Bayesian Regression

Morteza H. Chehreghani
morteza.chehreghani@chalmers.se

Chalmers Unniversity of Technology

January 26, 2024

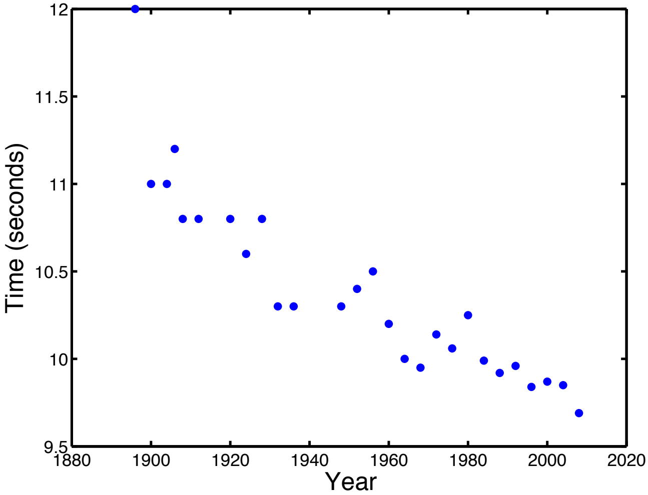
Reference

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

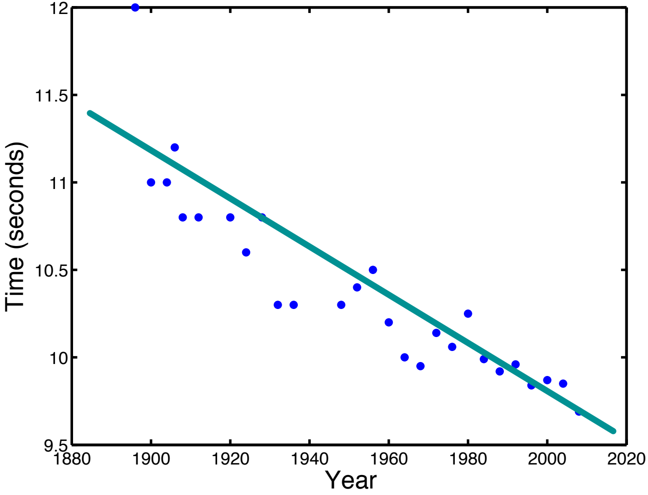
Some data and a problem

Predict the winning time for 2012!



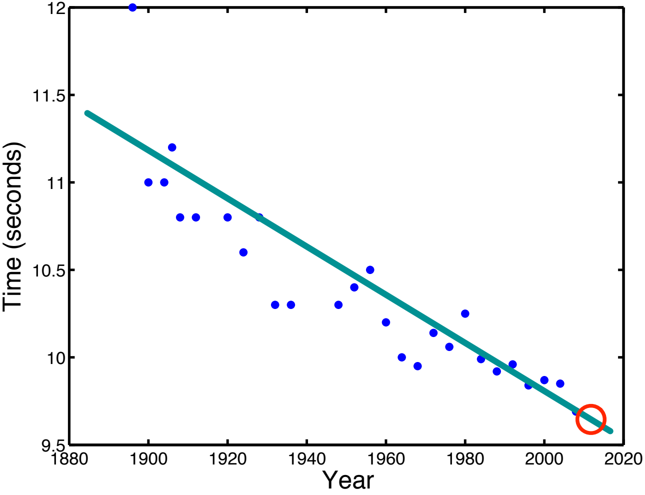
Some data and a problem

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

$$t_n = w_0 + w_1 x_{n,1} + w_2 x_{n,2} + w_3 x_{n,3} + \ldots + w_D x_{n,D}$$

$$\mathbf{x}_n = \begin{bmatrix} 1 \\ x_{n,1} \\ x_{n,2} \\ \vdots \\ x_{n,D} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \ldots & x_{1,D} \\ 1 & x_{2,1} & x_{2,2} & \ldots & x_{2,D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \ldots & x_{N,D} \end{bmatrix} \mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix},$$

Recipe for a linear model

$$t_n = w_0 + w_1 x_{n,1} + w_2 x_{n,2} + w_3 x_{n,3} + \ldots + w_D x_{n,D}$$

$$\mathbf{x}_n = \begin{bmatrix} 1 \\ x_{n,1} \\ x_{n,2} \\ \vdots \\ x_{n,D} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \ldots & x_{1,D} \\ 1 & x_{2,1} & x_{2,2} & \ldots & x_{2,D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \ldots & x_{N,D} \end{bmatrix} \mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix},$$

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

Recipe for linear model

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

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

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

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

Recipe for a linear model

Model

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

Parameters

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

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

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

Recipe for a *probabilistic* linear model

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

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

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

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

- Note that

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

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

Recipe for a *probabilistic* linear model

- The full likelihood is

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

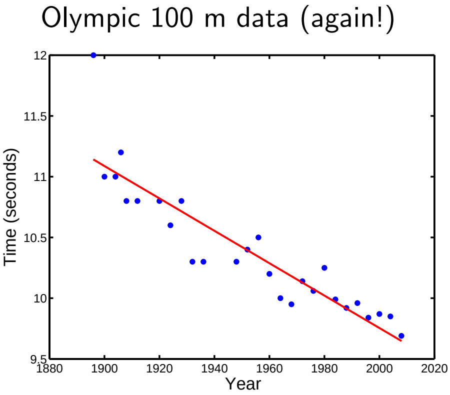
- We maximise the log-likelihood to obtain the parameters $\mathbf{w}$ and σ^2.
- Compute optimum $\widehat{\mathbf{w}}$ from:

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

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

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

Recipe for a *probabilistic* linear model



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

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

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

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

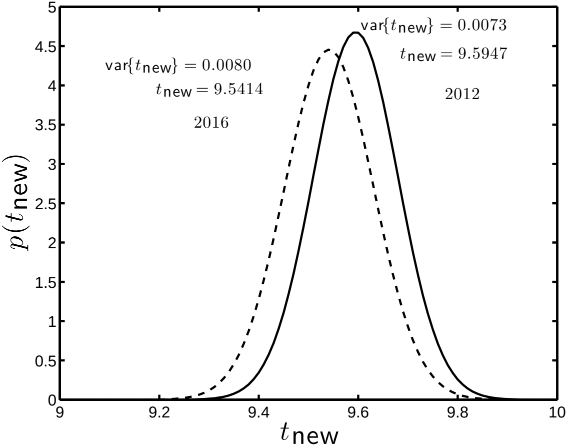
Prediction

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

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

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

Olympic prediction



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

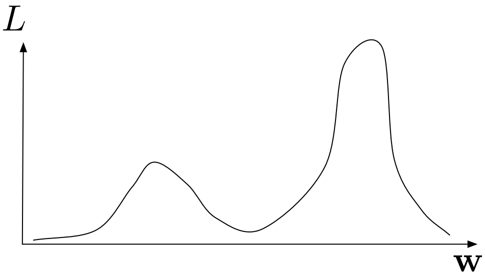
What is next?

- 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:

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

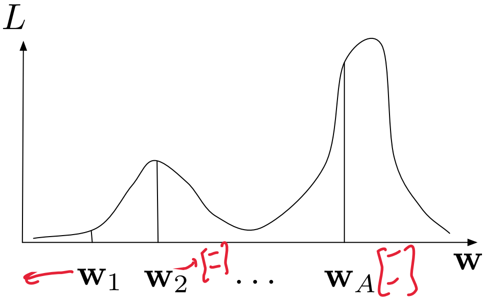
  - In the probabilistic linear regression, we also estimate σ^2.
- Is this the 'right' set of parameters?
- Is there a 'right' set of parameters?

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 $\mathbf{w}$. Say $f(\mathbf{w})$.
- Choose A different values – $\mathbf{w}_1, \ldots, \mathbf{w}_A$.
- Compute $\sum_{a=1}^{A} q_a f(\mathbf{w}_a)$ $w_A X_{new}$
- q_a is proportional to L (subject to $\sum_a q_a = 1$)
- Note that each $\mathbf{w}_a$ 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**:

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

  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**:

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

    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'? $\rightarrow X, t$
- How do we compute $p(\mathbf{w}|\underbrace{\text{stuff}}_{X,t})$?

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: $p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \cancel{\sigma^2})$ – the likelihood.
  - For simplicity, we ignore σ^2 for now (we can assume its value is known).

$p(t|w, X)$ Likelihood

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:

prior

posterior $\leftarrow$

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

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:

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

$$p(w) = p(w|X)$$

$$P(A,B) = P(A)P(B|A) \iff P(B|A) = \frac{P(A,B)}{P(A)}$$

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:

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

- **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:

$$\int p(\mathbf{w}|\mathbf{X},\mathbf{t})\,dw = \int \frac{\overbrace{p(\mathbf{t}|\mathbf{X},\mathbf{w})p(\mathbf{w})}^{p(t,w|X)}}{p(\mathbf{t}|\mathbf{X})}\,dw = \frac{\overbrace{\int p(t,w|X)\,dw}^{p(t|X)}}{p(t|X)}$$

- **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:

$$\overbrace{p(\mathbf{w}|\mathbf{X},\mathbf{t})}^{\text{Gaus.}} = \frac{\overbrace{p(\mathbf{t}|\mathbf{X},\mathbf{w})}^{\text{Gaus.}}\,\overbrace{p(\mathbf{w})}^{\text{Gaus.}}}{\cancel{p(\mathbf{t}|\mathbf{X})}}$$

- If prior and likelihood are conjugate, we **know** the form of $p(\mathbf{w}|\mathbf{X},\mathbf{t})$
- We know that the normalising constant does not have $\mathbf{w}$ terms.
- 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 $p(\mathbf{w}|\mathbf{X},\mathbf{t})$
- We know that the normalising constant does not have $\mathbf{w}$ terms.
- 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:

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

Example - Olympic data

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

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

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

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

  - 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:

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

- We only care about the terms that are related to **w**.

Finding posterior parameters

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

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

- Posterior (from previous slide):

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

$\Sigma^{-1} =$

Finding posterior parameters

- Equate individual terms on each side.
- Covariance:

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

- Mean:

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

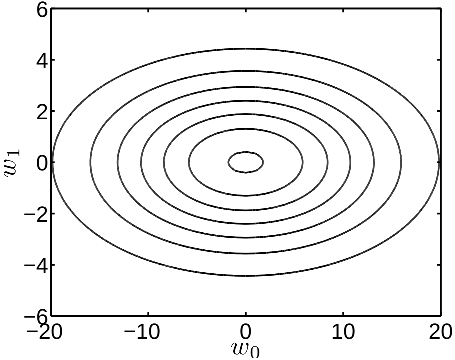
$p(w|X, t) = \mathcal{N}(\hat{\mu}, \hat{\Sigma})$

Olympic example

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

Olympic example

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

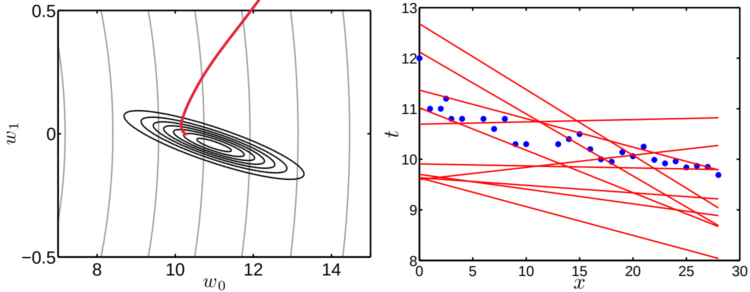


$p(w)$

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

Olympic example

$p(w \mid t, X) = \mathcal{N}(\hat{\mu}, \hat{\Sigma})$



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

Olympic example – predictions

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

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

- 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 $\mathbf{x}_{\text{new}}$) over all $\mathbf{w}$

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

- 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:

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

Olympic example – predictions

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

$$
\begin{aligned}
p(t_{\text{new}}|\mathbf{X},\mathbf{t},\mathbf{x}_{\text{new}},\sigma^2) &= \mathbf{E}_{p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma^2)}\left\{p(t_{\text{new}}|\mathbf{x}_{\text{new}},\mathbf{w},\sigma^2)\right\} \\
&= \int \underbrace{p(t_{\text{new}}|\mathbf{x}_{\text{new}},\mathbf{w},\sigma^2)}\underbrace{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2)}_{\text{posterior}: \mathcal{N}(\cdot,\cdot)}\,d\mathbf{w}
\end{aligned}
$$

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

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

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

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

Olympic example – predictions

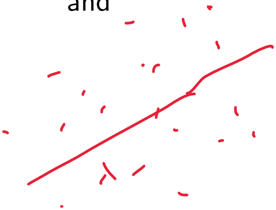
- Therefore, the predictive density is

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

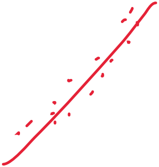
where,

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

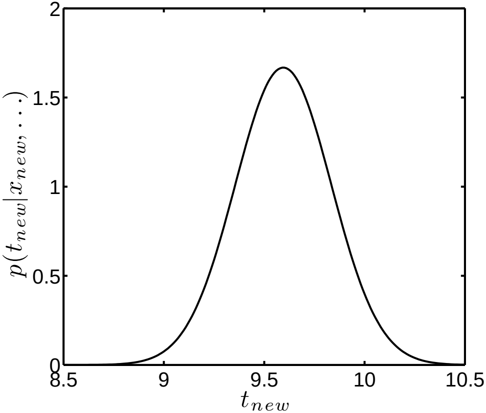
and



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



Olympic example – predictions



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

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

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

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

and

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

Recipe for a *Bayesian* linear model

- In the Bayesian linear regression, we compute a distribution over $\mathbf{w}$ instead of estimating it by $\widehat{\mathbf{w}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$.
- The model is
$$p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \sigma^2) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$
- Prediction (probabilistic predictions)
$$p(t_{\text{new}}|\mathbf{X}, \mathbf{t}, \mathbf{x}_{\text{new}}, \sigma^2) = \mathcal{N}(\mathbf{x}_{\text{new}}^\mathsf{T}\widehat{\boldsymbol{\mu}},\ \sigma^2 + \mathbf{x}_{\text{new}}^\mathsf{T}\widehat{\boldsymbol{\Sigma}}\mathbf{x}_{\text{new}})$$
where,
$$\widehat{\boldsymbol{\Sigma}} = \left(\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{S}^{-1}\right)^{-1}$$
and
$$\widehat{\boldsymbol{\mu}} = \frac{1}{\sigma^2}\widehat{\boldsymbol{\Sigma}}\mathbf{X}^\mathsf{T}\mathbf{t}.$$