Machine Learning (CS 181):
5. Bayesian Methods and Linear Regression

David C. Parkes and Sasha Rush

Spring 2017

Contents

Contents

Overview: Regularization vs. Bayesian Methods

- A regularization penalty, such as ridge regression

$$\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) + \lambda \mathbf{w}^\top \mathbf{w}$$

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.

Overview: Regularization vs. Bayesian Methods

- A regularization penalty, such as ridge regression

$$\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) + \lambda \mathbf{w}^\top \mathbf{w}$$

  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.

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.

Review: Maximum Likelihood Estimation

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

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

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

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 $p(D) = \int_{\mathbf{w}} p(D|\mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w}$.

- 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.

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 $p(D) = \int_{\mathbf{w}} p(D|\mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w}$.

- 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.

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 $p(D) = \int_{\mathbf{w}} p(D|\mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w}$.
- 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.

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 $p(D) = \int_{\mathbf{w}} p(D|\mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w}$.
- 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.

Maximum A Posteriori Estimator

- In the MAP approach, we find

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

- 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 ($\ln p(\mathbf{w}) = const$) is equal to MLE.

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 ($\ln p(\mathbf{w}) = const$) is equal to MLE.

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 ($\ln p(\mathbf{w}) = const$) is equal to MLE.

Posterior predictive (Full Bayes)

- In the <u>posterior predictive</u> approach, we work with

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

- 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).

Posterior predictive (Full Bayes)

- In the posterior predictive approach, we work with

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

- 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).

The Prior as Data Processor

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

$$prior \to data^{(1)} \to posterior \to data^{(2)} \to posterior \to \cdots$$

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

Simple Example: Discrete Parameter θ

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

- The prior on θ is:

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

- Data are $D = Lime, Lime, Lime, Lime, \ldots$.
- θ_{MLE} is 1, 1, 1, after 1, 2, 3 $Lime$s respectively; e.g., after 1 $Lime$ we can check $p(D|\theta=1) = (1)^2 > p(D|\theta=0.75) = (0.75)^2$ (and similarly for other θs)
- θ_{MAP} is 0.5, 0.75, 1.0, after 1, 2 and 3 $Lime$s respectively; e.g., after 2 $Lime$s we can check $p(D|\theta=0.75)p(\theta=0.75) = (0.75)^2(0.2) > p(D|\theta=1)p(\theta=1) = (1)^2(0.1)$ (and similarly for other θs)

Simple Example: Discrete Parameter θ

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

- The prior on θ is:

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

- Data are $D = Lime, Lime, Lime, Lime, \ldots$.
- θ_{MLE} is 1, 1, 1, after 1, 2, 3 $Lime$s respectively; e.g., after 1 $Lime$ we can check $p(D|\theta=1) = (1)^2 > p(D|\theta=0.75) = (0.75)^2$ (and similarly for other θs)
- θ_{MAP} is 0.5, 0.75, 1.0, after 1, 2 and 3 $Lime$s respectively; e.g., after 2 $Lime$s we can check $p(D|\theta=0.75)p(\theta=0.75) = (0.75)^2(0.2) > p(D|\theta=1)p(\theta=1) = (1)^2(0.1)$ (and similarly for other θs)

Simple Example: Discrete Parameter θ

- 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 $\theta \in \{0, 0.25, 0.5, 0.75, 1\}$.
- The prior on θ is:

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

- Data are $D = Lime, Lime, Lime, Lime, \ldots$.
- θ_{MLE} is 1, 1, 1, after 1, 2, 3 $Lime$s respectively; e.g., after 1 $Lime$ we can check $p(D|\theta = 1) = (1)^2 > p(D|\theta = 0.75) = (0.75)^2$ (and similarly for other θs)
- θ_{MAP} is 0.5, 0.75, 1.0, after 1, 2 and 3 $Lime$s respectively; e.g., after 2 $Lime$s we can check $p(D|\theta = 0.75)p(\theta = 0.75) = (0.75)^2(0.2) > p(D|\theta = 1)p(\theta = 1) = (1)^2(0.1)$ (and similarly for other θs)

Simple Example: Discrete Parameter θ

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

- The prior on θ is:

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

- Data are $D = Lime, Lime, Lime, Lime, \ldots$.
- θ_{MLE} is 1, 1, 1, after 1, 2, 3 $Lime$s respectively; e.g., after 1 $Lime$ we can check $p(D|\theta = 1) = (1)^2 > p(D|\theta = 0.75) = (0.75)^2$ (and similarly for other θs)
- θ_{MAP} is 0.5, 0.75, 1.0, after 1, 2 and 3 $Lime$s respectively; e.g., after 2 $Lime$s we can check $p(D|\theta = 0.75)p(\theta = 0.75) = (0.75)^2(0.2) > p(D|\theta = 1)p(\theta = 1) = (1)^2(0.1)$ (and similarly for other θs)

Simple Example: Discrete Parameter θ

- 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 $\theta \in \{0, 0.25, 0.5, 0.75, 1\}$.
- The prior on θ is:

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

- Data are $D = Lime, Lime, Lime, Lime, \ldots$.
- θ_{MLE} is 1, 1, 1, after 1, 2, 3 $Lime$s respectively; e.g., after 1 $Lime$ we can check $p(D|\theta = 1) = (1)^2 > p(D|\theta = 0.75) = (0.75)^2$ (and similarly for other θs)
- θ_{MAP} is 0.5, 0.75, 1.0, after 1, 2 and 3 $Lime$s respectively; e.g., after 2 $Lime$s we can check $p(D|\theta = 0.75)p(\theta = 0.75) = (0.75)^2(0.2) > p(D|\theta = 1)p(\theta = 1) = (1)^2(0.1)$ (and similarly for other θs)

Contents

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:

$$\frac{\mathrm{d}}{\mathrm{d}\theta}[\cdot] = \frac{n_1}{\theta} - \frac{n_0}{1-\theta} = 0$$

$$\Leftrightarrow \quad \theta_{\mathrm{MLE}} = \frac{n_1}{n_0+n_1}.$$

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:

$$\frac{\mathrm{d}}{\mathrm{d}\theta}[\cdot] = \frac{n_1}{\theta} - \frac{n_0}{1-\theta} = 0$$

$$\Leftrightarrow \quad \theta_{\mathrm{MLE}} = \frac{n_1}{n_0 + n_1}.$$

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:

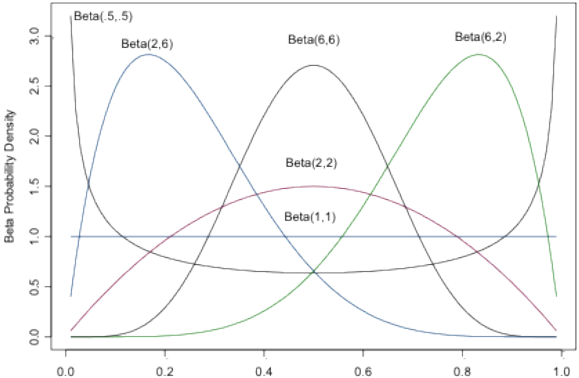
$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\theta}[\cdot] = \frac{n_1}{\theta} - \frac{n_0}{1-\theta} = 0 \\ \Leftrightarrow \quad \theta_{\mathrm{MLE}} = \frac{n_1}{n_0 + n_1}. \end{aligned}$$

Bernoulli: Bayesian approach

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

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

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



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

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:

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

  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.

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:

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

  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.

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:

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

  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.

Bernoulli-Beta: The MAP Estimate

- The mode of the Beta distribution is:

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

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

- Given posterior

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

we have

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

- For example, if data are $D = 1, 1, 0$, then $\theta_{\mathrm{MLE}} = n_1/n = 2/3$. Given prior $\mathrm{Beta}(2, 4)$, we have $\theta_{\mathrm{MAP}} = \frac{2+2-1}{2+4+3-2} = \frac{3}{7}$.

Bernoulli-Beta: The MAP Estimate

- The mode of the Beta distribution is:

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

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

- Given posterior

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

  we have

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

- For example, if data are $D = 1, 1, 0$, then $\theta_{\mathrm{MLE}} = n_1/n = 2/3$. Given prior $\mathrm{Beta}(2, 4)$, we have $\theta_{\mathrm{MAP}} = \frac{2+2-1}{2+4+3-2} = \frac{3}{7}$.

Bernoulli-Beta: The MAP Estimate

- The mode of the Beta distribution is:

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

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

- Given posterior

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

  we have

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

- For example, if data are $D = 1, 1, 0$, then $\theta_{\mathrm{MLE}} = n_1/n = 2/3$. Given prior $\mathrm{Beta}(2,4)$, we have $\theta_{\mathrm{MAP}} = \frac{2+2-1}{2+4+3-2} = \frac{3}{7}$.

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:

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

  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.

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:

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

  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.

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:

$$prior \to data^{(1)} \to posterior \to data^{(2)} \to posterior \to \cdots$$

  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.

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:

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

  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.

Bernoulli-Beta: Full Bayes

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

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

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

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

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

Bernoulli-Beta: Full Bayes

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

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

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

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

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

Bernoulli-Beta: Full Bayes

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

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

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

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

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

Contents

Warm-up: Univariate Normal

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

- Generative model:

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

given parameters μ, σ^2.

- Maximum likelihood estimation (known σ^2):

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

Warm-up: Univariate Normal

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

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

given parameters μ, σ^2.

- Maximum likelihood estimation (known σ^2):

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

MAP Estimator for Univariate Normal

- Model $\mathcal{N}(x \mid \mu, \sigma^2)$. 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:

$$m_n = \frac{\sigma^2}{ns_0^2 + \sigma^2} m_0 + \frac{ns_0^2}{ns_0^2 + \sigma^2} \mu_{\mathrm{MLE}} \tag{1}$$

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

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

MAP Estimator for Univariate Normal

- Model $\mathcal{N}(x \mid \mu, \sigma^2)$. 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:

$$m_n = \frac{\sigma^2}{ns_0^2 + \sigma^2} m_0 + \frac{ns_0^2}{ns_0^2 + \sigma^2} \mu_{\mathrm{MLE}} \tag{1}$$

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

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

MAP Estimator for Univariate Normal

- Model $\mathcal{N}(x \mid \mu, \sigma^2)$. 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:

$$m_n = \frac{\sigma^2}{ns_0^2 + \sigma^2} m_0 + \frac{ns_0^2}{ns_0^2 + \sigma^2} \mu_{\mathrm{MLE}} \tag{1}$$

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

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

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:

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

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:

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

Figuring out the Posterior (2 of 2)

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

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

Complete the square, moving additional terms into the constant

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

where

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

Figuring out the Posterior (2 of 2)

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

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

Complete the square, moving additional terms into the constant

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

where

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

Extension: Multivariate Normal

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

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

  with known $\boldsymbol{\Sigma}$.

- Prior $\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{m}_0, \boldsymbol{S}_0)$, for parameters $\boldsymbol{m}_0$ and $\boldsymbol{S}_0$.
- Posterior after n examples is $\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{m}_n, \boldsymbol{S}_n)$, and:

$$\boldsymbol{S}_n = \left(\boldsymbol{S}_0^{-1} + n\boldsymbol{\Sigma}^{-1}\right)^{-1} \tag{3}$$

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

Extension: Multivariate Normal

- $D = \{\mathbf{x}_i\}_{i=1}^n$, with $\mathbf{x}_i \in \mathbb{R}^m$
- Generative model:
$$\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right),$$
with known $\boldsymbol{\Sigma}$.
- Prior $\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{m}_0, \boldsymbol{S}_0)$, for parameters $\boldsymbol{m}_0$ and $\boldsymbol{S}_0$.
- Posterior after n examples is $\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{m}_n, \boldsymbol{S}_n)$, and:
$$\boldsymbol{S}_n = \left(\boldsymbol{S}_0^{-1} + n\boldsymbol{\Sigma}^{-1}\right)^{-1} \tag{3}$$
$$\boldsymbol{m}_n = \boldsymbol{S}_n \left(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + n\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\mathrm{MLE}}\right) \tag{4}$$

Interpretation of MAP estimator

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

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

- Prior is overwhelmed as n gets bigger, with Σ^{-1} and $\Sigma^{-1}\boldsymbol{\mu}_{\mathrm{MLE}}$ having a larger effect on $\boldsymbol{S}_n$ and $\boldsymbol{m}_n$, respectively.

- With a strong prior, then $\boldsymbol{S}_0$ has small positive numbers on diagonal and inverse would have large numbers on diagonal, would "compete" with n to center the posterior mean at $\boldsymbol{m}_0$ instead of μ_{MLE}.

- Posterior cov. depends on data only through amount of data n. Wouldn't be case if Σ was also unknown. (See Bishop 2.3.6).

Interpretation of MAP estimator

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

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

- Prior is overwhelmed as n gets bigger, with $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\mathrm{MLE}}$ having a larger effect on $\boldsymbol{S}_n$ and $\boldsymbol{m}_n$, respectively.

- With a strong prior, then 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 $\boldsymbol{m}_0$ instead of μ_{MLE}.

- Posterior cov. depends on data only through amount of data n. Wouldn't be case if Σ was also unknown. (See Bishop 2.3.6).

Interpretation of MAP estimator

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

$$
\begin{aligned}
\boldsymbol{S}_n &= \left(\boldsymbol{S}_0^{-1} + n\boldsymbol{\Sigma}^{-1}\right)^{-1} \\
\boldsymbol{m}_n &= \boldsymbol{S}_n \left(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + n\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\text{MLE}}\right)
\end{aligned}
$$

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

Interpretation of MAP estimator

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

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

- Prior is overwhelmed as n gets bigger, with $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\mathrm{MLE}}$ having a larger effect on $\boldsymbol{S}_n$ and $\boldsymbol{m}_n$, respectively.
- With a strong prior, then $\boldsymbol{S}_0$ has small positive numbers on diagonal and inverse would have large numbers on diagonal, would "compete" with n to center the posterior mean at $\boldsymbol{m}_0$ instead of μ_{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).

Figuring out the Posterior (v2)

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

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

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

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

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

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

Complete the square, moving additional terms into the constant

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

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

Contents

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.

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

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

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

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

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.

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

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

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

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

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.

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

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

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

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

Interpretation of Bayesian LR MAP Estimator

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

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

- The MAP estimate is $\theta_{\mathrm{MAP}} = \boldsymbol{m}_n$.

- With a weak prior, then S_0 has large entries on the diagonal, and S_0^{-1} is close to zero, and we have

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

In addition, we have

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

Interpretation of Bayesian LR MAP Estimator

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

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

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

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

  In addition, we have

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

Special case: Simple Prior on Weights

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

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

- We see that

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

  and we recover ridge regression!

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

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

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

Special case: Simple Prior on Weights

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

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

- We see that

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

and we recover ridge regression!

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

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

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

Special case: Simple Prior on Weights

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

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

- We see that

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

  and we recover ridge regression!

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

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

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

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 w into the constant, $\ln p(\mathbf{w}|D) =$

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

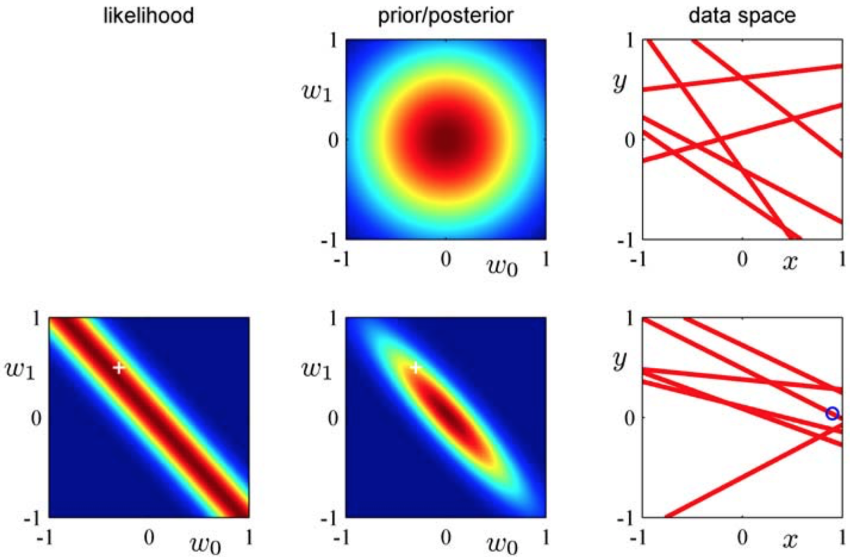
Collecting the quadratic and linear terms:

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

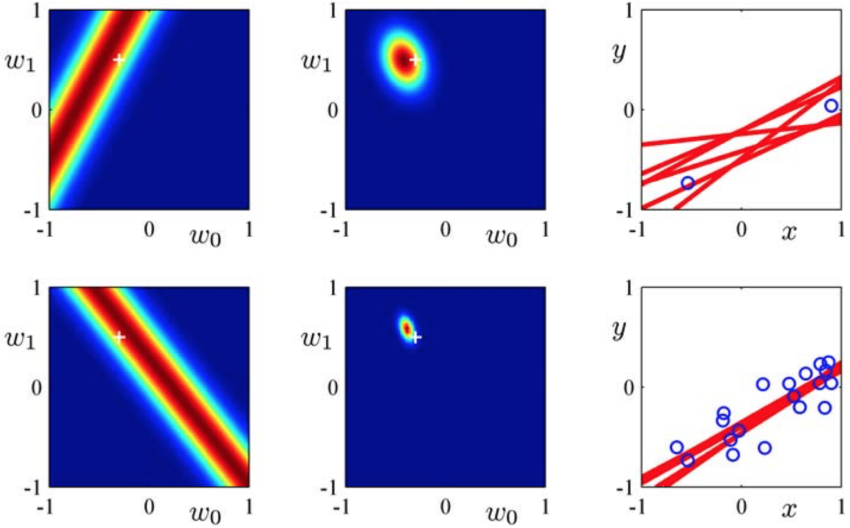
Completing the square, we have:

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

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



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



(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.

Posterior Predictive Bayesian LR

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

Interpretation:

- For a r.v. $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and transform $\mathbf{q} = \mathbf{A}\mathbf{z} + \mathbf{b}$, then $\mathbf{q}$ is distributed $\mathbf{q} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$
- (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:

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

Posterior Predictive Bayesian LR

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

Interpretation:

- For a r.v. $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and transform $\mathbf{q} = \mathbf{A}\mathbf{z} + \mathbf{b}$, then $\mathbf{q}$ is distributed $\mathbf{q} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$

- (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:

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

Posterior Predictive Bayesian LR

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

Interpretation:

- For a r.v. $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and transform $\mathbf{q} = \mathbf{A}\mathbf{z} + \mathbf{b}$, then $\mathbf{q}$ is distributed $\mathbf{q} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$
- (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:

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

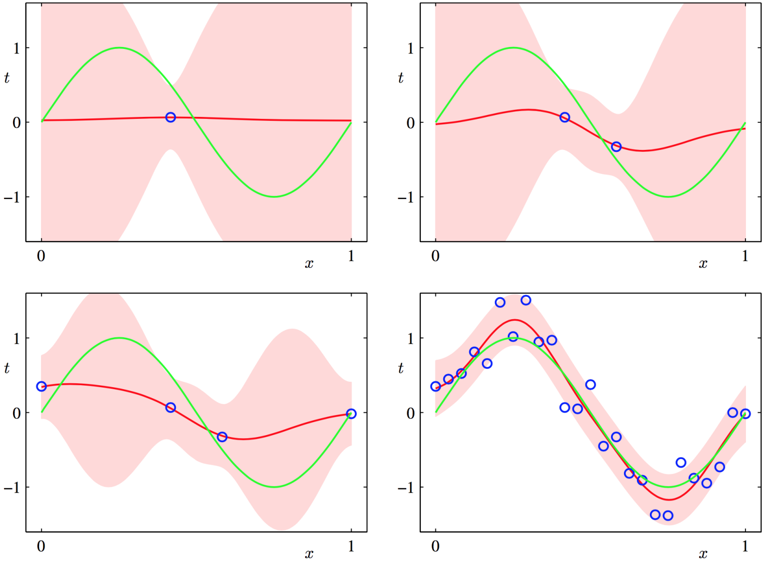
Posterior Predictive Bayesian LR

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

Interpretation:

- For a r.v. $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and transform $\mathbf{q} = \mathbf{A}\mathbf{z} + \mathbf{b}$, then $\mathbf{q}$ is distributed $\mathbf{q} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$
- (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:

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



(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 $= \pm 1$ sd of mean.

Contents

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.

Bayesian Model Selection (2 of 2)

- Suppose we have a collection of models, $\{m_1, \ldots, m_\ell\}$, 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:

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

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

Bayesian Model Selection (2 of 2)

- Suppose we have a collection of models, $\{m_1, \ldots, m_\ell\}$, 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:

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

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

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:

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

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

Bayesian Model Selection (2 of 2)

- Suppose we have a collection of models, $\{m_1, \ldots, m_\ell\}$, 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:

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

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

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.