

Bayesian Regression

Morteza H. Chehreghani

`morteza.chehreghani@chalmers.se`

Chalmers University of Technology

April 2, 2020

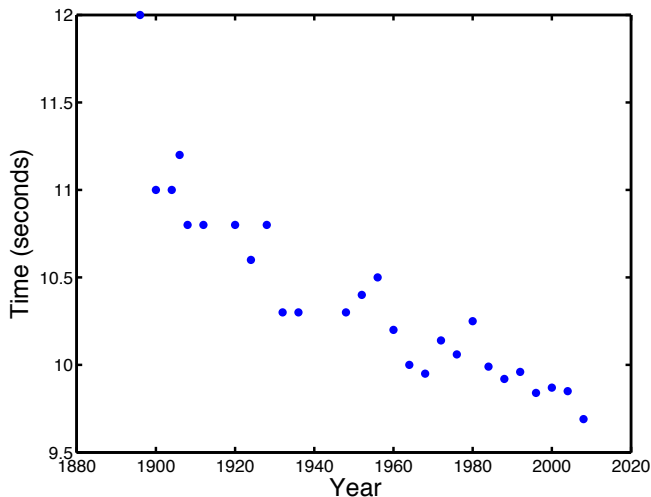
Reference

The content and the slides are adapted from

S. Rogers and M. Girolami, A First Course in Machine Learning (FCML), 2nd edition, Chapman & Hall/CRC 2016, ISBN: 9781498738484

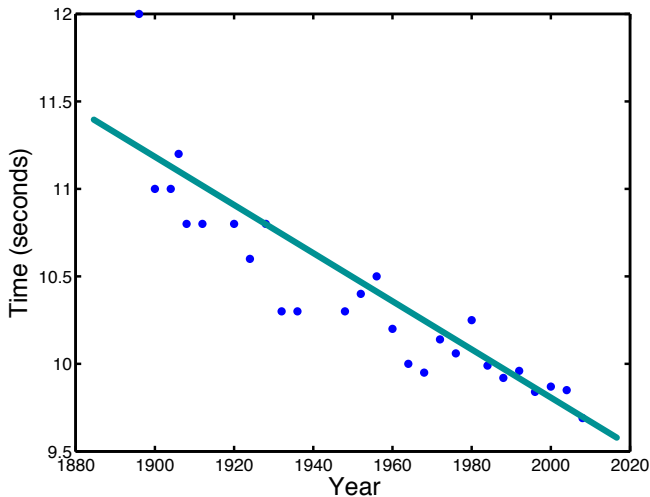
Some data and a problem

Predict the winning time for 2012!



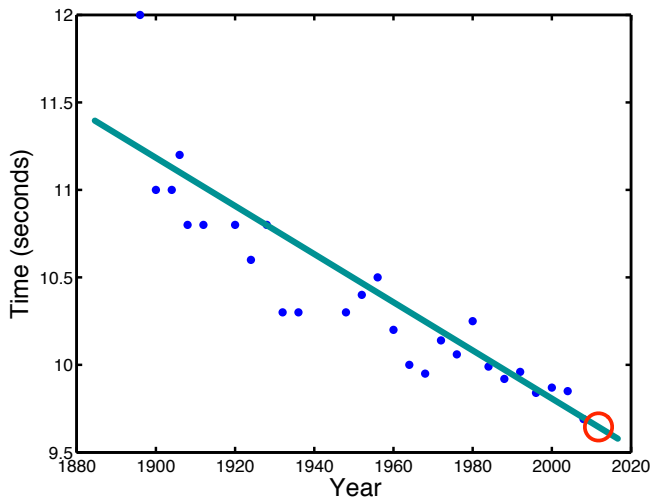
Some data and a problem

Fit a linear model (draw a line through the data)



Some data and a problem

Use the model (line) to *predict* the winning time in 2012.



Recipe for a linear model

More complex model: $t = w_0 + w_1x_1 + w_2x_2 + \dots + w_Dx_D$

$$\mathbf{x}_n = \begin{bmatrix} 1 \\ x_{n,1} \\ x_{n,2} \\ \vdots \\ x_{n,D} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,D} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \dots & x_{N,D} \end{bmatrix} \mathbf{t} = \begin{bmatrix} t_1 \\ t_n \\ \vdots \\ t_N \end{bmatrix},$$

Recipe for a linear model

More complex model: $t = w_0 + w_1x_1 + w_2x_2 + \dots + w_Dx_D$

$$\mathbf{x}_n = \begin{bmatrix} 1 \\ x_{n,1} \\ x_{n,2} \\ \vdots \\ x_{n,D} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,D} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \dots & x_{N,D} \end{bmatrix} \mathbf{t} = \begin{bmatrix} t_1 \\ t_n \\ \vdots \\ t_N \end{bmatrix},$$

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{bmatrix}, \quad \text{Model : } t_n = \mathbf{w}^T \mathbf{x}_n, \quad \text{or} \quad \mathbf{t} = \mathbf{X}\mathbf{w}$$

Recipe for linear model

$$\text{Model : } t_n = \mathbf{w}^T \mathbf{x}_n, \quad \text{or} \quad \mathbf{t} = \mathbf{X}\mathbf{w}$$

Usually, \mathbf{t} and $\mathbf{X}\mathbf{w}$ are not exactly equal. So, we try to minimise the difference.

$$\mathcal{L} = \frac{1}{N}(\mathbf{t} - \mathbf{X}\mathbf{w})^T(\mathbf{t} - \mathbf{X}\mathbf{w})$$

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

Recipe for a linear model

Model

$$t_n = \mathbf{w}^T \mathbf{x}_n, \quad \text{or} \quad \mathbf{t} = \mathbf{X} \mathbf{w}$$

Parameters

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

Prediction

$$\mathbf{x}_{\text{new}} = \begin{bmatrix} 1 \\ x_{\text{new},1} \\ x_{\text{new},2} \\ \vdots \\ x_{\text{new},D} \end{bmatrix}$$

then compute

$$t_{\text{new}} = \hat{\mathbf{w}}^T \mathbf{x}_{\text{new}}$$

Recipe for a *probabilistic* linear model

- In the probabilistic linear regression, we model the error, i.e.,

$$\text{Model : } t_n = \mathbf{w}^T \mathbf{x}_n + \epsilon_n, \quad \text{or} \quad \mathbf{t} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$$

In other words, we consider $p(t_n | \mathbf{w}, \mathbf{x}_n, \sigma^2) = \mathcal{N}(\mathbf{w}^T \mathbf{x}_n, \sigma^2)$

- The full likelihood is

$$p(\mathbf{t} | \mathbf{w}, \mathbf{X}, \sigma^2) = p(t_1, \dots, t_N | \mathbf{w}, \sigma^2, \mathbf{x}_1, \dots, \mathbf{x}_N)$$

- Note that

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

- And $p(\mathbf{t} | \mathbf{w}, \mathbf{X}, \sigma^2) = \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I})$
 \mathbf{I} is the identity matrix of size $N \times N$. The covariance matrix $\sigma^2 \mathbf{I}$ indicates i.i.d..

Recipe for a *probabilistic* linear model

- ▶ The full likelihood is

$$p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2) = p(t_1, \dots, t_N | \mathbf{w}, \sigma^2, \mathbf{x}_1, \dots, \mathbf{x}_N)$$

- ▶ We maximise the log-likelihood to obtain the parameters \mathbf{w} and σ^2 .
- ▶ Compute optimum $\hat{\mathbf{w}}$ from:

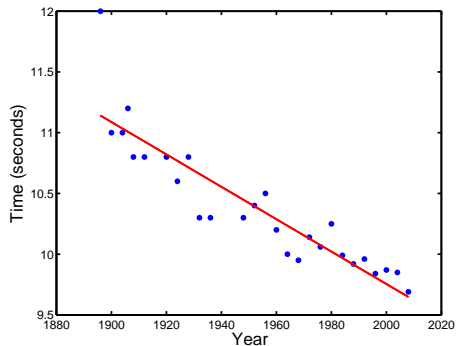
$$\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{t}$$

- ▶ Use this to compute optimum $\hat{\sigma}^2$ from:

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

Recipe for a *probabilistic* linear model

Olympic 100 m data (again!)



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

Recipe for a *probabilistic* linear model

Model

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

Parameters

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

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

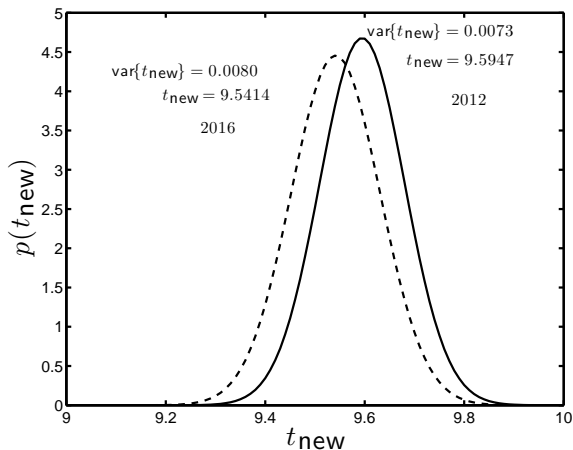
Prediction

$$t_{\text{new}} = \hat{\mathbf{w}}^T \mathbf{x}_{\text{new}}$$

$$\text{var}\{t_{\text{new}}\} = \hat{\sigma}^2 \mathbf{x}_{\text{new}}^T (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{x}_{\text{new}}$$

Hint: Always check the consistency of the dimensions (`numpy.shape()` in Python).

Olympic prediction



Predictive variance increases as we get further from the training data.

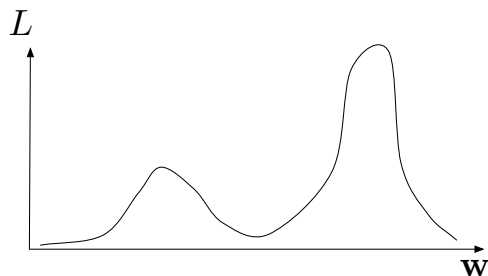
What is next?

- ▶ We have seen two ways of finding the ‘best’ parameter values:
 - ▶ Those that minimise the *loss* L .
 - ▶ Those that maximise the *likelihood* (probabilistic linear regression).
 - ▶ If the probabilistic model is Gaussian, both are the same:

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

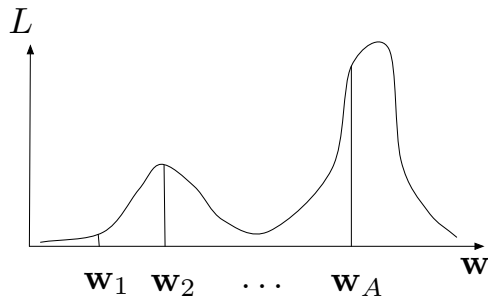
- ▶ In the probabilistic linear regression, we also estimate σ^2 .
- ▶ Is this the ‘right’ set of parameters?
- ▶ Is there a ‘right’ set of parameters?

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.
- ▶ Is there an alternative?

Averaging



- ▶ Prediction is some function of w . Say $f(w)$.
- ▶ Choose A different values – w_1, \dots, w_A .
- ▶ Compute $\sum_{a=1}^A q_a f(w_a)$
- ▶ q_a is proportional to L (subject to $\sum_a q_a = 1$)
- ▶ Note that each w_a is a vector.
- ▶ Increasing A seems like a good idea....

Example

- ▶ Olympic 100 m data.
- ▶ Want to predict winning time at London 2012 – t_{new} .
- ▶ Choose 2 '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 .
 - ▶ $q_1 + q_2 = 1$, $q_2 = 2q_1$.
 - ▶ Therefore: $q_1 = 1/3$, $q_2 = 2/3$
- ▶ Average prediction is $(1/3) \times 9.5 + (2/3) \times 9.2 = 9.3$

Averaging

- ▶ What if \mathbf{w} is a random variable with density $p(\mathbf{w}|\text{stuff})$?
- ▶ Imagine a weird die that chucks out values of \mathbf{w} .

Averaging

- ▶ What if \mathbf{w} is a random variable with density $p(\mathbf{w}|\text{stuff})$?
- ▶ Imagine a weird die that chucks out values of \mathbf{w} .
 - ▶ We can use every value of \mathbf{w} !
 - ▶ We do this with the following **expectation**:

$$\mathbf{E}_{p(\mathbf{w}|\text{stuff})} \{f(\mathbf{w})\} = \int f(\mathbf{w})p(\mathbf{w}|\text{stuff}) d\mathbf{w}$$

What is $f(\mathbf{w})$ is this course?

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

Averaging

- ▶ What if \mathbf{w} is a random variable with density $p(\mathbf{w}|\text{stuff})$?
- ▶ Imagine a weird die that chucks out values of \mathbf{w} .
 - ▶ We can use every value of \mathbf{w} !
 - ▶ We do this with the following **expectation**:

$$\mathbf{E}_{p(\mathbf{w}|\text{stuff})} \{f(\mathbf{w})\} = \int f(\mathbf{w})p(\mathbf{w}|\text{stuff}) d\mathbf{w}$$

What is $f(\mathbf{w})$ is this course?

- ▶ An average of predictions from each possible \mathbf{w} weighted by how likely that \mathbf{w} value is.
- ▶ 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.
 - ▶ For simplicity, we ignore σ^2 for now (we can assume its value is known).

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.
 - ▶ For simplicity, we ignore σ^2 for now (we can assume its value is known).
- ▶ 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.
 - ▶ For simplicity, we ignore σ^2 for now (we can assume its value is known).
- ▶ 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.
 - ▶ For simplicity, we ignore σ^2 for now (we can assume its value is known).
- ▶ 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

- 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

- ▶ 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're after.

Bayes rule

- ▶ 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're after.
- ▶ **Likelihood :** $p(\mathbf{t}|\mathbf{X}, \mathbf{w})$
 - ▶ We've used this before.

Bayes rule

- ▶ 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're after.

- ▶ **Likelihood :** $p(\mathbf{t}|\mathbf{X}, \mathbf{w})$

- ▶ We've used this before.

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

- ▶ This is new: do we know anything about the parameters before we see any data?

Bayes rule

- ▶ 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're after.

- ▶ **Likelihood :** $p(\mathbf{t}|\mathbf{X}, \mathbf{w})$

- ▶ We've used this before.

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

- ▶ This is new: do we know anything about the parameters before we see any data?

- ▶ **Marginal likelihood (or evidence or normalization):**
 $p(\mathbf{t}|\mathbf{X})$

- ▶ This is new: \mathbf{w} isn't in here. It is a normalisation constant.
Ensures $\int p(\mathbf{w}|\mathbf{X}, \mathbf{t}) d\mathbf{w} = 1$.

Computing the posterior

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

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

Computing the posterior

- ▶ Unfortunately, computing the posterior can be hard in general...
- ▶ ...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).

When can we compute the posterior?

Conjugacy (definition)

A prior $p(\mathbf{w})$ is said to be conjugate to a likelihood it results in a posterior of the same type of density as the prior.

- ▶ Example:
 - ▶ Prior: Gaussian; Likelihood: Gaussian; Posterior: Gaussian
 - ▶ Prior: Beta; Likelihood: Binomial; Posterior: Beta
 - ▶ Many others, e.g.
http://en.wikipedia.org/wiki/Conjugate_prior

Why is this important?

- ▶ Bayes rule:

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

- ▶ If prior and likelihood are conjugate, we **know** the form of $p(\mathbf{w}|\mathbf{X}, \mathbf{t})$
- ▶ Therefore, we **know** the form of the normalising constant.
- ▶ Therefore, we **don't need** to compute $p(\mathbf{t}|\mathbf{X})$

Why is this important?

- Bayes rule:

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

- If prior and likelihood are conjugate, we **know** the form of $p(\mathbf{w}|\mathbf{X}, \mathbf{t})$
- Therefore, we **know** the form of the normalising constant.
- Therefore, we **don't need** to compute $p(\mathbf{t}|\mathbf{X})$
- We just need to use some algebra to make $p(\mathbf{t}|\mathbf{X}, \mathbf{w})p(\mathbf{w})$ **look like** the correct density, ignoring all terms without \mathbf{w} .

Example - Olympic data

- ▶ Remember the (Gaussian) likelihood we used for maximum likelihood:

$$p(t|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{w}^\top \mathbf{x}_n, \sigma^2)$$

Example - Olympic data

- ▶ Remember the (Gaussian) likelihood we used for maximum likelihood:

$$p(t|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{w}^\top \mathbf{x}_n, \sigma^2)$$

- ▶ For the set of N observations (variables) $\{\mathbf{X}, \mathbf{t}\}$, we have

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

Example - Olympic data

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

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

- ▶ The prior conjugate to the Gaussian is Gaussian. So:

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

- ▶ Mean ($\mathbf{0}$) and covariance (\mathbf{S}) are design choices (prior knowledge).

Example - Olympic data

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

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

- ▶ The prior conjugate to the Gaussian is Gaussian. So:

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

- ▶ Mean ($\mathbf{0}$) and covariance (\mathbf{S}) are design choices (prior knowledge).
- ▶ Posterior **must be** Gaussian with unknown parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$:

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

Finding posterior parameters

- ▶ Ignoring normalising constant, the posterior is:

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

- ▶ We only care about the terms that are related to \mathbf{w} .

Finding posterior parameters

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

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

- Posterior (from previous slide):

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

Finding posterior parameters

- ▶ Equate individual terms on each side.
- ▶ Covariance:

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

- ▶ Mean:

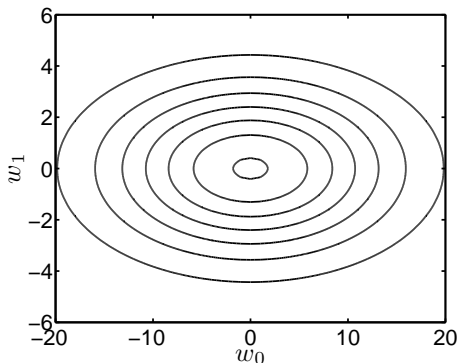
$$\begin{aligned}2\mathbf{w}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} &= \frac{2}{\sigma^2} \mathbf{w}^T \mathbf{X}^T \mathbf{t} \\ \hat{\boldsymbol{\mu}} &= \frac{1}{\sigma^2} \hat{\boldsymbol{\Sigma}} \mathbf{X}^T \mathbf{t}\end{aligned}$$

Olympic example

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

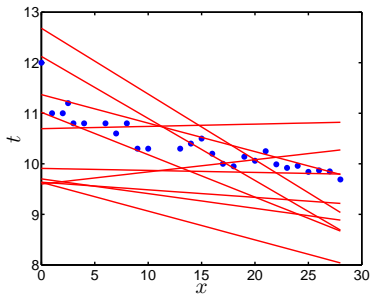
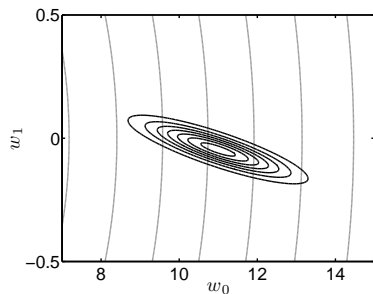
Olympic example

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



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

Olympic example



Posterior (left) (prior shown in grey, zoomed in) and functions corresponding to some \mathbf{w} sampled from posterior (right).

Olympic example – predictions

- ▶ Our motivation for being Bayesian was to be able to average predictions (at the test data \mathbf{x}_{new}) over all \mathbf{w}

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

- ▶ We have the full **posterior** distribution over all possible values of \mathbf{w} , it is also Gaussian and we computed the parameters.

Olympic example – predictions

- ▶ Our motivation for being Bayesian was to be able to average predictions (at the test data \mathbf{x}_{new}) over all \mathbf{w}

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

- ▶ We have the full **posterior** distribution over all possible values of \mathbf{w} , it is also Gaussian and we computed the parameters.
- ▶ We can compute exactly the **predictive density** to make **probabilistic predictions**:

$$\begin{aligned} p(t_{\text{new}}|\mathbf{X}, \mathbf{t}, \mathbf{x}_{\text{new}}, \sigma^2) &= \mathbf{E}_{p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \sigma^2)} \{p(t_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{w}, \sigma^2)\} \\ &= \int p(t_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{w}, \sigma^2) p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2) d\mathbf{w} \end{aligned}$$

Olympic example – predictions

- ▶ We can even compute exactly, the **predictive density** to make **probabilistic predictions**:

$$\begin{aligned} p(t_{\text{new}}|\mathbf{X}, \mathbf{t}, \mathbf{x}_{\text{new}}, \sigma^2) &= \mathbf{E}_{p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \sigma^2)} \{p(t_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{w}, \sigma^2)\} \\ &= \int p(t_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{w}, \sigma^2) p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2) d\mathbf{w} \end{aligned}$$

- ▶ $p(t_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{w}, \sigma^2)$ is defined by our model as the product of \mathbf{x}_{new} and \mathbf{w} with some additive Gaussian noise.

$$p(t_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{x}_{\text{new}}^T \mathbf{w}, \sigma^2)$$

- ▶ Because this expression and the posterior are both Gaussian, the result of expectation is another Gaussian.

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

Olympic example – predictions

- Therefore, the **predictive density** is

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

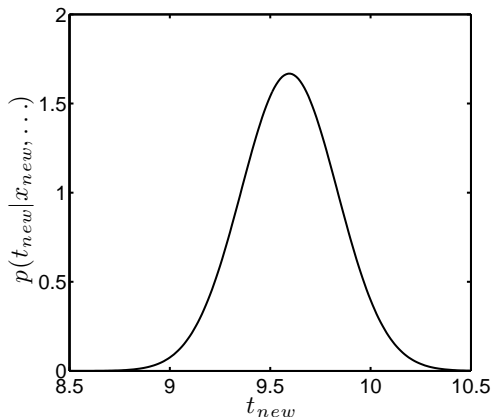
where,

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

and

$$\hat{\boldsymbol{\mu}} = \frac{1}{\sigma^2} \hat{\boldsymbol{\Sigma}} \mathbf{X}^T \mathbf{t}.$$

Olympic example – predictions



Predictive density at 2012 Olympics. Note that σ^2 was fixed at 0.05.

$$p(t_{new} | \mathbf{X}, \mathbf{t}, \mathbf{x}_{new}, \sigma^2) = \mathcal{N}(9.5951, 0.0572)$$

Computing posterior: recipe

- ▶ (Assuming prior conjugate to likelihood)
- ▶ Write down prior times likelihood (ignoring any constant terms, i.e., the term that are irrelevant to \mathbf{w})
- ▶ Write down posterior (ignoring any constant terms)
- ▶ Re-arrange them so they look like one another
- ▶ Equate terms on both sides to read off parameter values.

Choosing a prior

- ▶ How should we choose the prior?
 - ▶ Prior effect will diminish as more data arrive.
 - ▶ When we don't have much data, prior is very important.

Choosing a prior

- ▶ How should we choose the prior?
 - ▶ Prior effect will diminish as more data arrive.
 - ▶ When we don't have much data, prior is very important.
- ▶ Some influencing factors:
 - ▶ Data type: real, integer, string, etc.

Choosing a prior

- ▶ How should we choose the prior?
 - ▶ Prior effect will diminish as more data arrive.
 - ▶ When we don't have much data, prior is very important.
- ▶ Some influencing factors:
 - ▶ Data type: real, integer, string, etc.
 - ▶ Expert knowledge: 'the coin is fair', 'the model should be simple'

Choosing a prior

- ▶ How should we choose the prior?
 - ▶ Prior effect will diminish as more data arrive.
 - ▶ When we don't have much data, prior is very important.
- ▶ Some influencing factors:
 - ▶ Data type: real, integer, string, etc.
 - ▶ Expert knowledge: 'the coin is fair', 'the model should be simple'
 - ▶ Computational considerations (not as important as it used to be!)

Choosing a prior

- ▶ How should we choose the prior?
 - ▶ Prior effect will diminish as more data arrive.
 - ▶ When we don't have much data, prior is very important.
- ▶ Some influencing factors:
 - ▶ Data type: real, integer, string, etc.
 - ▶ Expert knowledge: 'the coin is fair', 'the model should be simple'
 - ▶ Computational considerations (not as important as it used to be!)
 - ▶ If we know nothing, can use a broad prior – e.g. uniform density.

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*.

Recipe for a *Bayesian* linear model

- ▶ In the Bayesian linear regression, we compute a distribution over \mathbf{w} instead of estimating it by $\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{t}$.

- ▶ The model is

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

- ▶ We use the Gaussian prior $p(\mathbf{w})$ and the likelihood $p(\mathbf{t} | \mathbf{w}, \mathbf{X}, \sigma^2) = \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I})$ to compute the model parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

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

and

$$\hat{\boldsymbol{\mu}} = \frac{1}{\sigma^2} \hat{\boldsymbol{\Sigma}} \mathbf{X}^T \mathbf{t}.$$

Recipe for a *Bayesian* linear model

- ▶ In the Bayesian linear regression, we compute a distribution over \mathbf{w} instead of estimating it by $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{t}$.
- ▶ The model is

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

- ▶ Prediction (**probabilistic predictions**)

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

where,

$$\hat{\boldsymbol{\Sigma}} = \left(\frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X} + \mathbf{S}^{-1} \right)^{-1}$$

and

$$\hat{\boldsymbol{\mu}} = \frac{1}{\sigma^2} \hat{\boldsymbol{\Sigma}} \mathbf{X}^\top \mathbf{t}.$$