

ST451 - Lent term

Bayesian Machine Learning

Kostas Kalogeropoulos

Bayesian Inference Concepts - Linear Regression

Summary of last lecture

- **Linear regression:** $y \sim N(X\beta, \sigma^2)$.
- **MLE - Least Squares** estimator: $\hat{\beta} = (X^T X)^{-1} X^T y$. Minimises

$$\sum_{i=1}^n (y_i - \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})^2,$$

- **Ridge regression estimator** $\hat{\beta}^r = (X^T X + \lambda^2 I)^{-1} X^T y$ corrects for overfit by minimising

$$\sum_{i=1}^n (y_i - \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})^2 + \lambda \sum_{i=1}^M \beta_i^2.$$

- Same as Bayesian estimator obtained by assigning $N(0, \sigma^2 g I)$ as prior for β - **Bayesian linear regression**

Outline

- 1 Bayesian Linear Regression - unknown σ^2
- 2 Bayesian Model Selection
- 3 Model/Prior Selection based on cross-validation

Outline

- 1 Bayesian Linear Regression - unknown σ^2
- 2 Bayesian Model Selection
- 3 Model/Prior Selection based on cross-validation

Normal with unknown mean vs Linear Regression

Null model: Assume $y = (y_1, \dots, y_n)$ (y_i independent) from $N(\theta, \sigma^2)$ with σ^2 known. Assign $N(\mu_0, \sigma^2 \tau_0^2)$ as **prior** for θ .

The **posterior** $\pi(\theta|y, \sigma^2)$ is then $N(\mu_n, \sigma^2 \tau_n^2)$ where

$$\mu_n = \frac{\frac{1}{n}\mu_0 + \tau^2 \bar{y}}{\tau^2 + \frac{1}{n}}, \quad \tau_n^2 = \frac{1}{\frac{1}{\tau^2} + n}$$

Linear Regression model: Assume y is $N(X\beta, \sigma^2 I_n)$. Assign $N(\mu_0, \sigma^2 \Omega_0)$ as **prior** for β .

The **posterior** $\pi(\beta|y, X, \sigma^2)$ is then $N(\mu_n, \sigma^2 \Omega_n^2)$ where

$$\mu_n = (X^T X + \Omega_0^{-1})^{-1}(\Omega_0^{-1} \mu_0 + X^T y), \quad \Omega_n = (X^T X + \Omega_0^{-1})^{-1}$$

What if σ^2 is **unknown**?

Bayesian inference for multiparameter models

We may have **more than one** parameters, say $\theta = (\theta_1, \theta_2)$. As before assign a prior $\pi(\theta_1, \theta_2)$ and obtain the posterior

$$\pi(\theta_1, \theta_2 | y) = \frac{f(y | \theta_1, \theta_2) \pi(\theta_1, \theta_2)}{\int \int f(y | \theta_1, \theta_2) \pi(\theta_1, \theta_2) d\theta_1 d\theta_2}$$

If interest mainly lies in θ_1 , the **marginal posterior** of θ_1 may be used by averaging over θ_2

$$\pi(\theta_1 | y) = \int \pi(\theta_1, \theta_2 | y) d\theta_2$$

Example 1: Normal with both μ and σ^2 unknown

Let $y = (y_1, \dots, y_n)$ independent rv's from the $N(\theta, \sigma^2)$

Likelihood: The **likelihood** is (s_y^2 denotes the sample variance)

$$f(y|\theta, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\left(-\frac{(n-1)s_y^2 + n(\bar{y} - \theta)^2}{2\sigma^2}\right)$$

Prior: Assume an **improper** prior $\pi(\theta, \sigma^2) \propto (\sigma^2)^{-1}$

Example 1: Posterior

Posterior: factorised as $\pi(\theta, \sigma^2|y) = \pi(\sigma^2|y)\pi(\theta|y, \sigma^2)$

$$\begin{aligned}\pi(\theta, \sigma^2|y) &\propto (\sigma^2)^{-n/2-1} \exp\left(-\frac{(n-1)s_y^2 + n(\bar{y} - \theta)^2}{2\sigma^2}\right) \\&= (\sigma^2)^{-n/2-1} \exp\left(-\frac{(n-1)s_y^2}{2\sigma^2}\right) \exp\left(-\frac{(\bar{y} - \theta)^2}{2\frac{\sigma^2}{n}}\right) \\&\propto (\sigma^2)^{-n/2-1} \exp\left(-\frac{(n-1)s_y^2}{2\sigma^2}\right) \frac{(\frac{\sigma^2}{n})^{1/2}}{(\frac{\sigma^2}{n})^{1/2}} \exp\left(-\frac{(\theta - \bar{y})^2}{2\frac{\sigma^2}{n}}\right) \\&= (\sigma^2)^{-(n-1)/2-1} \exp\left(-\frac{\frac{(n-1)s_y^2}{2}}{\sigma^2}\right) \text{N}\left(\bar{y}, \frac{\sigma^2}{n}\right) \\&= \text{IGamma}\left(\frac{n-1}{2}, \frac{(n-1)s_y^2}{2}\right) \times \text{N}\left(\bar{y}, \frac{\sigma^2}{n}\right)\end{aligned}$$

Note: the above IGamma(), N() refer to the corresponding pdfs

Example 1: Proper priors

Aiming for a **proper** prior we can set as in the previous slide

$$\pi(\theta, \sigma^2) = \pi(\sigma^2)\pi(\theta|\sigma^2) = \text{IGamma}(\alpha_0, \beta_0) \times \text{N}(\mu_0, \sigma^2\tau_0^2)$$

Similar calculations the yield that $\pi(\theta, \sigma^2|y)$ can be **factorised** as

$$\pi(\theta, \sigma^2|y) = \pi(\sigma^2|y)\pi(\theta|\sigma^2, y) = \text{IGamma}(\alpha_n, \beta_n) \times \text{N}(\mu_n, \sigma^2\tau_n^2)$$

where

$$\mu_n = \frac{\frac{1}{n}\mu_0 + \tau_0^2\bar{y}}{\tau_0^2 + \frac{1}{n}}, \quad \tau_n^2 = \frac{1}{\frac{1}{\tau_0^2} + n},$$

$$\alpha_n = \alpha_0 + n/2, \quad \beta_n = \beta_0 + \frac{n-1}{2}\mathbf{s}_y^2 + \frac{\frac{1}{\tau_0^2}n(\bar{y}-\mu_0)^2}{2(1/\tau_0^2+n)}$$

Bayesian Linear Regression model with unknown σ^2

Assume y is $N(X\beta, \sigma^2 I_n)$. Assign $N(\mu_0, \sigma^2 \Omega_0)$ as **prior** for β (given σ^2) and the $\text{IGamma}(\alpha_0, \beta_0)$ for σ^2 .

The **posterior** for $\pi(\beta, \sigma^2 | y, X)$ is then the product of the $\text{IGamma}(\alpha_n, \beta_n)$ and the $N(\mu_n, \sigma^2 \Omega_n^2)$ where

$$\mu_n = (X^T X + \Omega_0^{-1})^{-1} (\Omega_0^{-1} \mu_0 + X^T y)$$

$$\Omega_n = (X^T X + \Omega_0^{-1})^{-1},$$

$$\alpha_n = \alpha_0 + \frac{n}{2},$$

$$\beta_n = \beta_0 + \frac{1}{2} (y^T y + \mu_0^T \Omega_0^{-1} \mu_0 + \mu_n^T \Omega_n^{-1} \mu_n).$$

Example 1: Marginal posterior of μ

For the **improper** prior case, The joint posterior is

$$\pi(\theta, \sigma^2 | y) \propto (\sigma^2)^{-n/2-1} \exp \left(-\frac{(n-1)s_y^2 + n(\bar{y} - \theta)^2}{2\sigma^2} \right)$$

Integrating σ^2 out yield the **marginal** posterior of θ to be the **t distribution** with $n - 1$ degrees of freedom, location \bar{y} and scale s_y/\sqrt{n}

Bayesian Linear Regression model with unknown σ^2

Assume y is $N(X\beta, \sigma^2 I_n)$. Assign $N(\mu_0, \sigma^2 \Omega_0)$ as **prior** for β (given σ^2) and the $\text{IGamma}(\alpha_0, \beta_0)$.

The **posterior** for $\pi(\beta, \sigma^2 | y, X)$ is then the product of the $\text{IGamma}(\alpha_n, \beta_n)$ and the $N(\mu_n, \sigma^2 \Omega_n^2)$ where

$$\mu_n = (X^T X + \Omega_0^{-1})^{-1} (\Omega_0^{-1} \mu_0 + X^T y)$$

$$\Omega_n = (X^T X + \Omega_0^{-1})^{-1},$$

$$\alpha_n = \alpha_0 + \frac{n}{2},$$

$$\beta_n = \beta_0 + \frac{1}{2} (y^T y + \mu_0^T \Omega_0^{-1} \mu_0 + \mu_n^T \Omega_n^{-1} \mu_n).$$

The marginal posterior for β , $\pi(\beta | y, X)$ is the **multivariate t** distribution with $2\alpha_n$ degrees of freedom, location μ_n and scale $\frac{\beta_n}{\alpha_n} \Omega_n$

Marginal Posterior of β and Predictive distribution

To obtain credible intervals for β we could use the t distribution. But we would use Monte Carlo instead as this will cover more general models. e.g. logistic regression.

So we will sample N Monte Carlo samples from $\pi(\beta|y)$ and use them for Monte Carlo inference (credible intervals, density plots etc)

Monte Carlo Samples can be drawn by

- 1 Generating samples σ_i^2 from $\text{IGamma}(\alpha_n, \beta_n)$, $i = 1, \dots, N$,
- 2 Draw β_i sample based on each σ_i^2 from $N(\mu_n, \sigma^2 \Omega_n^2)$,

For the predictive distribution for a new observation y_* based on covariates X_* we can use the additional step of drawing y_{*i} from $N(X_*\beta_i, \sigma_i^2)$ for each β_i, σ_i^2 .

Outline

- 1 Bayesian Linear Regression - unknown σ^2
- 2 Bayesian Model Selection**
- 3 Model/Prior Selection based on cross-validation

Example: Automobile Bodily Injuries Claim data

The data are automobile **injury claims** from the Insurance Research Council (IRC) collected in 2002.

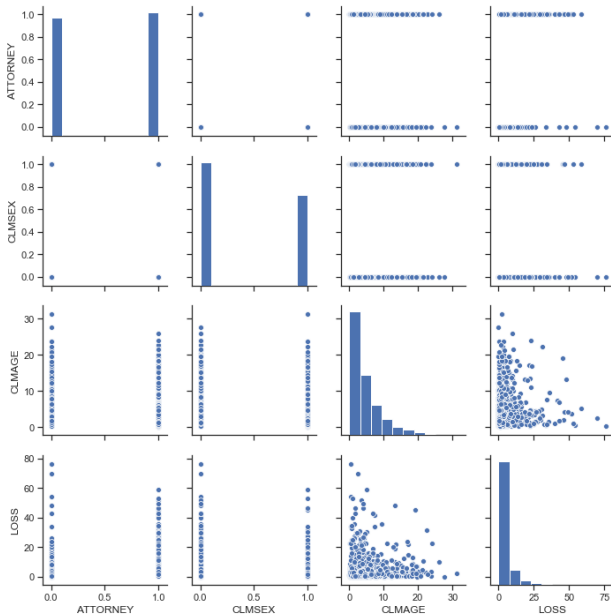
They contain information on the **gender** of the claimant, attorney involvement, years of driving experience and the economic loss (in thousands \$).

Is the information on the claimant's gender an important predictor? As of now insurers are **not allowed** to use gender information.

The **hypothesis testing / model selection** problem can be formulated as the Bayes factor between the model including all the variables and the model without gender. Or else if we fit the model with all variables

$$H_0 : \beta_{gender} = 0 \quad \text{vs} \quad H_1 : \beta_{gender} \neq 0$$

Example: Automobile Bodily Injuries Claim data



Bayesian Hypothesis Testing / Model Choice

Hypotheses: Let $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$.

Define the **Bayes factor** in favour of H_1 as $B_{10}(x) = f(x|H_1)/f(x|H_0)$.
The **Bayes** test rule is **Choose H_1 if $B_{10}(x) > 1$** .

Notes

- No control of type I error probability.
- Simple outcome, reference to both hypotheses, choice of H_0 vs H_1 doesn't matter, easily extended to more hypotheses.

The Bayes factor can be computed from either of the expressions below

$$B_{10}(x) = \frac{\int_{\Theta_1} f(x|\theta, H_1)\pi(\theta|H_1)d\theta}{\int_{\Theta_0} f(x|\theta, H_0)\pi(\theta|H_0)d\theta} = \frac{P(H_1|x)/P(H_0|x)}{P(H_1)/P(H_0)}$$

Bayes factor - interpretation

In terms of interpretation the following guidelines are available

| | |
|---------------------------|--|
| $1 < B_{10}(x) \leq 3$ | evidence against H_0 is poor |
| $3 < B_{10}(x) \leq 20$ | evidence against H_0 is substantial |
| $20 < B_{10}(x) \leq 150$ | evidence against H_0 is strong |
| $B_{10}(x) > 150$ | evidence against H_0 is decisive |

Bayesian Occam's razor

Bayesian Occam's razor: Models with more parameters (more complex models) will not necessarily have higher marginal likelihood.

Conservation of probability mass: More complex models will handle more complex datasets adequately. But the probabilities over all these datasets will have to sum to one.

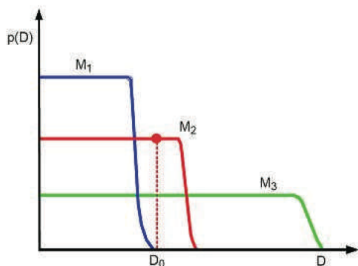


Figure 5.6 A schematic illustration of the Bayesian Occam's razor. The broad (green) curve corresponds to a complex model, the narrow (blue) curve to a simple model, and the middle (red) curve is just right. Based on Figure 3.13 of (Bishop 2006a). See also (Murray and Ghahramani 2005, Figure 2) for a similar plot produced on real data.

Jeffreys-Lindley-(Bartlett) paradox - example 1

Real data example: A person claimed to possess extrasensory capacities (ESP) and can alter the outcome of a machine that output 0, 1 with probability $\theta = 0.5$ (H_0). H_1 is $\theta \neq 0.5$.

In 104.490.000 trials, there were 52.263.471 ones.



Jeffreys-Lindley-(Bartlett) paradox - example 1

Frequentist inference gives a very small **p-value**, much less than 0.01, pointing towards H_1 .

Bayes factor, based on a Uniform(0, 1) prior on θ supports H_0 .

Maybe not a paradox. Frequentist testing ask the question is $\theta = 0.5$?

Bayesian testing compares a model with $\theta = 0.5$ and a model with θ drawn uniformly from (0, 1) as to how they explain the data.

Jeffreys-Lindley-(Bartlett) paradox - example 2

Let $y = (y_1, \dots, y_n)$ iid from the $N(\theta, 1)$ distribution, and consider testing $H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$.

The Bayes factor is

$$B_{01} = \frac{\exp\{-n(\bar{y}_n)^2/2\}}{\int_{-\infty}^{+\infty} \exp\{-n(\bar{y}_n - \theta)^2/2\} \pi(\theta) d\theta}$$

Assume the improper **Jeffreys prior** $p(\theta) = c$. Then

$$B_{01} = \frac{\exp\{-n(\bar{y}_n)^2/2\}}{c \int_{-\infty}^{+\infty} \exp\{-n(\bar{y}_n - \theta)^2/2\} d\theta} = \frac{\exp\{-n(\bar{y}_n)^2/2\}}{c\sqrt{2\pi/n}}$$

The decision depends on the arbitrary constant c ! Should use proper priors.

Jeffreys-Lindley-(Bartlett) paradox - example 2 (cont'd)

Consider the **low informative** prior $N(0, \tau^2)$ with some big τ^2 . The Bayes factor is

$$B_{01} = \frac{\exp\{-n(\bar{y}_n)^2/2\}}{\int_{-\infty}^{+\infty} \exp\{-n(\bar{y}_n - \theta)^2/2\} (2\pi\tau^2)^{-1/2} \exp(-\theta^2/2\tau^2) d\theta}$$

As $\tau \rightarrow \infty$, $B_{01} \rightarrow \infty$ regardless of \bar{y}_n (except if $\bar{y}_n = 0$). So we for a near-infinite value of τ^2 we will always choose H_0 .

It is therefore clear that more thought should be put on the choice of $\pi(\theta)$ when it come to testing.

If we don't have information we still need to put **some** information but not **too much**.

Unit information priors

In the previous example the **unit information** prior is the $N(\mu_0, 1)$, i.e. putting the same prior variance as the variance of each data point.

The posterior is $N(\mu_n, \tau_n^2)$ with

$$\mu_n = \frac{1}{n+1}(\mu_0 + n\bar{y}), \quad \tau_n^2 = \frac{1}{n+1}$$

This prior is like adding **one** more observation equal to μ_0 . In fact σ^2 corresponds to Fisher information from one data point.

Cheat (add information), but as **little** as possible.

Unit information prior for Linear Regression

In the linear regression case remember from the derivation of MLE that

$$\frac{\partial}{\partial \beta} \log f(y|X, \hat{\beta}, \sigma^2) = \frac{1}{\sigma^2} \left(\sum_{i=1}^n X_i^T Y_i - \sum_{i=1}^n X_i^T X_i \hat{\beta} \right),$$

Hence the Fisher information for β based on n observations is

$$I(\beta) = -E \left[-\frac{1}{\sigma^2} X^T X \right] = \frac{X^T X}{\sigma^2}$$

Unit information takes the average over n observations so the variance is set to $n\sigma^2(X^T X)^{-1}$.

This implies setting $\Omega_0 = n(X^T X)^{-1}$, so $(X^T X)^{-1}$ instead of I_p and with $g = n$.

Bayesian Model Selection

To compare models we will need to compute the **marginal likelihood** or evidence **for each model**.

We can then use the model with the **highest marginal likelihood**. The use of unit information prior is the default option to guard against the Jeffreys-Lindley paradox.

Computing the marginal likelihood is generally a very difficult task but here we can use the following trick. We can write

$$\begin{aligned}\pi(\beta, \sigma^2 | y, X) &= \frac{\pi(y | \beta, \sigma^2, X) \pi(\beta, \sigma^2)}{\pi(y | X)}, \text{ or else} \\ \pi(y | X) &= \frac{\pi(y | \beta, \sigma^2, X) \pi(\beta, \sigma^2)}{\pi(\beta, \sigma^2 | y, X)}, \text{ for all } \beta, \sigma^2.\end{aligned}$$

The expression above contains known Normal and Inverse Gamma pdfs so we can just evaluate for -say- the posterior mean of β, σ^2 .

Outline

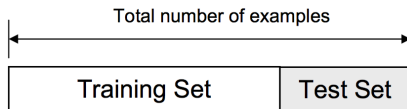
- 1 Bayesian Linear Regression - unknown σ^2
- 2 Bayesian Model Selection
- 3 Model/Prior Selection based on cross-validation

Training error versus Test error

- In machine learning the choice between models or priors (tuning parameters) is made on the basis of their **out of sample performance**.
- Models are estimated in part(s) of the data, the **training set**. The **training error** can be found by checking predictions of each model on the the training set data.
- The **test** set consists of data not used in the training set. The **test error** is the obtained by checking the predictions of each model, estimated (trained) on the training data.
- Training error often is **quite different** from the test error; the former can dramatically underestimate the latter.

Hold out method

The method described in the previous slide is also known as **holdout** method.



The holdout method has two basic drawbacks:

- In problems where we have a sparse dataset we may not be able to afford the **luxury** of setting aside a portion of the dataset for testing
- Since it is a single train-and-test experiment, the holdout estimate of error rate will be misleading if we happen to get an **unfortunate** split

Cross Validation

The drawback above can be overcome by using resampling methods at the expense of a **higher** computational cost.

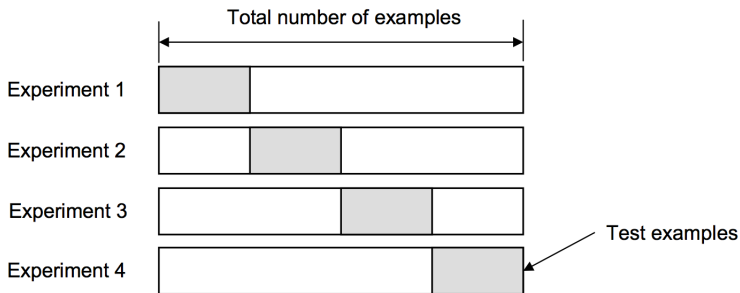
One of them is **cross-validation**. The following are the most frequently used cross-validation techniques:

- Random Subsampling
- K-Fold Cross-Validation
- Leave-one-out Cross-Validation

In all of the above schemes **several** train-test splits are constructed and the test error is **averaged** over them.

K-Fold Cross Validation

- For each of K experiments, use $K - 1$ folds for training and a **different fold** for testing.
- K-Fold Cross validation is **similar** to Random Subsampling. The advantage is that **all** the examples in the dataset are eventually used for both training and testing



How many folds are needed? (cont'd)

In **practice**, the choice of the number of folds depends on the **size** of the dataset

- For **large datasets**, even 3-Fold Cross Validation will be quite accurate
- For very **sparse datasets**, we may have to use leave-one-out in order to train on as many examples as possible

Test error and model selection

- If model selection and test error estimates are to be computed simultaneously, the data needs to be divided into **three disjoint sets** [Ripley, 1996]:
 - 1 **Training set:** to train different versions of each model.
 - 2 **Validation set:** to tune the parameters of each model and compare.
 - 3 **Test set:** For estimating the test error of the best model.
- Test and validation sets are separated because the error rate estimate of the final model on validation data will be **biased** downwards.
- After assessing the final model on the test set, **you must not** tune the model any further.

Today's lecture - Reading

Murphy: 5.3 and 7.6

Bishop: 2.3.6, 3.3.1, 3.3.2, 3.4, 3.5.1 and 3.5.2