

Machine Learning (CS 181):

5. Bayesian Methods and Linear Regression

David C. Parkes and Sasha Rush

Spring 2017

1 / 38

Contents

- 1 Introduction
- 2 Beta/Bernoulli model
- 3 Normal-Normal Model
- 4 Bayesian Linear Regression
- 5 Bayesian Model Selection

2 / 38

- 1 Introduction
- 2 Beta/Bernoulli model
- 3 Normal-Normal Model
- 4 Bayesian Linear Regression
- 5 Bayesian Model Selection

3 / 38

Overview: Regularization vs. Bayesian Methods

- A regularization penalty, such as ridge regression

$$\min_{\mathbf{w}} \mathbf{w}^\top \mathbf{x} + \lambda \mathbf{w}^\top \mathbf{w}$$

is one approach to avoid over-fitting. Use a validation set to choose penalty $\lambda > 0$. Effective, but a bit ad hoc and inflexible.

- A Bayesian approach puts a prior on parameters, and views data D as [evidence for updating our beliefs](#) (get a posterior).
- By changing the prior, we change the way we learn from data.

4 / 38

Review: Maximum Likelihood Estimation

- Start with a generative model of the data, $p(D|\mathbf{w})$. Select parameters that maximize the likelihood:

$$\mathbf{w}_{\text{MLE}} = \arg \max_{\mathbf{w}} p(D|\mathbf{w})$$

- Taking logs and negating, equivalent to minimizing loss function $\mathcal{L}_D(\mathbf{w}) = -\ln p(D|\mathbf{w})$.
- For linear regression, we model the target

$$y_i \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i, \beta^{-1}),$$

and $\mathbf{w}_{\text{MLE}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$. Will tend to over-fit.

5 / 38

Bayesian Basics

- View parameters \mathbf{w} as a random variable. Adopt a prior $p(\mathbf{w})$, and a generative model $p(D|\mathbf{w})$ (the likelihood of data D)
- Use Bayes rule to update posterior based on observed data:

$$p(\mathbf{w}|D) = \frac{p(D|\mathbf{w})p(\mathbf{w})}{p(D)} \propto p(D|\mathbf{w})p(\mathbf{w}).$$

- $p(D)$ is the marginal likelihood, obtained as $p(D) = \int_{\mathbf{w}} p(D|\mathbf{w})p(\mathbf{w})d\mathbf{w}$.
- Can do various things with the posterior:
 - Obtain the **maximum a posteriori** estimate, w_{MAP} , which maximizes $p(\mathbf{w}|D)$.
 - “Full Bayes” (or **posterior predictive**), which considers the uncertainty on \mathbf{w} when making a prediction.

6 / 38

Maximum A Posteriori Estimator

- In the MAP approach, we find

$$\mathbf{w}_{\text{MAP}} = \arg \max_{\mathbf{w}} p(\mathbf{w}|D) = \arg \max_{\mathbf{w}} p(D|\mathbf{w})p(\mathbf{w})$$

- Equivalent to minimizing loss function:

$$-\ln p(D|\mathbf{w}) - \ln p(\mathbf{w})$$

- A small prior corresponds to a large regularization penalty. Provides a principled approach to regularization.
- Note: MAP with uniform prior ($\ln p(\mathbf{w}) = \text{const}$) is equal to MLE.

7 / 38

Posterior predictive (Full Bayes)

- In the posterior predictive approach, we work with

$$\begin{aligned} p(y|D, \mathbf{x}) &= \int_{\mathbf{w}} p(y, \mathbf{w}|D, \mathbf{x}) d\mathbf{w} \\ &= \int_{\mathbf{w}} p(y|\mathbf{w}, D, \mathbf{x}) p(\mathbf{w}|D, \mathbf{x}) d\mathbf{w} \\ &= \int_{\mathbf{w}} \underbrace{p(y|\mathbf{w}, \mathbf{x})}_{\text{predictive distribution}} \underbrace{p(\mathbf{w}|D)}_{\text{posterior}} d\mathbf{w} \end{aligned}$$

- Tractable when posterior has simple form ([conjugate property](#)).
- Can also use sample-based approaches such as Markov chain Monte Carlo, or variational methods (out of scope).

8 / 38

The Prior as Data Processor

We can view Bayes rule, and the use of a prior, as providing a framework for processing data:

$$\text{prior} \rightarrow \text{data}^{(1)} \rightarrow \text{posterior} \rightarrow \text{data}^{(2)} \rightarrow \text{posterior} \rightarrow \dots$$

The posterior carries forward our current belief, ready to be used to “process” more data.

9 / 38

Example (Warm-up)

- Sample $x_i \in \{Cherry, Lime\}$, candy from an opaque bag.

- Generative model:

$$p(x|\theta) = \begin{cases} \theta & \text{if } x = Lime \\ (1 - \theta) & \text{if } x = Cherry \end{cases}$$

for parameter $\theta \in \{0, 0.25, 0.5, 0.75, 1\}$.

■	Prior	θ	0	0.25	0.5	0.75	1
		$p(\theta)$	0.1	0.2	0.4	0.2	0.1

- Data are $D = Lime, Lime, Lime, Lime, \dots$

- After 1 *Lime*, $\theta_{MLE} = 1$; $p(D|\theta = 1) = (1)^2 > p(D|\theta = 0.75) = (0.75)^2$

- θ_{MAP} is 0.5, 0.75, 1.0,; e.g., after *Lime, Lime*, we have $p(D|\theta = 0.75)p(\theta = 0.75) = (0.75)^2(0.2) > p(D|\theta = 1)p(\theta = 1) = (1)^2(0.1)$.

10 / 38

- 1 Introduction
- 2 Beta/Bernoulli model
- 3 Normal-Normal Model
- 4 Bayesian Linear Regression
- 5 Bayesian Model Selection

11 / 38

Bernoulli model

- Coin flip = Bernoulli distribution. '1' w.p. θ , '0' otherwise.

$$p(x|\theta) = \theta^x(1 - \theta)^{1-x}$$

- Likelihood function:

$$p(D|\theta) = \prod_{i=1}^n \theta^{x_i}(1 - \theta)^{1-x_i} = \theta^{n_1}(1 - \theta)^{n_0}$$

where n_1 is number 1s and n_0 is number 0s.

- Taking the log, we have $n_1 \ln \theta + n_0 \ln(1 - \theta)$. Optimizing:

$$\begin{aligned} \frac{d}{d\theta}[\cdot] &= \frac{n_1}{\theta} - \frac{n_0}{1 - \theta} = 0 \\ \Leftrightarrow \theta_{\text{MLE}} &= \frac{n_1}{n_0 + n_1}. \end{aligned}$$

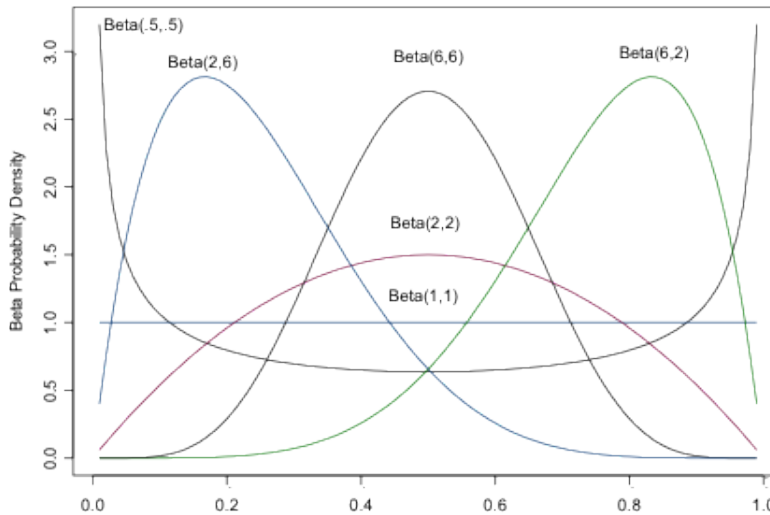
12 / 38

Bernoulli: Bayesian approach

Put a Beta prior on parameter θ , with probability density:

$$p(\theta) = \text{Beta}(\theta|\alpha, \beta) = \frac{1}{Z} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

where Z is a normalization constant, and $\alpha > 0, \beta > 0$.



Mean $\mathbb{E}[\theta] = \alpha/(\alpha + \beta)$; more peaked as $\alpha + \beta$ increases.

13 / 38

Beta-Bernoulli: Conjugate Pair

- Given Bernoulli likelihood and Beta prior, we have

$$\begin{aligned} p(\theta|D) &\propto p(D|\theta)p(\theta) \\ &= \theta^{n_1} (1-\theta)^{n_0} \theta^{\alpha-1} (1-\theta)^{\beta-1} \\ &= \theta^{n_1+\alpha-1} (1-\theta)^{n_0+\beta-1} \end{aligned}$$

- The posterior is:

$$p(\theta|D) = \text{Beta}(\theta|n_1 + \alpha, n_0 + \beta),$$

and the same form as the prior. This is the **conjugate** property.

- Beta-Bernoulli are a conjugate pair.
- Interpret α as the number of **pseudocounts** of 1s seen before, and β as the number of **pseudocounts** of 0s seen before.

14 / 38

Bernoulli-Beta: The MAP Estimate

- The mode of the Beta distribution is:

$$\arg \max_{\theta} \text{Beta}(\theta|\alpha, \beta) = \frac{\alpha - 1}{\alpha + \beta - 2}$$

for $\alpha > 1, \beta > 1$ (which we assume).

- Given posterior

$$p(\theta|D) = \text{Beta}(\theta|n_1 + \alpha, n_0 + \beta),$$

we have

$$\theta_{\text{MAP}} = \frac{\alpha + n_1 - 1}{\alpha + \beta + n_1 + n_0 - 2}.$$

- For example, if data are $D = 1, 1, 0$, then $\theta_{\text{MLE}} = n_1/n = 2/3$.

Given prior $\text{Beta}(2, 4)$, we have $\theta_{\text{MAP}} = \frac{2+2-1}{2+4+3-2} = \frac{3}{7}$.

15 / 38

Conjugate distributions

Definition (Conjugate property)

$p(\theta)$ is a **conjugate prior** on parameter θ for likelihood $p(D|\theta)$ if posterior $p(\theta|D)$ has the same form as the prior.

- Beta-Bernoulli form a conjugate pair.
- With a conjugate prior, we can easily use Bayes for data processing:

$$\text{prior} \rightarrow \text{data}^{(1)} \rightarrow \text{posterior} \rightarrow \text{data}^{(2)} \rightarrow \text{posterior} \rightarrow \dots$$

where the distributions on parameters are all from the same family.

- Other conjugate pair examples (all in the **exponential family**) are Gamma-Poisson, Dirichlet-Multinomial, and Normal-Normal.

16 / 38

- Can also compute the posterior predictive for a new example:

$$\begin{aligned} p(x = 1|D) &= \int_{\theta} p(x = 1|\theta)p(\theta|D)d\theta \\ &= \int_{\theta} \theta \cdot p(\theta|D)d\theta = \mathbb{E}_{\theta|D}[\theta] \\ &= \frac{\alpha + n_1}{\alpha + \beta + n_1 + n_0} \end{aligned}$$

- With $D = 1, 1, 0$ and prior $\text{Beta}(2, 4)$, this is

$$p(x = 1|D) = \frac{2 + 2}{2 + 4 + 3} = \frac{4}{9}$$

- Comparing with $P(x = 1|\theta_{\text{MAP}}) = 3/7$ and $P(x = 1|\theta_{\text{MLE}}) = 2/3$, this is inbetween, with $3/7 < 4/9 < 2/3$.

17 / 38

Contents

- 1 Introduction
- 2 Beta/Bernoulli model
- 3 Normal-Normal Model
- 4 Bayesian Linear Regression
- 5 Bayesian Model Selection

Warm-up: Univariate Normal

- $D = \{x_i\}_{i=1}^n$, with $x_i \in \mathbb{R}$.

- Generative model:

$$\mathcal{N}(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

given parameters μ, σ^2 .

- Maximum likelihood estimation (known σ^2):

$$\mu_{\text{MLE}} = \arg \max_{\mu} \sum_{i=1}^n \ln \mathcal{N}(x_i | \mu, \sigma^2) = \frac{\sum_{i=1}^n x_i}{n}$$

19 / 38

MAP Estimator for Univariate Normal

- Model $\mathcal{N}(x | \mu, \sigma^2)$. Assume variance known, and treat μ as a r.v.
- Conjugate pair for mean is Normal-Normal, and thus adopt $\mu \sim \mathcal{N}(m_0, s_0^2)$, for parameters m_0 and s_0^2 .
- After n examples, write posterior $\mu \sim \mathcal{N}(m_n, s_n^2)$. We have:

$$m_n = \frac{\sigma^2}{ns_0^2 + \sigma^2} m_0 + \frac{ns_0^2}{ns_0^2 + \sigma^2} \mu_{\text{MLE}} \quad (1)$$

$$s_n^2 = \left(\frac{1}{s_0^2} + \frac{n}{\sigma^2} \right)^{-1} \quad (2)$$

- Thus $\theta_{\text{MAP}} = m_n$ (since mode of Normal = mean). We see:
 - As $n \rightarrow \infty$, $\theta_{\text{MAP}} \rightarrow \mu_{\text{MLE}}$; As $s_0 \rightarrow \infty$, $\theta_{\text{MAP}} \rightarrow \mu_{\text{MLE}}$;
 - As $\sigma \rightarrow \infty$, $\theta_{\text{MAP}} \rightarrow m_0$.

20 / 38

Figuring out the Posterior (1 of 2)

Posterior

$$\begin{aligned} p(\mu|D) &\propto p(\mu)P(D|\mu) \\ &= \mathcal{N}(\mu|m_0, s_0^2) \prod_{i=1}^n \mathcal{N}(x_i|\mu, \sigma^2) \\ &= \frac{1}{\sqrt{2\pi s_0^2}} \exp\left(\frac{-(\mu - m_0)^2}{2s_0^2}\right) \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x_i - \mu)^2}{2\sigma^2}\right) \end{aligned}$$

Taking logs, and collecting constant terms, we have:

$$\ln p(\mu|D) \propto \text{const} - \frac{1}{2} \left[\frac{(\mu - m_0)^2}{s_0^2} + \sum_{i=1}^n \frac{(x_i - \mu)^2}{\sigma^2} \right]$$

21 / 38

Figuring out the Posterior (2 of 2)

Expand, fold terms that don't depend on μ into the constant, collect quadratic and linear terms:

$$p(\mu|D) \propto \text{const} - \frac{1}{2} \left[\mu^2 \left(\frac{1}{s_0^2} + \frac{n}{\sigma^2} \right) - 2\mu \left(\frac{m_0}{s_0^2} + \frac{\sum x_i}{\sigma^2} \right) \right]$$

Complete the square, moving additional terms into the constant

$$p(\mu|D) \propto \text{const} - \frac{1}{2} \left[\frac{(\mu - m_n)^2}{s_n^2} \right],$$

where

$$\frac{1}{s_n^2} = \left(\frac{1}{s_0^2} + \frac{n}{\sigma^2} \right); \quad m_n = \left(\frac{m_0}{s_0^2} + \frac{n \cdot \mu_{\text{MLE}}}{\sigma^2} \right).$$

22 / 38

Extension: Multivariate Normal

- $D = \{\mathbf{x}_i\}_{i=1}^n$, with $\mathbf{x}_i \in \mathbb{R}^m$

- Generative model:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right),$$

with known $\boldsymbol{\Sigma}$.

- Prior $\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{m}_0, \mathbf{S}_0)$, for parameters \mathbf{m}_0 and \mathbf{S}_0 .

- Posterior after n examples is $\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{m}_n, \mathbf{S}_n)$, and:

$$\mathbf{S}_n = (\mathbf{S}_0^{-1} + n\boldsymbol{\Sigma}^{-1})^{-1} \quad (3)$$

$$\mathbf{m}_n = \mathbf{S}_n (\mathbf{S}_0^{-1}\mathbf{m}_0 + n\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\text{MLE}}) \quad (4)$$

23 / 38

Interpretation of MAP estimator

- Posterior $\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{m}_n, \mathbf{S}_n)$, and:

$$\mathbf{S}_n = (\mathbf{S}_0^{-1} + n\boldsymbol{\Sigma}^{-1})^{-1}$$

$$\mathbf{m}_n = \mathbf{S}_n (\mathbf{S}_0^{-1}\mathbf{m}_0 + n\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\text{MLE}})$$

- Prior is overwhelmed as n gets bigger, with $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\text{MLE}}$ having a larger effect on \mathbf{S}_n and \mathbf{m}_n , respectively.
- With a strong prior, then \mathbf{S}_0 has small positive numbers on diagonal and inverse would have large numbers on diagonal, would “compete” with n to center the posterior mean at \mathbf{m}_0 instead of $\boldsymbol{\mu}_{\text{MLE}}$.
- Posterior cov. depends on data only through amount of data n . Wouldn't be case if $\boldsymbol{\Sigma}$ was also unknown. (See Bishop 2.3.6).

24 / 38

Figuring out the Posterior (v2)

Posterior $p(\boldsymbol{\mu}|D) \propto p(\boldsymbol{\mu})P(D|\boldsymbol{\mu})$. Taking logs, we have $\ln p(\boldsymbol{\mu}|D) =$

$$\text{const} - \frac{1}{2} \left[(\boldsymbol{\mu} - \mathbf{m}_0)^\top \mathbf{S}_0^{-1} (\boldsymbol{\mu} - \mathbf{m}_0) + \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right]$$

Expand, folding terms that don't depend on $\boldsymbol{\mu}$ into the constant:

$$= \text{const} - \frac{1}{2} \left[\boldsymbol{\mu}^\top \mathbf{S}_0^{-1} \boldsymbol{\mu} - 2\boldsymbol{\mu}^\top \mathbf{S}_0^{-1} \mathbf{m}_0 - 2\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \sum_{i=1}^n \mathbf{x}_i + n\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right]$$

Now collect quadratic and linear terms, and write $\sum_i \mathbf{x}_i = n\boldsymbol{\mu}_{\text{MLE}}$.

$$= \text{const} - \frac{1}{2} \left[\boldsymbol{\mu}^\top (\mathbf{S}_0^{-1} + n\boldsymbol{\Sigma}^{-1}) \boldsymbol{\mu} - 2\boldsymbol{\mu}^\top (\mathbf{S}_0^{-1} \mathbf{m}_0 + n\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{\text{MLE}}) \right]$$

Complete the square, moving additional terms into the constant

$$= \text{const} - \frac{1}{2} (\boldsymbol{\mu} - \mathbf{m}_n)^\top \mathbf{S}_n^{-1} (\boldsymbol{\mu} - \mathbf{m}_n),$$

where we can check that we obtain \mathbf{S}_n as in (3) and \mathbf{m}_n as in (4).

25 / 38

Contents

1 Introduction

2 Beta/Bernoulli model

3 Normal-Normal Model

4 Bayesian Linear Regression

5 Bayesian Model Selection

26 / 38

Bayesian Linear Regression

- $D = \{(\mathbf{x}_i, y)\}_{i=1}^n$, $\mathbf{x}_i \in \mathbb{R}^m$, $y_i \in \mathbb{R}$. Generative model:

$$y_i \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i, \beta^{-1}).$$

- Likelihood for data:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \beta^{-1}\mathbf{I})$$

- Put prior on weights \mathbf{w} , assume precision β known.

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

- Write posterior after n examples as $\mathbf{w} \sim \mathcal{N}(\mathbf{m}_n, \mathbf{S}_n)$. We show:

$$\mathbf{S}_n = \left(\mathbf{S}_0^{-1} + \beta \mathbf{X}^\top \mathbf{X}\right)^{-1} \quad (5)$$

$$\mathbf{m}_n = \mathbf{S}_n \left(\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \mathbf{X}^\top \mathbf{y}\right) \quad (6)$$

27 / 38

Interpretation of Bayesian LR MAP Estimator

Posterior $\mathbf{w} \sim \mathcal{N}(\mathbf{m}_n, \mathbf{S}_n)$, with:

$$\mathbf{S}_n = \left(\mathbf{S}_0^{-1} + \beta \mathbf{X}^\top \mathbf{X}\right)^{-1}$$

$$\mathbf{m}_n = \mathbf{S}_n \left(\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \mathbf{X}^\top \mathbf{y}\right)$$

- The MAP estimate is $\theta_{\text{MAP}} = \mathbf{m}_n$.
- With a weak prior, then \mathbf{S}_0 has large entries on the diagonal, and \mathbf{S}_0^{-1} is close to zero, and we have

$$\mathbf{S}_n \approx \beta^{-1}(\mathbf{X}^\top \mathbf{X})^{-1}$$

In addition, we have

$$\mathbf{m}_n \approx \beta^{-1}(\mathbf{X}^\top \mathbf{X})^{-1} \beta \mathbf{X}^\top \mathbf{y} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \theta_{\text{MLE}}$$

28 / 38

Special case: Simple Prior on Weights

- Suppose $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$. Posterior is $\mathbf{w} \sim \mathcal{N}(\mathbf{m}_n, \mathbf{S}_n)$, with

$$\mathbf{S}_n = (\alpha\mathbf{I} + \beta\mathbf{X}^\top\mathbf{X})^{-1}, \quad \mathbf{m}_n = \beta\mathbf{S}_n\mathbf{X}^\top\mathbf{y}.$$

- We see that

$$\mathbf{w}_{\text{MAP}} = \mathbf{m}_n = \beta(\alpha\mathbf{I} + \beta\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = (\mathbf{X}^\top\mathbf{X} + \frac{\alpha}{\beta}\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y},$$

and we recover ridge regression!

- Can also check the log posterior, which is $\ln \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) +$

$$\sum_{i=1}^n \ln \mathcal{N}(y_i|\mathbf{w}^\top\mathbf{x}_i, \beta^{-1}) = \text{const} - \frac{\alpha}{2}\mathbf{w}^\top\mathbf{w} - \frac{\beta}{2} \sum_{i=1}^n (y_i - \mathbf{w}^\top\mathbf{x}_i)^2,$$

and takes form of ridge penalty plus sum-of-squares error.

29 / 38

Figuring out the Posterior (v3!)

Posterior $p(\mathbf{w}|D) \propto p(\mathbf{w})p(\mathbf{y}|\mathbf{X}, \mathbf{w})$. Taking logs, expanding and folding terms that don't depend on \mathbf{w} into the constant, $\ln p(\mathbf{w}|D) =$

$$\begin{aligned} & \text{const} - \frac{1}{2} \left[(\mathbf{w} - \mathbf{m}_0)^\top \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0) + \beta (\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w}) \right] \\ & = \text{const} - \frac{1}{2} \left[\mathbf{w}^\top \mathbf{S}_0^{-1} \mathbf{w} - 2\mathbf{w}^\top \mathbf{S}_0^{-1} \mathbf{m}_0 - 2\beta \mathbf{w}^\top \mathbf{X}^\top \mathbf{y} + \beta \mathbf{w}^\top \mathbf{X}^\top \mathbf{X} \mathbf{w} \right] \end{aligned}$$

Collecting the quadratic and linear terms:

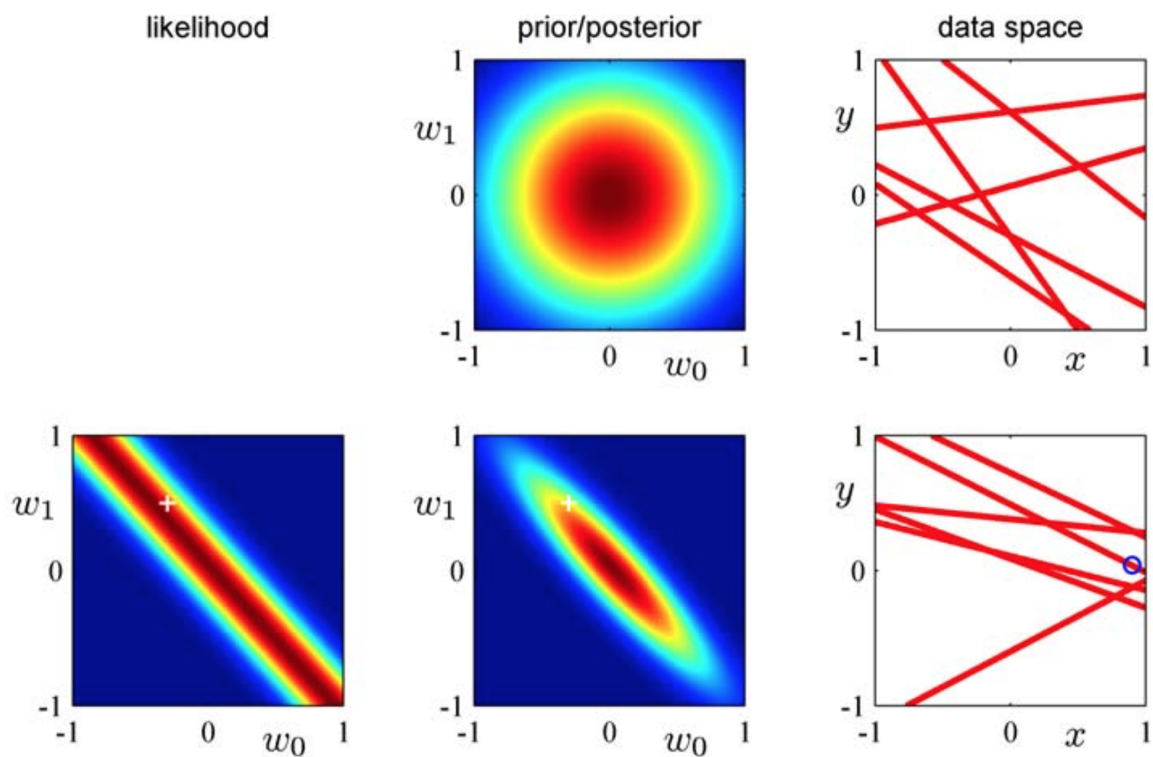
$$= \text{const} - \frac{1}{2} \left[\mathbf{w}^\top \left(\mathbf{S}_0^{-1} + \beta \mathbf{X}^\top \mathbf{X} \right) \mathbf{w} - 2\mathbf{w}^\top \left(\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \mathbf{X}^\top \mathbf{y} \right) \right]$$

Completing the square, we have:

$$= \text{const} - \frac{1}{2} (\mathbf{w} - \mathbf{m}_n)^\top \mathbf{S}_n^{-1} (\mathbf{w} - \mathbf{m}_n),$$

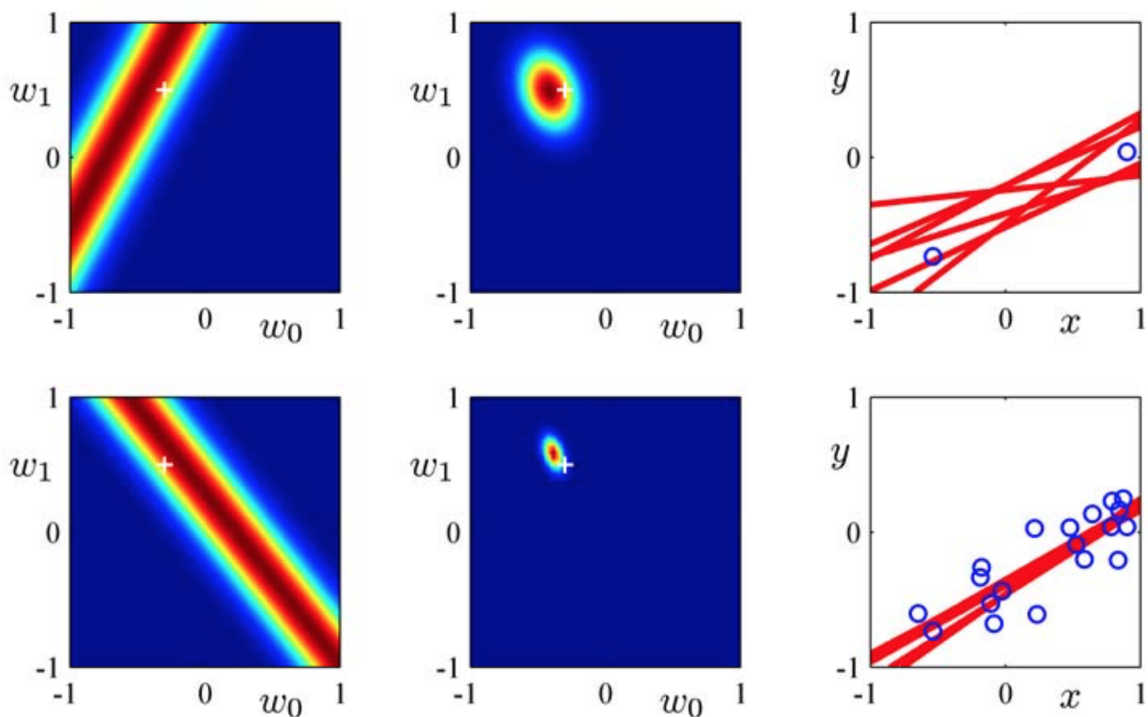
where we can check that we obtain \mathbf{S}_n as in (5) and \mathbf{m}_n as in (6).

30 / 38



(Bishop) w_0 offset. First example, see likelihood, product with prior giving new posterior, and new sample of possible relationships.

31 / 38



(Bishop) Observe second data point, see likelihood, product with most recent posterior giving new posterior, and new sample of possible relationships. Finally after 20 examples.

32 / 38

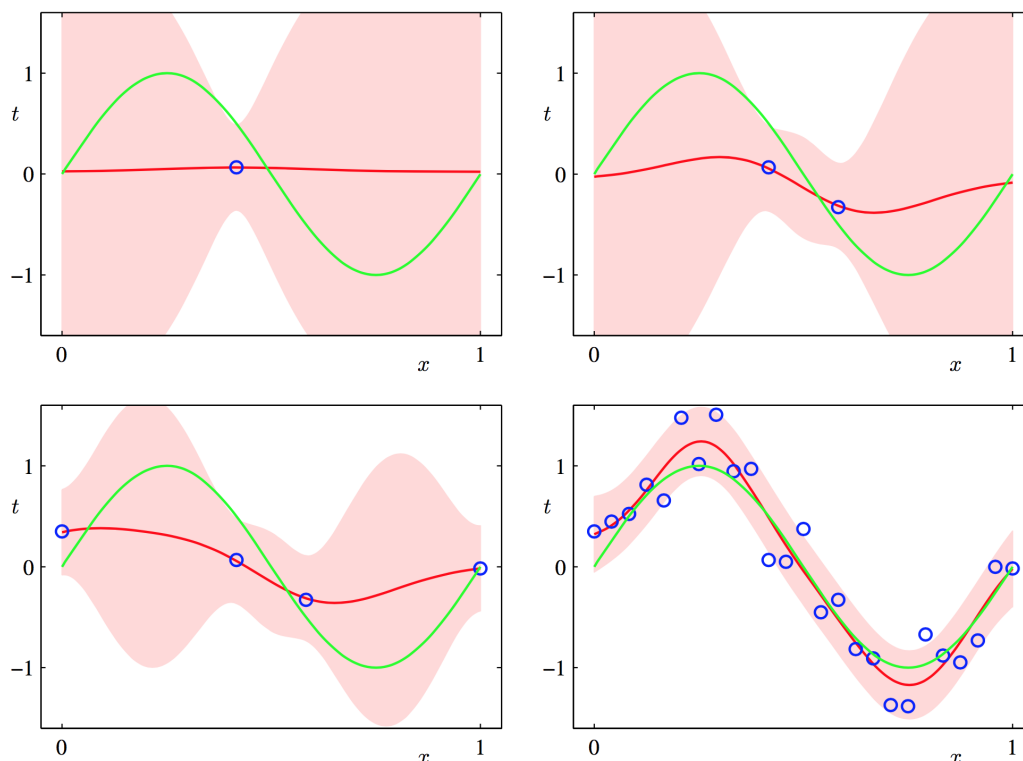
$$\begin{aligned}
 p(\mathbf{y}|\mathbf{x}, D) &= \int_{\mathbf{w}} p(\mathbf{y}, \mathbf{w}|\mathbf{x}, D) = \int_{\mathbf{w}} p(\mathbf{y}|\mathbf{x}, \mathbf{w})p(\mathbf{w}|D)d\mathbf{w} \\
 &= \int_{\mathbf{w}} \mathcal{N}(\mathbf{y}|\mathbf{w}^\top \mathbf{x}, \beta^{-1})\mathcal{N}(\mathbf{w}|\mathbf{m}_n, \mathbf{S}_n)d\mathbf{w}
 \end{aligned} \tag{7}$$

Interpretation:

- For a r.v. $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and transform $\mathbf{y} = \mathbf{A}\mathbf{z} + \mathbf{b}$, then \mathbf{y} is distributed $\mathbf{y} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$
- (7) draws \mathbf{w} from the posterior, and linearly transforms it with \mathbf{x}^\top (and adds some noise). Also: when we add two Normal r.v.s, the covariance of sum is some of covariance matrices.
- Predict the target value as follows:

$$p(\mathbf{y}|\mathbf{x}, D) = \mathcal{N}(\mathbf{y}|\mathbf{x}^\top \mathbf{m}_n, \mathbf{x}^\top \mathbf{S}_n \mathbf{x} + \beta^{-1})$$

33 / 38



(Bishop). Posterior predictive for a model with 9 Gaussian basis functions. Green = true model. 1, 2, 4 then 25 points. Red curve is mean of posterior predictive distribution. Red shaded region = ± 1 sd of mean.

34 / 38

- 1 Introduction
- 2 Beta/Bernoulli model
- 3 Normal-Normal Model
- 4 Bayesian Linear Regression
- 5 Bayesian Model Selection

35 / 38

Bayesian Model Selection (1 of 2)

- We have focused on using the Bayesian method to avoid over-fitting when learning parameters.
- Can also be used for model selection. The idea is to also introduce a prior on models, along with a prior on parameters for each model.
- This provides an alternative to using a validation set (or cross-validation) for model selection.

36 / 38

Bayesian Model Selection (2 of 2)

- Suppose we have a collection of models, $\{m_1, \dots, m_\ell\}$, and we want to use the data to form a posterior on models.

- True model M is a r.v., and has prior $p(m_k)$. We can evaluate

$$p(M = m_k | D) \propto \underbrace{p(D | M = m_k)}_{\text{model evidence}} \underbrace{p(M = m_k)}_{\text{model prior}}$$

- Second term expands as:

$$\begin{aligned} p(D | M = m_k) &= \int_{\boldsymbol{\theta}} p(D, \boldsymbol{\theta} | M = m_k) d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} \underbrace{p(D | \boldsymbol{\theta}, M = m_k)}_{\text{likelihood data}} \underbrace{p(\boldsymbol{\theta} | M = m_k)}_{\text{prior on parameters}} d\boldsymbol{\theta} \end{aligned}$$

- A complex model will tend to increase the first term, but decrease the second term. Also have a lower model prior.

37 / 38

Summary

- The Bayesian approach balances old data against new, accumulates information in the posterior.
- We think about the effect of data on a posterior on parameters.
- Given this posterior, we can extract a point estimate or compute the full posterior predictive.
- It is extremely helpful when the prior and likelihood functions form conjugate pairs, so that posterior in same form as prior.
- The MAP estimate in Bayesian LR reduces to MLE (and min-squared-error) when the prior on weights is uninformative, and to ridge regression when the prior on weights is zero mean and isotropic.

38 / 38