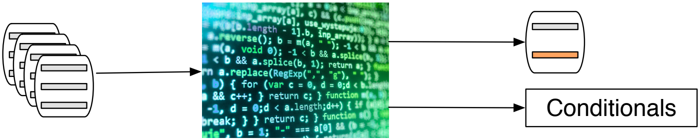
Machine Learning
A Bayesian View

Rajesh Ranganath

Last Class



$$p(y \mid \mathbf{x})$$

[Image of code from Atlantic]

Linear Regression

Model: linear functions

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\theta}^\top \mathbf{x}$$

Distance: Squared Error

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} d(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i)$$

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{\theta}^\top \mathbf{x}_i - y_i)^2$$

- Intercepts handled by including a column of 1 in $\mathbf{x}$

At the Highest Level

- Pick a loss function
- Pick a parametric ($\boldsymbol{\theta}$) model like linear functions
- Minimize the loss with respect to $\boldsymbol{\theta}$

Is optimization of a deterministic, parametric function based on a loss "learning"?

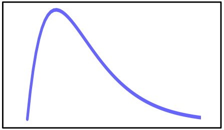
Is this the only way to think about learning?

The Bayesian Perspective

- Knowledge about the world encoded as a probability
- Data improves the knowledge of the world

A Sketch

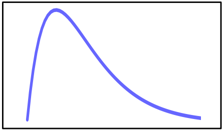
Belief



$\boldsymbol{\theta}_{sqft}$

A Sketch

Belief



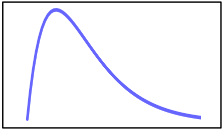
θ_{sqft}

Data

- 700 sqft, 800K
- 750 sqft, 1.25M
- 1842 sqft, 4.25M
- 3145 sqft, 3.5M

A Sketch

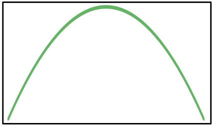
Belief



θ_{sqft}

- 700 sqft, 800K
- 750 sqft, 1.25M
- 1842 sqft, 4.25M
- 3145 sqft, 3.5M

New Belief



θ_{sqft}

A Formalism

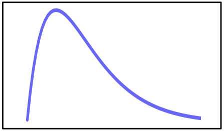
- *Prior*: $p(\boldsymbol{\theta})$
- *Likelihood*: $p(y \mid \boldsymbol{\theta}, \mathbf{x})$
- *Posterior*: $p(\boldsymbol{\theta} \mid y, \mathbf{x})$

A Formalism

- *Prior*: $p(\boldsymbol{\theta})$ — Prior belief over parameters
- *Likelihood*: $p(y \mid \boldsymbol{\theta}, \mathbf{x})$ — Assess data fit for specific prior
- *Posterior*: $p(\boldsymbol{\theta} \mid y, \mathbf{x})$ — Belief in parameters after seeing the data

A Formalism

Prior

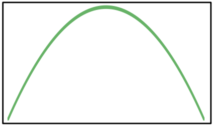


θ_{sqft}

Likelihood

- 700 sqft, 800K
- 750 sqft, 1.25M
- 1842 sqft, 4.25M
- 3145 sqft, 3.5M

Posterior



θ_{sqft}

A Formalism

- *Prior*: $p(\boldsymbol{\theta})$ – Does not depend on the features $\mathbf{x}$
- *Likelihood*: $p(y \mid \boldsymbol{\theta}, \mathbf{x})$
- *Posterior*: $p(\boldsymbol{\theta} \mid y, \mathbf{x})$

Overspecified?

Only Need Prior and Likelihood

A Formalism

Joint Distribution:

$$p(\boldsymbol{\theta}, y \mid \mathbf{x})$$

Decomposes to prior and likelihood

$$p(\boldsymbol{\theta}, y \mid \mathbf{x}) = \underbrace{p(\boldsymbol{\theta})}_{\text{prior}} \underbrace{p(y \mid \mathbf{x}, \boldsymbol{\theta})}_{\text{likelihood}}$$

Posterior

$$p(\boldsymbol{\theta} \mid y, \mathbf{x}) = \frac{p(\boldsymbol{\theta}, y \mid \mathbf{x})}{\int p(\boldsymbol{\theta}, y \mid \mathbf{x}) d\boldsymbol{\theta}}$$

- White + are the weights the data is sampled from
- Red lines are samples from the current belief
- Blue rings are data samples

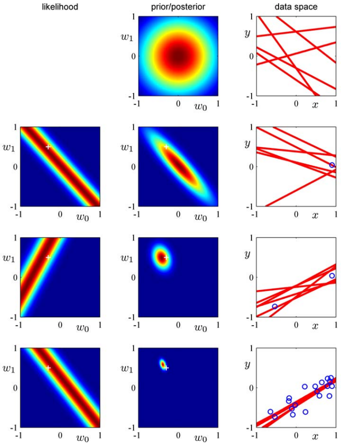


Figure 3.7 Illustration of sequential Bayesian learning for a simple linear model of the form $y(x, \mathbf{w}) = w_0 + w_1 x$. A detailed description of this figure is given in the text.

[Bishop 2006]

Making Predictions

Joint Distribution:

$$p(\boldsymbol{\theta}, y \mid \mathbf{x})$$

Posterior

$$p(\boldsymbol{\theta} \mid y, \mathbf{x}) = \frac{p(\boldsymbol{\theta}, y \mid \mathbf{x})}{\int p(\boldsymbol{\theta}, y \mid \mathbf{x}) d\boldsymbol{\theta}} = \frac{p(y \mid \mathbf{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int p(\boldsymbol{\theta}, y \mid \mathbf{x}) d\boldsymbol{\theta}}$$

How to predict on a new point $\mathbf{x}^*$?

Making Predictions

Joint Distribution:

$$p(\boldsymbol{\theta}, y \mid \mathbf{x})$$

Posterior

$$p(\boldsymbol{\theta} \mid y, \mathbf{x}) = \frac{p(\boldsymbol{\theta}, y \mid \mathbf{x})}{\int p(\boldsymbol{\theta}, y \mid \mathbf{x}) d\boldsymbol{\theta}} = \frac{p(y \mid \mathbf{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int p(\boldsymbol{\theta}, y \mid \mathbf{x}) d\boldsymbol{\theta}}$$

How to predict on a new point $\mathbf{x}^*$?
Probabilistic calculation

$$p(y^* \mid \mathbf{x}^*, \mathbf{x}, y)$$

Making Predictions

Joint Distribution:

$$p(\boldsymbol{\theta}, y \mid \mathbf{x})$$

Posterior

$$p(\boldsymbol{\theta} \mid y, \mathbf{x}) = \frac{p(\boldsymbol{\theta}, y \mid \mathbf{x})}{\int p(\boldsymbol{\theta}, y \mid \mathbf{x}) d\boldsymbol{\theta}} = \frac{p(y \mid \mathbf{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int p(\boldsymbol{\theta}, y \mid \mathbf{x}) d\boldsymbol{\theta}}$$

How to predict on a new point $\mathbf{x}^*$?
Probabilistic calculation

$$p(y^* \mid \mathbf{x}^*, \mathbf{x}, y) = \int p(y^* \mid \mathbf{x}^*, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid y, \mathbf{x}) d\boldsymbol{\theta}$$

Making Predictions

Joint Distribution:

$$p(\boldsymbol{\theta}, y \,|\, \mathbf{x})$$

Posterior

$$p(\boldsymbol{\theta} \,|y, \mathbf{x}) = \frac{p(\boldsymbol{\theta}, y \,|\, \mathbf{x})}{\int p(\boldsymbol{\theta}, y \,|\, \mathbf{x}) d\boldsymbol{\theta}} = \frac{p(y \,|\, \mathbf{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int p(\boldsymbol{\theta}, y \,|\, \mathbf{x}) d\boldsymbol{\theta}}$$

How to predict on a new point $\mathbf{x}^*$?
Probabilistic calculation

$$p(y^* \,|\, \mathbf{x}^*, \mathbf{x}, y) = \int p(y^* \,|\, \mathbf{x}^*, \boldsymbol{\theta}) p(\boldsymbol{\theta} \,|y, \mathbf{x}) d\boldsymbol{\theta}$$

What assumption did we make?

Observing another data point

Start with a prior

$$p(\boldsymbol{\theta})$$

Observe an $\mathbf{x}_1, y_1$, get a posterior

$$p(\boldsymbol{\theta} \mid \mathbf{x}_1, y_1)$$

What if we get another $\mathbf{x}_2, y_1$?

$$p(\boldsymbol{\theta} \mid \mathbf{x}_1, y_1, \mathbf{x}_2, y_2) = \frac{p(\boldsymbol{\theta} \mid y_1, \mathbf{x}_1) p(y_2 \mid \mathbf{x}_2, \boldsymbol{\theta})}{\int p(\boldsymbol{\theta} \mid y_1, \mathbf{x}_1) p(y_2 \mid \mathbf{x}_2, \boldsymbol{\theta}) d\boldsymbol{\theta}}$$

Posterior after one point became prior for second

Learning cast as probabilistic calculations

Was this just an intellectual exercise?

Where does a prior come from?

Model: linear functions

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\theta}^\top \mathbf{x}$$

Model: linear functions

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\theta}^\top \mathbf{x}$$

Simple example

- $\mathbf{x}$: p dimensional vector of features (house age, square feet, number of rooms)
- y: house price
- $\boldsymbol{\theta}$: p dimensional regression coefficients

Prior Information:

- House prices are bounded
- Coefficient for square-feet should be smaller than bound

Model: linear functions

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\theta}^\top \mathbf{x}$$

Complex example

- $\mathbf{x}$: Number of red blood cells

- y: blood volume

Physiology imposes restrictions on $\boldsymbol{\theta}$

Priors can save the day

Suppose we want to rank foods based on ratings (1-10)

- pizza: 9.8 (from 10,000 ratings)

- boiled potato: 2.3 (from 1490 ratings)

Priors can save the day

Suppose we want to rank foods based on ratings (1-10)

- pizza: 9.8 (from 10,000 ratings)
- boiled potato: 2.3 (from 1490 ratings)

New food comes: natto with one rating of 10

- Should it be ranked as the top food?

A peaked prior can resolve this. How?

Another Motivation

Suppose the goal is to

$$\min_{\hat{\boldsymbol{\theta}}} \mathbb{E}_{p(\boldsymbol{\theta})p(y\,|\,\mathbf{x},\boldsymbol{\theta})}[(\hat{\boldsymbol{\theta}}(y,\mathbf{x})-\boldsymbol{\theta})^2]$$

Expectation over possible "environments" and data from that environment

- Possible environments: $p(\boldsymbol{\theta})$
- Data from environments: $p(y\,|\,\boldsymbol{\theta},\mathbf{x})$

Another Motivation

Suppose the goal is to

$$\min_{\hat{\boldsymbol{\theta}}} \mathbb{E}_{p(\boldsymbol{\theta})p(y\,|\,\mathbf{x},\boldsymbol{\theta})}[(\hat{\boldsymbol{\theta}}(y,\mathbf{x}) - \boldsymbol{\theta})^2]$$

Expectation over possible “environments” and data from that environment

Best possible is

$$\boldsymbol{\theta}^*(y,\mathbf{x}) = \mathbb{E}[\boldsymbol{\theta}\,|y,\mathbf{x}]$$

Another Motivation

Suppose the goal is to

$$\min_{\hat{\boldsymbol{\theta}}} \mathbb{E}_{p(\boldsymbol{\theta})p(y\,|\,\mathbf{x},\boldsymbol{\theta})}[(\hat{\boldsymbol{\theta}}(y,\mathbf{x}) - \boldsymbol{\theta})^2]$$

Expectation over possible “environments” and data from that environment

Best possible is

$$\boldsymbol{\theta}^*(y,\mathbf{x}) = \mathbb{E}[\boldsymbol{\theta}\,|\,y,\mathbf{x}]$$

Posterior expectation minimizes loss

Another Motivation

Suppose the goal is to

$$\min_{\hat{\boldsymbol{\theta}}} \mathbb{E}_{p(\boldsymbol{\theta})p(y \mid \mathbf{x}, \boldsymbol{\theta})}[(\hat{\boldsymbol{\theta}}(y, \mathbf{x}) - \boldsymbol{\theta})^2]$$

Expectation over possible “environments” and data from that environment

Where does $\mathbf{x}$ come from?

Posterior is optimal? What happened last class?

A Conceptual Difference

Bayesian view

- World is a belief over parameters $\boldsymbol{\theta}$
- This is the prior
- Observe data from some $\boldsymbol{\theta}$ drawn from belief

More on this later

Frequentist view

- World has a fixed parameter $\boldsymbol{\theta}^*$
- Observe data from that fixed $\boldsymbol{\theta}^*$

Bayesian Linear Regression

Linear model

$$f(\mathbf{x}_i) = \boldsymbol{\theta}^\top \mathbf{x}_i$$

Needs a prior

Bayesian Linear Regression

Linear model

$$f(\mathbf{x}_i) = \boldsymbol{\theta}^\top \mathbf{x}_i$$

Needs a prior

$$\boldsymbol{\theta} \sim \mathrm{Normal}(0, 1)$$

Needs a likelihood

Bayesian Linear Regression

Linear model

$$f(\mathbf{x}_i) = \boldsymbol{\theta}^\top \mathbf{x}_i$$

Needs a prior

$$\boldsymbol{\theta} \sim \mathrm{Normal}(0, 1)$$

Needs a likelihood

$$y \,|\, \boldsymbol{\theta}, \mathbf{x} \sim \mathrm{Normal}(\boldsymbol{\theta}^\top \mathbf{x}, \sigma^2)$$

What about more than one data point?

Bayesian Linear Regression

Linear model

$$f(\mathbf{x}_i) = \boldsymbol{\theta}^\top \mathbf{x}_i$$

Needs a prior

$$\boldsymbol{\theta} \sim \text{Normal}(0, 1)$$

Needs a likelihood

$$y \mid \boldsymbol{\theta}, \mathbf{x} \sim \text{Normal}(\boldsymbol{\theta}^\top \mathbf{x}, \sigma^2)$$

What about more than one data point?

$$p(y_{1\ldots n} \mid \mathbf{x}_{1\ldots n}, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i \mid \boldsymbol{\theta}, \mathbf{x}_i)$$

Bayesian Linear Regression: Posterior Computation Sketch

Given n data points $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, compute

$$p(\boldsymbol{\theta} \,|\, y_{1\ldots n}, \mathbf{x}_{1\ldots n})$$

How?

Bayesian Linear Regression: Posterior Computation Sketch

Given n data points $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, compute

$$p(\boldsymbol{\theta} \mid y_{1\ldots n}, \mathbf{x}_{1\ldots n})$$

How? *Bayes Rule*

$$p(\boldsymbol{\theta} \mid y_{1\ldots n}, \mathbf{x}_{1\ldots n}) = \frac{p(\boldsymbol{\theta}, y_{1\ldots n} \mid \mathbf{x}_{1\ldots n})}{p(y_{1\ldots n} \mid \mathbf{x}_{1\ldots n})}$$

Does the denominator matter?

Bayesian Linear Regression: Posterior Computation Sketch

Given n data points $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, compute

$$p(\boldsymbol{\theta} \,|\, y_{1\ldots n}, \mathbf{x}_{1\ldots n})$$

How? *Bayes Rule*

$$p(\boldsymbol{\theta} \,|\, y_{1\ldots n}, \mathbf{x}_{1\ldots n}) = \frac{p(\boldsymbol{\theta}, y_{1\ldots n} \,|\, \mathbf{x}_{1\ldots n})}{p(y_{1\ldots n} \,|\, \mathbf{x}_{1\ldots n})}$$

Does the denominator matter? *Posterior proportional to joint*

Bayesian Linear Regression: Posterior Computation Sketch

$$p(\boldsymbol{\theta} \,|\, y_{1\ldots n}, \mathbf{x}_{1\ldots n}) \propto p(\boldsymbol{\theta}, y_{1\ldots n} \,|\, \mathbf{x}_{1\ldots n})$$

Substitute the model

$$p(\boldsymbol{\theta} \,|\, y_{1\ldots n}, \mathbf{x}_{1\ldots n}) \propto p(\boldsymbol{\theta}, y_{1\ldots n} \,|\, \mathbf{x}_{1\ldots n}) = \mathcal{N}(\boldsymbol{\theta}; 0, 1) \prod_{i=1}^{n} \mathcal{N}(y; \boldsymbol{\theta}^\top \mathbf{x}_i, \sigma^2)$$

Bayesian Linear Regression: Posterior Computation Sketch

$$p(\boldsymbol{\theta} \mid y_{1\ldots n}, \mathbf{x}_{1\ldots n}) \propto p(\boldsymbol{\theta}, y_{1\ldots n} \mid \mathbf{x}_{1\ldots n})$$

Substitute the model

$$p(\boldsymbol{\theta} \mid y_{1\ldots n}, \mathbf{x}_{1\ldots n}) \propto p(\boldsymbol{\theta}, y_{1\ldots n} \mid \mathbf{x}_{1\ldots n}) = \mathcal{N}(\boldsymbol{\theta}; 0, 1) \prod_{i=1}^{n} \mathcal{N}(y; \boldsymbol{\theta}^\top \mathbf{x}_i, \sigma^2)$$

Now fill out the functional forms

$$p(\boldsymbol{\theta} \mid y_{1\ldots n}, \mathbf{x}_{1\ldots n}) = C \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\theta}\right) \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - \boldsymbol{\theta}^\top \mathbf{x}_i\right)^2\right)$$

Bayesian Linear Regression: Posterior Computation Sketch

$$p(\boldsymbol{\theta} \,|\, y_{1\ldots n}, \mathbf{x}_{1\ldots n}) = C \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\theta}\right) \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - \boldsymbol{\theta}^\top \mathbf{x}_i\right)^2\right)$$

What distribution is this?

Bayesian Linear Regression: Posterior Computation Sketch

$$p(\boldsymbol{\theta} \,|\, y_{1\ldots n}, \mathbf{x}_{1\ldots n}) = C \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\theta}\right) \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - \boldsymbol{\theta}^\top \mathbf{x}_i\right)^2\right)$$

What distribution is this?

$$\begin{aligned} p(\boldsymbol{\theta} \,|\, y_{1\ldots n}, \mathbf{x}_{1\ldots n}) &= C \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\theta} + \sum_{i=1}^{n} -\frac{1}{2\sigma^2}\left(y_i - \boldsymbol{\theta}^\top \mathbf{x}_i\right)^2\right) \\ &= C \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\theta} + \sum_{i=1}^{n} -\frac{1}{2\sigma^2}\left(y_i^2 - 2y_i \boldsymbol{\theta}^\top \mathbf{x}_i + \mathbf{x}_i^\top \boldsymbol{\theta}\boldsymbol{\theta}^\top \mathbf{x}_i\right)\right) \end{aligned}$$

Distribution function of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}\boldsymbol{\theta}^\top$?

Bayesian Linear Regression: Posterior Computation Sketch

$$p(\boldsymbol{\theta} \,|\, y_{1\ldots n}, \mathbf{x}_{1\ldots n}) = C \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{\theta}\right) \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - \boldsymbol{\theta}^\top \mathbf{x}_i\right)^2\right)$$

What distribution is this?

$$\begin{aligned} p(\boldsymbol{\theta} \,|\, y_{1\ldots n}, \mathbf{x}_{1\ldots n}) &= C \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{\theta} + \sum_{i=1}^{n} -\frac{1}{2\sigma^2}\left(y_i - \boldsymbol{\theta}^\top \mathbf{x}_i\right)^2\right) \\ &= C \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{\theta} + \sum_{i=1}^{n} -\frac{1}{2\sigma^2}\left(y_i^2 - 2y_i\boldsymbol{\theta}^\top \mathbf{x}_i + \mathbf{x}_i^\top \boldsymbol{\theta}\boldsymbol{\theta}^\top \mathbf{x}_i\right)\right) \end{aligned}$$

Distribution function of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}\boldsymbol{\theta}^\top$?

Looks like a Normal!

Multivariate Gaussian

$$p(a;\mu,\Sigma) \propto \exp\left(-\frac{1}{2}(a-\mu)^\top\Sigma^{-1}(a-\mu)\right) = C\exp\left(-\frac{1}{2}(a^\top\Sigma^{-1}a - 2a^\top\Sigma^{-1}\mu)\right)$$

Define $\Sigma_n = \left(I + \frac{1}{\sigma^2}\sum_{i=1}^n x_i x_i^\top\right)^{-1}$

$$\begin{aligned}
p(\boldsymbol{\theta}\,|y_{1\ldots n},\mathbf{x}_{1\ldots n}) &= C\exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{\theta} + \sum_{i=1}^n -\frac{1}{2\sigma^2}\left(y_i^2 - 2y_i\boldsymbol{\theta}^\top\mathbf{x}_i + \mathbf{x}_i^\top\boldsymbol{\theta}\boldsymbol{\theta}^\top\mathbf{x}_i\right)\right) \\
&= C\exp\left(-\frac{1}{2}\left(\boldsymbol{\theta}^\top\boldsymbol{\theta} + \sum_{i=1}^n \frac{1}{\sigma^2}\left(-2y_i\boldsymbol{\theta}^\top\mathbf{x}_i + \mathbf{x}_i^\top\boldsymbol{\theta}\boldsymbol{\theta}^\top\mathbf{x}_i\right)\right)\right) \\
&= C\exp\left(-\frac{1}{2}\left(\boldsymbol{\theta}^\top I\boldsymbol{\theta} + \sum_{i=1}^n \frac{1}{\sigma^2}\left(-2y_i\boldsymbol{\theta}^\top\mathbf{x}_i + \boldsymbol{\theta}^\top\mathbf{x}_i\mathbf{x}_i^\top\boldsymbol{\theta}\right)\right)\right) \\
&= C\exp\left(-\frac{1}{2}\left(\boldsymbol{\theta}^\top\left(I + \frac{1}{\sigma^2}\sum_{i=1}^n x_i x_i^\top\right)\boldsymbol{\theta} - 2\boldsymbol{\theta}^\top\sum_{i=1}^n\frac{1}{\sigma^2}y_i\mathbf{x}_i\right)\right) \\
&= C\exp\left(-\frac{1}{2}\left(\boldsymbol{\theta}^\top\Sigma_n^{-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}^\top\Sigma_n^{-1}\Sigma_n\sum_{i=1}^n\frac{1}{\sigma^2}y_i\mathbf{x}_i\right)\right)
\end{aligned}$$

Matching $\boldsymbol{\theta}$ with a from above, $\mu_n = \Sigma_n\sum_{i=1}^n \frac{1}{\sigma^2}y_i\mathbf{x}_i$

Bayesian Linear Regression: Posterior

Posterior for Bayesian linear regression

$$p(\boldsymbol{\theta} \,|\, y_{1\dots n}, \mathbf{x}_{1\dots n}) = \mathrm{Normal}(\mu_n, \Sigma_n),$$

where

$$\Sigma_n = \left(I + \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X}\right)^{-1}$$
$$\mu_n = \Sigma_n \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{y}$$

- Do sizes work out?
- What happens with no data?
- What happens with lots of data?

Bayesian Linear Regression: Posterior With More Data

$$p(\boldsymbol{\theta} \,|\, y_{1\ldots n}, \mathbf{x}_{1\ldots n}) = \text{Normal}(\mu_n, \Sigma_n),$$

where

$$\Sigma_n = \left(I + \frac{1}{\sigma^2}\mathbf{X}^\top \mathbf{X}\right)^{-1}$$
$$\mu_n = \Sigma_n \frac{1}{\sigma^2}\mathbf{X}^\top \mathbf{y}$$

For large n

$$\Sigma_n = \left(I + \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1} \approx \left(\frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1} = \sigma^2\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1}$$

Bayesian Linear Regression: Posterior With More Data

$$p(\boldsymbol{\theta} \,|y_{1\ldots n}, \mathbf{x}_{1\ldots n}) = \mathrm{Normal}(\mu_n, \Sigma_n),$$

where

$$\Sigma_n = \left(I + \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X}\right)^{-1}$$

$$\mu_n = \Sigma_n \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{y}$$

For large n

$$\Sigma_n = \left(I + \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1} \approx \left(\frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1} = \sigma^2\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1}$$

$$\boldsymbol{\mu}_n = \Sigma_n\sigma^2\mathbf{X}^\top\mathbf{y} = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1}\sigma^2\mathbf{X}^\top\mathbf{y} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top y$$

Imagine m different schools

In each school collect:

- n_j students
- $\mathbf{x}_{i,j}$ student traits (gpa, school year, math classes)
- $y_{i,j}$ SAT score

Goal predict SAT scores in each school

Build one big model or build a different model for each school?

Fitting One Big Model

$$p(\boldsymbol{\theta}) = \text{Normal}(0, I)$$
$$p(y_{i,j} \mid \mathbf{x}_{i,j}, \boldsymbol{\theta}) = \text{Normal}(\boldsymbol{\theta}^\top \mathbf{x}_{i,j}, \sigma^2)$$

- Advantage: More data, posterior will be more certain
- Disadvantage: Coefficients may vary in each school

Fitting Many Individual Models

$$p(\boldsymbol{\theta}_j) = \text{Normal}(0, I)$$
$$p(y_{i,j} \mid \mathbf{x}_{i,j}, \boldsymbol{\theta}) = \text{Normal}(\boldsymbol{\theta}_j^\top \mathbf{x}_{i,j}, \sigma^2)$$

- Advantage: Each school can have own coefficients
- Disadvantage: Coefficients may vary in each school

Want something in between

Hierarchical Linear Regression

Idea: Change the prior on $p(\boldsymbol{\theta}_j)$ to relate groups

$$p(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}) = \text{Normal}(\boldsymbol{\theta}, I)$$
$$p(y_{i,j} \mid \mathbf{x}_{i,j}, \boldsymbol{\theta}) = \text{Normal}(\boldsymbol{\theta}_j^\top \mathbf{x}_{i,j}, \sigma^2)$$

Place prior on new parameter:

$$p(\boldsymbol{\theta}) = N(0, 1)$$

- $\boldsymbol{\theta}_j$ shrunk toward $\boldsymbol{\theta}$
- $\boldsymbol{\theta}$ posterior "averages" each $\boldsymbol{\theta}_j$

Hierarchical Linear Regression: Picking the Variances

$$p(\boldsymbol{\theta}) = \mathrm{Normal}(0, I)$$
$$p(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}) = \mathrm{Normal}(\boldsymbol{\theta}, I)$$
$$p(y_{i,j} \mid \mathbf{x}_{i,j}, \boldsymbol{\theta}) = \mathrm{Normal}(\boldsymbol{\theta}_j^\top \mathbf{x}_{i,j}, \sigma^2)$$

What controls the relationship between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_j$?

Hierarchical Linear Regression: Picking the Variances

$$p(\boldsymbol{\theta}) = \text{Normal}(0, I)$$
$$p(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}) = \text{Normal}(\boldsymbol{\theta}, I)$$
$$p(y_{i,j} \mid \mathbf{x}_{i,j}, \boldsymbol{\theta}) = \text{Normal}(\boldsymbol{\theta}_j^\top \mathbf{x}_{i,j}, \sigma^2)$$

What controls the relationship between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_j$?

Partly, it's the 1 in the I in $p(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}) = \text{Normal}(\boldsymbol{\theta}, I)$

Why 1? Can we do something else?

Hierarchical Linear Regression: Picking the Variances

$$
\begin{aligned}
p(\boldsymbol{\theta}) &= \text{Normal}(0, I) \\
p(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}) &= \text{Normal}(\boldsymbol{\theta}, I) \\
p(y_{i,j} \mid \mathbf{x}_{i,j}, \boldsymbol{\theta}) &= \text{Normal}(\boldsymbol{\theta}_j^\top \mathbf{x}_{i,j}, \sigma^2)
\end{aligned}
$$

What controls the relationship between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_j$?

Introduce $\tau_j \sim p(\tau_j)$. Constraints on $p(\tau_j)$?

Hierarchical Linear Regression: Picking the Variances

$$p(\boldsymbol{\theta}) = \text{Normal}(0, I)$$
$$p(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}) = \text{Normal}(\boldsymbol{\theta}, I)$$
$$p(y_{i,j} \mid \mathbf{x}_{i,j}, \boldsymbol{\theta}) = \text{Normal}(\boldsymbol{\theta}_j^\top \mathbf{x}_{i,j}, \sigma^2)$$

What controls the relationship between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_j$?

Introduce $\tau_j \sim p(\tau_j)$. Constraints on $p(\tau_j)$?

- Gamma
- Inverse Gamma
- Exponential
- Log-Normal, Log-T
- Half Normal, Half-T

Hierarchical Linear Regression: Picking the Variances

$$p(\boldsymbol{\theta}) = \text{Normal}(0, I)$$
$$p(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}) = \text{Normal}(\boldsymbol{\theta}, I)$$
$$p(y_{i,j} \mid \mathbf{x}_{i,j}, \boldsymbol{\theta}) = \text{Normal}(\boldsymbol{\theta}_j^\top \mathbf{x}_{i,j}, \sigma^2)$$

What controls the relationship between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_j$?

$$p(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}, \tau_j) = \text{Normal}(\boldsymbol{\theta}, \tau_j I)$$

Hierarchical Linear Regression: Predictions

Given a new point $\mathbf{x}_j^*$ in group j, how do we predict?

$$p(y_j^* \mid \mathcal{D}_1 \ldots \mathcal{D}_m, \mathbf{x}_j^*)$$
$$= \int p(y_j^* \mid \mathbf{x}_j^*, \boldsymbol{\theta}_j) p(\boldsymbol{\theta}_j \mid \mathcal{D}_j, \boldsymbol{\theta}, \tau_j) p(\tau_j \mid \boldsymbol{\theta}, \mathcal{D}_j) p(\boldsymbol{\theta} \mid \mathcal{D}_1 \ldots \mathcal{D}_m) d\boldsymbol{\theta}_j d\tau_j d\boldsymbol{\theta}$$

Again write down the probability of interest and compute it!

Simple Recipe

- Introduce all the hidden variables and integrate them
- Use independence assumptions to simplify

Hierarchical Linear Regression: Intuition

$$
\begin{aligned}
p(\boldsymbol{\theta}) &= \text{Normal}(0, 1) \\
p(\tau_j) &= \text{p} \\
p(\boldsymbol{\theta}_j \,|\, \boldsymbol{\theta}, \tau_j) &= \text{Normal}(\boldsymbol{\theta}, \tau_j) \\
p(y_{i,j} \,|\, \mathbf{x}_{i,j}, \boldsymbol{\theta}) &= \text{Normal}(\boldsymbol{\theta}_j^\top \mathbf{x}_{i,j}, \sigma^2)
\end{aligned}
$$

What happens if there's lots of data in one group?

Hierarchical Linear Regression: Intuition

$$
\begin{aligned}
p(\boldsymbol{\theta}) &= \text{Normal}(0, 1) \\
p(\tau_j) &= \text{p} \\
p(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}, \tau_j) &= \text{Normal}(\boldsymbol{\theta}, \tau_j) \\
p(y_{i,j} \mid \mathbf{x}_{i,j}, \boldsymbol{\theta}) &= \text{Normal}(\boldsymbol{\theta}_j^\top \mathbf{x}_{i,j}, \sigma^2)
\end{aligned}
$$

What happens if there's lots of data in one group?

$$
p(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}, \tau_j, \mathscr{D}_j) \propto p(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}, \tau_j) \prod_{i=1}^{n_j} p(y_{i,j} \mid \mathbf{x}_{i,j}, \boldsymbol{\theta}_j)
$$

Looks like linear regression for that group

Hierarchical Linear Regression: Intuition

$$
\begin{aligned}
p(\boldsymbol{\theta}) &= \mathrm{Normal}(0, 1) \\
p(\tau_j) &= \mathrm{p} \\
p(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}, \tau_j) &= \mathrm{Normal}(\boldsymbol{\theta}, \tau_j) \\
p(y_{i,j} \mid \mathbf{x}_{i,j}, \boldsymbol{\theta}) &= \mathrm{Normal}(\boldsymbol{\theta}_j^\top \mathbf{x}_{i,j}, \sigma^2)
\end{aligned}
$$

What happens if one group has little data?

Hierarchical Linear Regression: Intuition

$$
\begin{aligned}
p(\boldsymbol{\theta}) &= \mathrm{Normal}(0, 1) \\
p(\tau_j) &= \mathrm{p} \\
p(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}, \tau_j) &= \mathrm{Normal}(\boldsymbol{\theta}, \tau_j) \\
p(y_{i,j} \mid \mathbf{x}_{i,j}, \boldsymbol{\theta}) &= \mathrm{Normal}(\boldsymbol{\theta}_j^\top \mathbf{x}_{i,j}, \sigma^2)
\end{aligned}
$$

What happens if one group has little data?

$$
p(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}, \tau_j, \mathcal{D}_j) \propto p(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}, \tau_j) \prod_{i=1}^{n_j} p(y_{i,j} \mid \mathbf{x}_{i,j}, \boldsymbol{\theta}_j)
$$

Looks like prior given other groups $p(\boldsymbol{\theta} \mid \mathcal{D}_{-j})$

Hierarchical Linear Regression: Intuition

$$
\begin{aligned}
p(\boldsymbol{\theta}) &= \text{Normal}(0, 1) \\
p(\tau_j) &= \text{p} \\
p(\boldsymbol{\theta}_j \,|\, \boldsymbol{\theta}, \tau_j) &= \text{Normal}(\boldsymbol{\theta}, \tau_j) \\
p(y_{i,j} \,|\, \mathbf{x}_{i,j}, \boldsymbol{\theta}) &= \text{Normal}(\boldsymbol{\theta}_j^\top \mathbf{x}_{i,j}, \sigma^2)
\end{aligned}
$$

What happens if one group is really different?

Hierarchical Linear Regression: Intuition

$$
\begin{aligned}
p(\boldsymbol{\theta}) &= \mathrm{Normal}(0, 1) \\
p(\tau_j) &= \mathrm{p} \\
p(\boldsymbol{\theta}_j \,|\, \boldsymbol{\theta}, \tau_j) &= \mathrm{Normal}(\boldsymbol{\theta}, \tau_j) \\
p(y_{i,j} \,|\, \mathbf{x}_{i,j}, \boldsymbol{\theta}) &= \mathrm{Normal}(\boldsymbol{\theta}_j^\top \mathbf{x}_{i,j}, \sigma^2)
\end{aligned}
$$

What happens if one group is really different?

$$
p(\tau_j \,|\, \boldsymbol{\theta}_j, \boldsymbol{\theta}) \propto \tau_j^{-d/2} \exp\left(-\frac{1}{\tau_j} (\boldsymbol{\theta} - \boldsymbol{\theta}_j)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_j) \right)
$$

Encourages τ_j to get big

Hierarchical Linear Regression: Intuition

$$
\begin{aligned}
p(\boldsymbol{\theta}) &= \mathrm{Normal}(0, 1) \\
p(\tau_j) &= \mathrm{p} \\
p(\boldsymbol{\theta}_j \,|\, \boldsymbol{\theta}, \tau_j) &= \mathrm{Normal}(\boldsymbol{\theta}, \tau_j) \\
p(y_{i,j} \,|\, \mathbf{x}_{i,j}, \boldsymbol{\theta}) &= \mathrm{Normal}(\boldsymbol{\theta}_j^\top \mathbf{x}_{i,j}, \sigma^2)
\end{aligned}
$$

All handled by Bayesian computation!

The posterior is a distribution. Is that useful?

Posterior Credible Intervals

Posterior distribution

$$p(\boldsymbol{\theta} \mid \mathbf{x}, y)$$

- Can compute the cumulative distribution function to find where $\boldsymbol{\theta}$ lies with 95% probability under the posterior
- Provides range of likely $\boldsymbol{\theta}$
- Why 95%?

Thompson Sampling

- Imagine 10 different random lotteries sampled from Gaussians with unknown mean
- Collect data by pulling a particular arm
- Goal to maximize earnings

Thompson Sampling

- Imagine 10 different random lotteries sampled from Gaussians with unknown mean
- Collect data by pulling a particular arm
- Goal to maximize earnings

Strategy:

- Place prior on reward for each arm

$$r_j \sim \mathcal{N}(\theta_j, 1), \theta_j \sim \mathcal{N}(0, 1)$$

- Sample hypothetical expected rewards from posterior

$$\hat{\theta}_j \sim p(\theta_j \,|\, r_{j,1\ldots n_j})$$

- Pick largest $\hat{\theta}_j$

Balances exploration and exploitation

Confidence Intervals

Where's the uncertainty in standard linear regression?

Confidence Intervals

Where's the uncertainty in standard linear regression?

A Conceptual Difference

Bayesian view

- World is a belief over parameters $\boldsymbol{\theta}$
- This is the prior
- Observe data from some $\boldsymbol{\theta}$ drawn from belief
- Randomness inherent in belief about the world

Frequentist view

- World has a fixed parameter $\boldsymbol{\theta}^*$
- Observe data from that fixed $\boldsymbol{\theta}^*$
- Randomness comes from finite sampling

How do we decide between models in Bayesian way?

How do we decide between models in Bayesian way?

Assume two model classes Model_1, Model_2

$$p(\text{Model}_1 \mid \mathscr{D}) = \frac{p(\mathscr{D} \mid \text{Model}_1)p(\text{Model}_1)}{p(\mathscr{D})}$$
$$= \frac{p(\mathscr{D} \mid \text{Model}_1)p(\text{Model}_1)}{p(\mathscr{D} \mid \text{Model}_1)p(\text{Model}_1) + p(\mathscr{D} \mid \text{Model}_2)p(\text{Model}_2)}$$

- Only needs a prior on models
- Bigger model classes have to spread prior on more models
- A type of regularization

Bayesian computation has lots of advantages

- Composability
- Uncertainty
- Optimality under prior
- Matches Maximum Likelihood with large data

But why not use it everywhere?

Bayesian computation has lots of advantages

- Composability
- Uncertainty
- Optimality under prior
- Matches Maximum Likelihood with large data

But why not use it everywhere?

- Needs a prior
- Computation

Thinking about the data generating process does not mean things are "Bayesian"