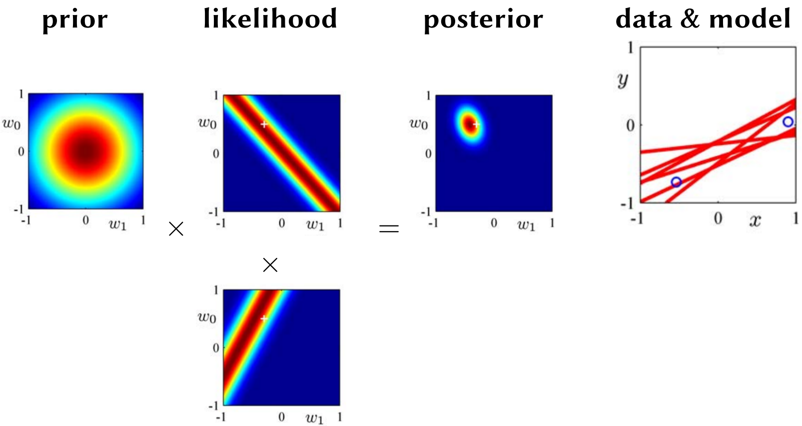


data & model



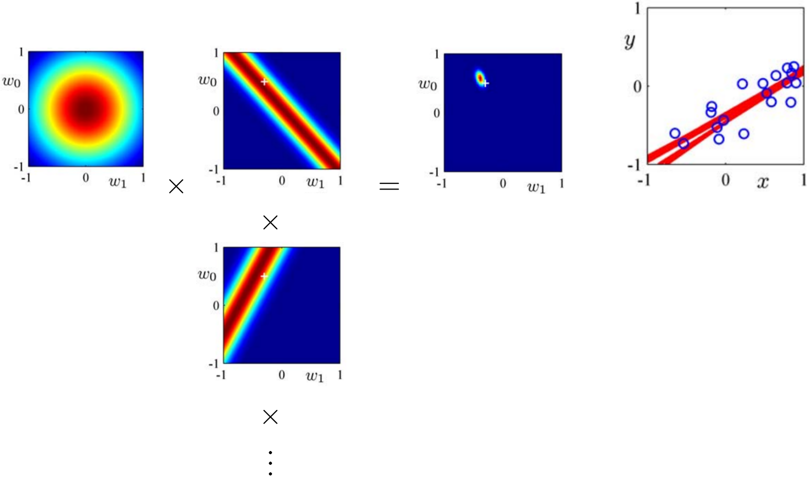
prior **likelihood** **posterior** **data & model**





prior
likelihood
posterior
data & model
1
w_0
0
-1
-1
0
w_1
1
×
×
=
y
x

prior **likelihood** **posterior** **data & model**



Computing the posterior

- Bayes rule:

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

- Unfortunately, computing the posterior is hard...
- ...because marginal likelihood $p(\mathbf{t}|\mathbf{X})$ is hard to compute:

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

- In some cases we can do it (this lecture)
- In most we can't and need some trick/alternative

When can we compute the posterior?

- Bayes rule:

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

- A prior $p(\mathbf{w})$ is said to be **conjugate** to a likelihood $p(\mathbf{t}|\mathbf{X}, \mathbf{w})$ if their product has the same type of density as the prior
- In our case: Gaussian prior $\times$ Gaussian likelihood gives Gaussian posterior

When can we compute the posterior?

- Bayes rule:

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

- A prior $p(\mathbf{w})$ is said to be **conjugate** to a likelihood $p(\mathbf{t}|\mathbf{X}, \mathbf{w})$ if their product has the same type of density as the prior
- In our case: Gaussian prior $\times$ Gaussian likelihood gives Gaussian posterior
  - Therefore, we **know** the form of the normalising constant
  - Therefore, we **don't need** to compute $p(\mathbf{t}|\mathbf{X})$

Example – Olympic data

- We'll use the (Gaussian) likelihood we used for maximum likelihood:
$$p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2) = f_{\mathcal{N}}(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I})$$

Example – Olympic data

- We'll use the (Gaussian) likelihood we used for maximum likelihood:

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

- $\mathbf{t}$ is a vector containing all the t_n
- $\mathbf{X}$ is a matrix containing all the $\mathbf{x}_n$
- Joint likelihood is given by

$$p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{w}, \mathbf{x}_n, \sigma^2)$$

Example – Olympic data

- We'll use the (Gaussian) likelihood we used for maximum likelihood:

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

- $\mathbf{t}$ is a vector containing all the t_n
- $\mathbf{X}$ is a matrix containing all the $\mathbf{x}_n$
- Joint likelihood is given by

$$p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{w}, \mathbf{x}_n, \sigma^2)$$

- Ignoring a constant, this is

$$p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2) \propto \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{t} - \mathbf{X}\mathbf{w})^\mathsf{T}(\mathbf{t} - \mathbf{X}\mathbf{w})\right\}$$

Example – Olympic data

- Choose a Gaussian prior:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \mathbf{S}),\ \mathbf{S} = \begin{bmatrix} 100 & 0 \\ 0 & 5 \end{bmatrix}$$

- Mean ($\mathbf{0}$) and covariance ($\mathbf{S}$) are design choices
- Lets us inject our own knowledge about what $\mathbf{w}$ are likely

Example – Olympic data

- Choose a Gaussian prior:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \mathbf{S}), \; \mathbf{S} = \left[\begin{array}{cc} 100 & 0 \\ 0 & 5 \end{array} \right]$$

- Mean ($\mathbf{0}$) and covariance ($\mathbf{S}$) are design choices
- Lets us inject our own knowledge about what $\mathbf{w}$ are likely

- Density is (ignoring a constant)

$$p(\mathbf{w}) \propto \exp\left\{-\frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{S}^{-1}\mathbf{w}\right\}$$

Example – Olympic data

- Ignoring non **w** terms, the prior multiplied by the likelihood is:

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

- Can be rearranged to (yet another) Gaussian
- It has parameters:

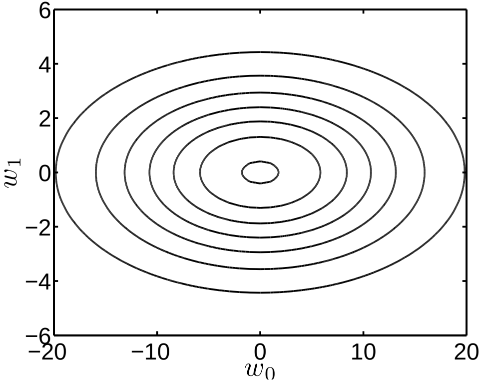
$$\boldsymbol{\mu} = \frac{1}{\sigma^2}\boldsymbol{\Sigma}\mathbf{X}^\mathsf{T}\mathbf{t} \qquad \boldsymbol{\Sigma} = \left(\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X}+\mathbf{S}^{-1}\right)^{-1}$$

Olympic data – Prior

- To make numbers better, rescale Olympic year:
  - $1896 = 1, 1900 = 2, \ldots, 2008 = 27, 2012 = 28$

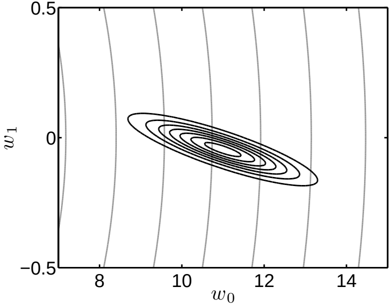
Olympic data – Prior

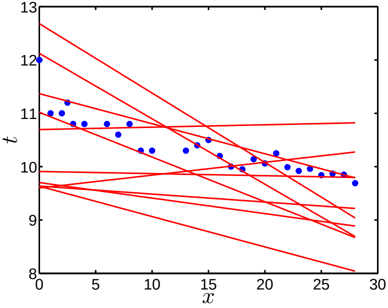
- To make numbers better, rescale Olympic year:
  - $1896 = 1, 1900 = 2, \ldots, 2008 = 27, 2012 = 28$
- Prior density:



- Mean ($\mathbf{0}$) and covariance ($\mathbf{S}$)
- Quite a *vague* prior

Olympic data – Posterior





- *Left*: posterior (black) and prior (grey), zoomed in
- *Right*: functions corresponding to some **w** sampled from posterior

Olympic data – Predictions

- Our motivation for being Bayesian was to be able to average predictions (at $\mathbf{x}_{\text{new}}$) over all $\mathbf{w}$:

$$\mathbf{E}_{p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma^2)}\left\{f(\mathbf{w})\right\} = \int f(\mathbf{w})p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2)\ d\mathbf{w}$$

Olympic data – Predictions

- Our motivation for being Bayesian was to be able to average predictions (at $\mathbf{x}_{\text{new}}$) over all $\mathbf{w}$:

$$\mathbf{E}_{p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma^2)}\left\{f(\mathbf{w})\right\} = \int f(\mathbf{w}) p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2)\ d\mathbf{w}$$

- For our model, $f(\mathbf{w})$ is another Gaussian

$$\mathcal{N}(\mathbf{w}^{\mathsf{T}}\mathbf{x}_{\text{new}},\sigma^2)$$

Olympic data – Predictions

- Our motivation for being Bayesian was to be able to average predictions (at $\mathbf{x}_{\text{new}}$) over all $\mathbf{w}$:

$$\mathbf{E}_{p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma^2)}\left\{f(\mathbf{w})\right\} = \int f(\mathbf{w})p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2)\;d\mathbf{w}$$

- For our model, $f(\mathbf{w})$ is another Gaussian

$$\mathcal{N}(\mathbf{w}^{\mathsf{T}}\mathbf{x}_{\text{new}},\sigma^2)$$

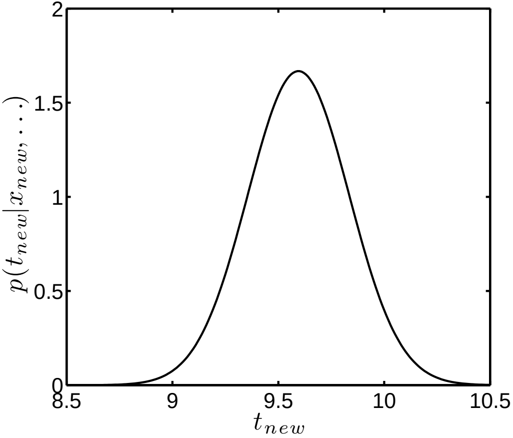
- We can compute this expectation exactly, to give predictive **density**:

$$p(t_{\text{new}}|\mathbf{X},\mathbf{t},\mathbf{x}_{\text{new}},\sigma^2) = \mathcal{N}(\mathbf{x}_{\text{new}}^{\mathsf{T}}\boldsymbol{\mu},\sigma^2 + \mathbf{x}_{\text{new}}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{x}_{\text{new}})$$

...where posterior parameters were:

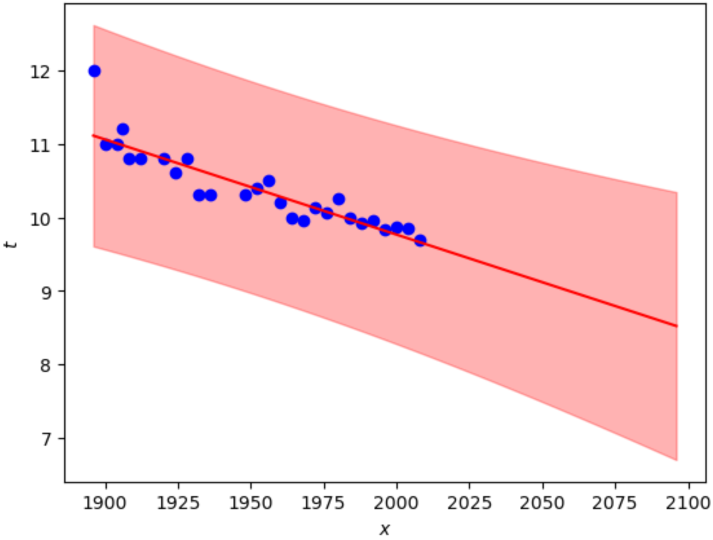
$$\boldsymbol{\mu} = \frac{1}{\sigma^2}\boldsymbol{\Sigma}\mathbf{X}^{\mathsf{T}}\mathbf{t} \qquad\qquad \boldsymbol{\Sigma} = \left(\frac{1}{\sigma^2}\mathbf{X}^{\mathsf{T}}\mathbf{X} + \mathbf{S}^{-1}\right)^{-1}$$

Olympic data – Predictions



Predictive density for 2012 Olympics

Olympic data – Predictions



Summary

- Moved away from a single parameter value
- Saw how predictions could be made by averaging over all possible parameter values – Bayesian
- Saw how Bayes' rule allows us to get a density for $\mathbf{w}$ conditioned on the data (and other stuff)

Summary

- Moved away from a single parameter value
- Saw how predictions could be made by averaging over all possible parameter values – Bayesian
- Saw how Bayes' rule allows us to get a density for $\mathbf{w}$ conditioned on the data (and other stuff)
- Computing the posterior is hard except in some cases...
- ...we can do it when things are *conjugate*