Bayesian Linear Regression

Advanced Machine Learning 2019
Tutorial 2

Stefan Stark

ETH Zurich

October 21, 2019

Table of Contents

Bayes vs Frequentist

Estimators

Bayesian Linear Regression

Intro to GPs

Bayesian Approach

Parameters: θ
Data: D

Prior: $\mathrm{p}(\theta)$
Evidence: $\mathrm{p}(D)$

Likelihood: $\mathrm{p}(D|\theta)$
Posterior: $\mathrm{p}(\theta|D)$

Treat θ as a random variable and compute the posterior

$$\mathrm{p}(\theta|D) = \frac{\mathrm{p}(D|\theta)\,\mathrm{p}(\theta)}{\mathrm{p}(D)} = \frac{\mathrm{p}(D|\theta)\,\mathrm{p}(\theta)}{\int \mathrm{p}(D|\theta)\,\mathrm{p}(\theta)d\theta}$$

Standard Bayesian workflow

$$\mathrm{p}(\theta|D) = \frac{\mathrm{p}(D|\theta)\,\mathrm{p}(\theta)}{\int \mathrm{p}(D|\theta)\,\mathrm{p}(\theta)d\theta}$$

- Design $\mathrm{p}(D|\theta)$ for your problem
- Design $\mathrm{p}(\theta)$ based on your prior notions of the problem
- Compute $\mathrm{p}(\theta|D)$
- Realize $\int \mathrm{p}(D|\theta)\,\mathrm{p}(\theta)d\theta$ is super hard to compute
  - Choose simple $\mathrm{p}(D|\theta)$ and $\mathrm{p}(\theta)$ to give an analytic solution
    - e.g. $\mathrm{p}(\theta)$ congjugate to $\mathrm{p}(D|\theta)$
  - Or use expensive approximation methods
- Analyze $\mathrm{p}(\theta|D)$

Frequentist Approach

- There is one true θ^*
- Datasets are random samples: $D \sim \mathrm{p}(\cdot|\theta^*)$
- Estimate θ^* with $\hat{\theta}$ by proposing and applying an estimator δ
  - $\hat{\theta} = \delta(D)$; e.g. $\hat{\mu} = \delta_\mu(X) = \frac{1}{N}\sum x_i$
- Sampling distribution: distribution induced on $\hat{\theta}$ by applying δ to different datasets
  - estimated by e.g. bootstraping D

There is not an automatic δ that falls out of this approach. Sometimes you would prefer one estimator over another. What are some desirable properties of estimators?

Table of Contents

Bayes vs Frequentist

Estimators

Bayesian Linear Regression

Intro to GPs

Estimator Properties

Consistent Estimators

- $\hat{\theta}(D) \rightarrow \theta^*$ as $|D| \rightarrow \infty$

Unbiased Estimators

- $bias(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta^*$
- An unbiased estimator has bias $= 0$

In words, if the bias is zero then the sampling distribution is centered on the true value.

Consider estimating the mean of a Gaussian distribution from samples

$$D = \{x_1, \ldots, x_N\}$$

δ_1 estimates by the first datapoint x_1
δ_N estimates by $\frac{1}{N}\sum_{i=1}^{N} x_i$

What are the biases of these estimators?
Why do you prefer δ_N? (You should prefer δ_N)

Estimator Properties II

Variance of an estimator

$$var(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$$

Cramer-Rao lower bound: the minimum variance of an unbiased estimator

$$var(\hat{\theta}) \geq \mathcal{I}(\theta_*)^{-1}$$

- $\mathcal{I}$ is the fisher information
- (See the lecture for details)

Efficiency: how close is the variance to the lower bound

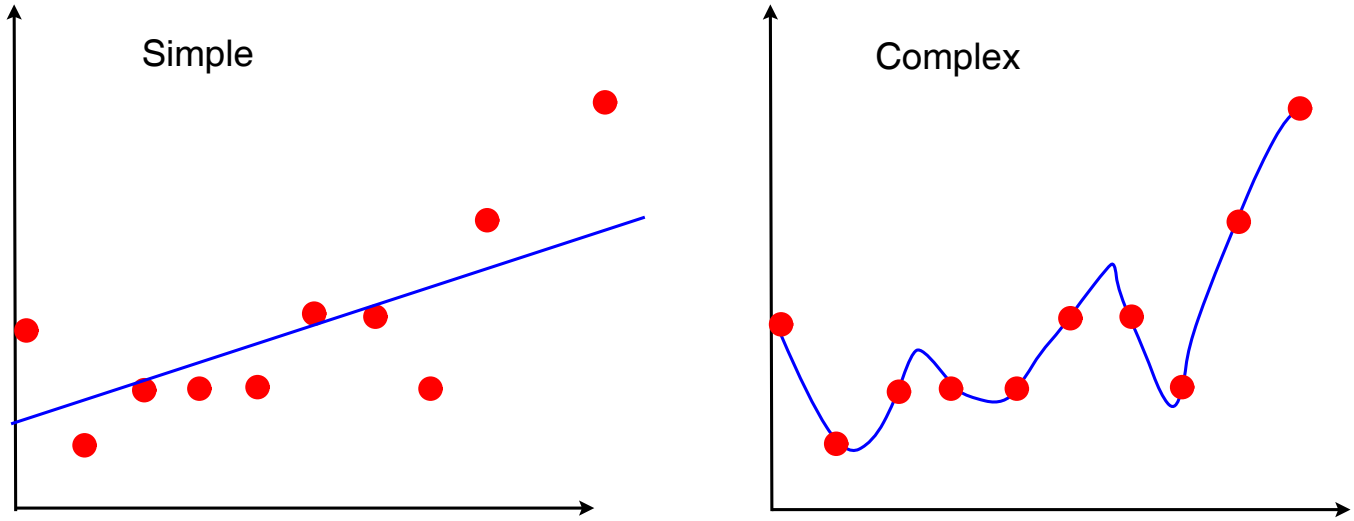
$$e(\hat{\theta}) = \frac{\mathcal{I}(\theta_*)^{-1}}{var(\hat{\theta})}$$

Bias-Variance Tradeoff

Evaluate $\hat{\theta}$ by the MSE to the true θ^*. Let $\bar{\theta} = \mathbb{E}[\hat{\theta}]$

$$\begin{aligned}\mathbb{E}[(\hat{\theta} - \theta^*)^2] &= \mathbb{E}[(\hat{\theta} - \bar{\theta} + \bar{\theta} - \theta^*)^2] \\ &= \mathbb{E}[(\hat{\theta} - \bar{\theta})^2] + (\bar{\theta} - \theta^*)^2 \\ &= \mathit{var}(\hat{\theta}) + \mathit{bias}^2(\hat{\theta})\end{aligned}$$

Why does this motivate regularization as a strategy?



- Complex: low bias high variance
- Simple: high bias low variance

Matrix Differentiation Cheatsheet

- $\frac{\partial \mathbf{a}^T \beta}{\partial \beta} = \mathbf{a}$
- $\frac{\partial \beta^T \mathbf{A} \beta}{\partial \beta} = 2\mathbf{A}\beta$, where $\mathbf{A}$ is symmetric
- $\frac{\partial |\mathbf{X}|}{\partial \mathbf{X}} = |\mathbf{X}|\mathbf{X}^{-T}$
- $\frac{\partial Tr(\mathbf{X}^T A)}{\partial \mathbf{X}} = A$

Along the learning of ML (and related courses), more matrix differentiations will be needed.

Reference on Matrix Differentiation

- Matrix cookbook: `https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`
- Useful topics that maybe used in ML:
  - Matrix differentiation (chapter 2).
  - Matrix operations: inverse ...
  - Matrix decompositions (chapter 5)
  - Statistics & probabilities (chapters 6, 7)
- Note: just a reference book. For further details, can check the references therein or textbooks.
- Sanity checking dimensions is always a good practice

Table of Contents

Bayes vs Frequentist

Estimators

Bayesian Linear Regression

Intro to GPs

Regression Revisted

We implicitly assumed a likelihood in the RSS derivation

$$y = X\beta + \epsilon$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I})$$

implies the likelihood

$$\mathrm{p}(y|X, \beta, \sigma) = \mathcal{N}(y|X\beta, \sigma^2 \mathbb{I})$$

RSS solution is the MLE solution

$$
\begin{aligned}
l(\beta) &= \log \mathrm{p}(y|X, \beta, \sigma) \\
&\propto -\frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta) + const \\
&\propto \beta^T X^T X \beta - 2\beta^T X y + const
\end{aligned}
$$

$$
\frac{\partial l(\beta)}{\partial \beta} := 0 \implies \hat{\beta} = \left(X^T X\right)^{-1} X^T y
$$

Bayesian Linear Regression

Let's adopt a bayesian approach to modelling β and compute its posterior (assume σ is known)

$$\mathrm{p}(\beta|y; X, \sigma) = \frac{\mathrm{p}(y|\beta; X, \sigma)\,\mathrm{p}(\beta)}{\int \mathrm{p}(y|\beta; X, \sigma)\,\mathrm{p}(\beta) d\beta}$$

We already have the likelihood form

$$\mathrm{p}(y|\beta; X, \sigma) = \mathcal{N}(y|X\beta, \sigma^2\mathbb{I})$$

I propose to use a normal prior (with mean zero for simplicity)

$$\mathrm{p}(\beta) = \mathcal{N}(\beta|0, \Sigma)$$

Why do I propose to use a Gaussian?

Partioned Gaussian Identities

Given a joint Gaussian $\mathcal{N}(\mathbf{x}|\mu, \Sigma)$, with the partitioning

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$$

The conditional distribution takes the form:

$$\mathrm{p}(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\mu_{a|b}, \Sigma_{a|b})$$

$$\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(\mathbf{x}_b - \mu_b)$$

$$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$$

The marginal takes the form:

$$\mathrm{p}(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a|\mu_a, \Sigma_{aa})$$

Completing the Square

Condiser a Guassian $\mathcal{N}(\mathbf{x}|\mu, \Sigma)$

$$\begin{aligned} log\mathcal{N}(\mathbf{x}|\mu, \Sigma) &\propto (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \\ &\propto \mathbf{x}^T \Sigma^{-1} \mathbf{x} - 2\mathbf{x}^T \Sigma^{-1} \mu + const \end{aligned}$$

Where the constant term is independent of $\mathbf{x}$.
What does it consist of?
This means if

$$log\ \mathrm{p}(\mathbf{x}) \propto \mathbf{x}^T \mathbf{A} \mathbf{x} - 2\mathbf{x}^T \mathbf{b} + c$$

then $\mathrm{p}(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{A}^{-1}\mathbf{b}, \mathbf{A}^{-1})$

Schur Complement

We will see that it is more convienent to work with the precision matrix $\Lambda = \Sigma^{-1}$.

$$\Lambda = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$$

We can move between the precision and covariance matrices as:

$$\Lambda_{aa} = \left(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}\right)^{-1}$$
$$\Lambda_{ab} = -\Lambda_{aa}\Sigma_{ab}\Sigma_{bb}^{-1}$$

See Bishop 2.3.1 for more details.

Derive Conditional

Lets assume $\mu = 0$ for simplicity, and derive $\mathrm{p}(\mathbf{x}_a|\mathbf{x}_b)$.

1. Write down the log of the joint distribution.

$$\log \mathrm{p}(\mathbf{x}) \propto \mathbf{x}_a^T \Lambda_{aa} \mathbf{x}_a + \mathbf{x}_a^T \Lambda_{ab} \mathbf{x}_b + \mathbf{x}_b^T \Lambda_{ba} \mathbf{x}_a + \mathbf{x}_b^T \Lambda_{bb} \mathbf{x}_b$$

2. Treat $\mathbf{x}_b$ as constant

$$\log \mathrm{p}(\mathbf{x}_a|\mathbf{x}_b) \propto \mathbf{x}_a^T \Lambda_{aa} \mathbf{x}_a + 2\mathbf{x}_a^T \Lambda_{ab} \mathbf{x}_b + c$$

3. Complete the square

$$\mathrm{p}(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a| - \Lambda_{aa}^{-1} \Lambda_{ab} \mathbf{x}_b, \Lambda_{aa}^{-1})$$

Derive Conditional

Lets assume $\mu = 0$ for simplicity, and derive $\mathrm{p}(\mathbf{x}_a|\mathbf{x}_b)$.

1. Write down the log of the joint distribution.

$$\log \mathrm{p}(\mathbf{x}) \propto \mathbf{x}_a^T \Lambda_{aa} \mathbf{x}_a + \mathbf{x}_a^T \Lambda_{ab} \mathbf{x}_b + \mathbf{x}_b^T \Lambda_{ba} \mathbf{x}_a + \mathbf{x}_b^T \Lambda_{bb} \mathbf{x}_b$$

2. Treat $\mathbf{x}_b$ as constant

$$\log \mathrm{p}(\mathbf{x}_a|\mathbf{x}_b) \propto \mathbf{x}_a^T \Lambda_{aa} \mathbf{x}_a + 2\mathbf{x}_a^T \Lambda_{ab} \mathbf{x}_b + c$$

3. Complete the square (Apply Schur complements)

$$\mathrm{p}(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\Sigma_{ab}\Sigma_{bb}^{-1}\mathbf{x}_b, \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})$$

Bayesian Regression Solution

1. Write down the log joint distribtion

$$\begin{aligned}\log \mathrm{p}(\beta, y) &\propto \log \mathrm{p}(y|\beta) + \log \mathrm{p}(\beta) \\ &\propto \frac{1}{\sigma^2}(y - X\beta)^T (y - X\beta) + \beta^T \Sigma_\beta^{-1} \beta\end{aligned}$$

2. Treat y as constant

$$\log \mathrm{p}(\beta|y) \propto \beta^T \left[\Sigma_\beta^{-1} + \frac{1}{\sigma^2} X^T X\right] \beta - \frac{2}{\sigma^2} \beta^T X^T y + c$$

3. Complete the square

$$\mathrm{p}(\beta|y) = \mathcal{N}(\beta|\mu_{\beta|y}, \Sigma_{\beta|y})$$

$$\Sigma_{\beta|y} = \left(\sigma^2 \Sigma_\beta^{-1} + X^T X\right)^{-1} \qquad \mu_{\beta|y} = \Sigma_{\beta|y} X^T y$$

I highly recommend going through Bishop Section 2.3
The identities 2.94 and 2.113 are extremely useful

Table of Contents

Bayes vs Frequentist

Estimators

Bayesian Linear Regression

Intro to GPs

A Very Breif Intro to Gaussian Processes

In the linear regression approaches we took so far, we

1. Observed data $D = \{y_i, \mathbf{x}_i\}$
2. Modelled y_i as corrupted observations of some $f(\mathbf{x}_i)$
3. Assumed a parametric form of f, with parameters θ

Then we took one of two approaches.

- Classical: find the "best" θ
- Bayesian: define $\mathrm{p}(\theta)$, compute $\mathrm{p}(\theta|D)$

Gaussian Processes adopt a Bayesian approach to directly model $\mathrm{p}(f|D)$ non-parametrically.

How to represent a distribution over functions

GPs assume that for any finite set $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$,

$$\mathrm{p}(\{f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N)\}) = \mathcal{N}(\mathbf{f}|\mu(\mathbf{x}), \Sigma(\mathbf{x}))$$

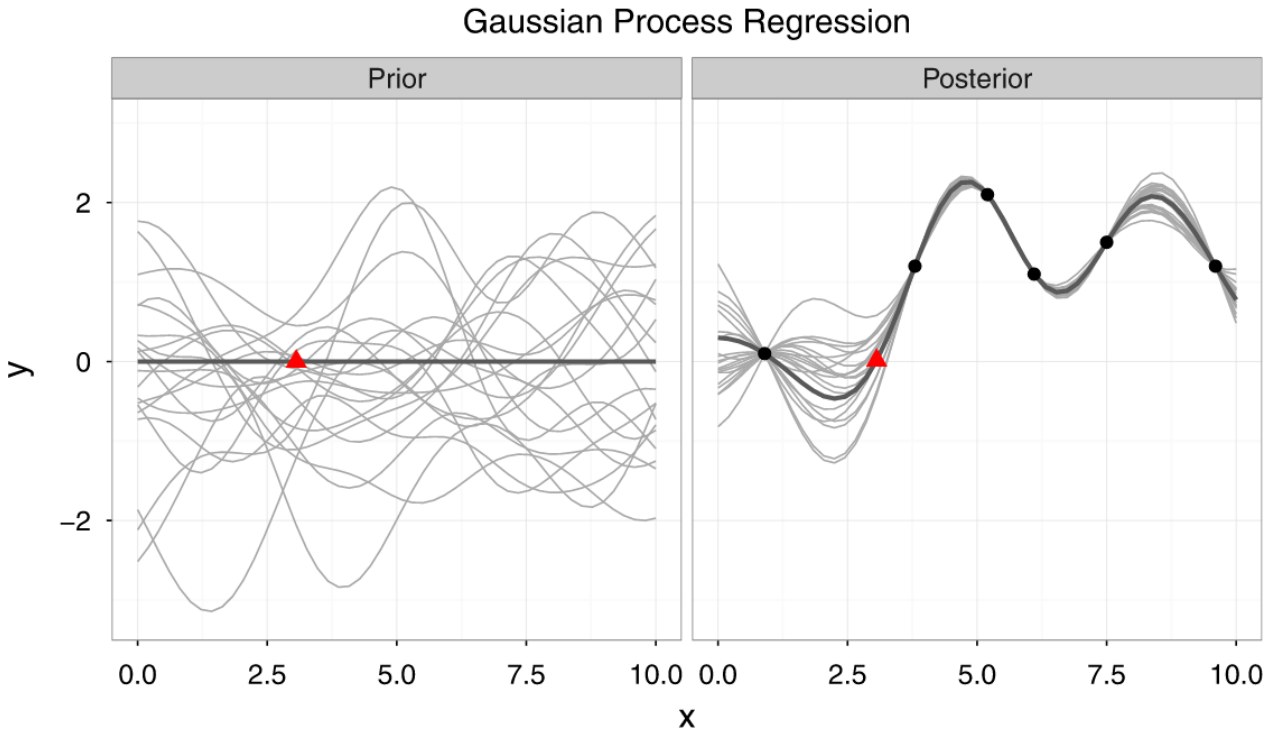
The covariance is computed as $\Sigma_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$
k is called the kernel function
it has some restrictions to keep Σ p.s.d.
(more later in the course)

Prediction in Gaussian Processes

GPs define a prior over functions, called the GP prior.
Consider the noise-free observations $\mathbf{f}$ at locations $\mathbf{x}$.
Its prior in a zero-mean GP is:

$$\mathrm{p}(\mathbf{f}) = \mathcal{N}(0, K(\mathbf{x}, \mathbf{x}))$$

The basic task of regression is to predict the values $\mathbf{f}_*$ at new locations $\mathbf{x}_*$ given observations $\mathbf{f}$ and $\mathbf{x}$.



A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions

Eric Schulz [a,*], Maarten Speekenbrink [b], Andreas Krause [c]

[a] *Department of Psychology, Harvard University, United States*
[b] *Department of Experimental Psychology, University College London, United Kingdom*
[c] *Department of Computer Science, Swiss Federal Institute of Technology, Zürich, Switzerland*

What is $p(\mathbf{f}_*|\mathbf{f})$?
We already know the joint

$$p(\begin{pmatrix}\mathbf{f}\\ \mathbf{f}_*\end{pmatrix}) = \mathcal{N}(\begin{pmatrix}\mathbf{f}\\ \mathbf{f}_*\end{pmatrix}|0, \begin{pmatrix}K & K_*\\ K_*^T & K_{**}\end{pmatrix})$$

And we already know how to condition Gaussian distributions:

$$\begin{aligned} p(\mathbf{f}_*|\mathbf{f}) &= \mathcal{N}(\mathbf{f}_*|\mu_*, \Sigma_*) \\ \mu_* &= K_*^T K^{-1}\mathbf{f} \\ \Sigma_* &= K_{**} - K_*^T K^{-1} K_* \end{aligned}$$

Good GP References

- Tutorial: https://www.youtube.com/watch?v=92-98SYOdlY
- Interactive blog post: https://distill.pub/2019/visual-exploration-gaussian-processes/
- Paper from earlier: https://www.biorxiv.org/content/10.1101/095190v3