Statistical Learning and Data Mining

Module 7: Estimation Methods

Semester 2, 2017

Discipline of Business Analytics, The University of Sydney Business School

Module 7: Estimation Methods

1. Empirical risk and regularised risk minimisation

2. Maximum likelihood

3. Bayesian statistics

Estimation methods

This is a reference module that provides an overview of estimation methods that we use in the unit.

We also include a brief discussion of Bayesian estimation, since anyone getting a degree in a data-related area must at least know what Bayesian statistics is (Bayesian approaches are highly effective in many settings).

Overview of estimation methods

- Empirical risk minimisation.
- Regularised risk minimisation.
- Maximum likelihood.
- Bayesian estimation.
- Maximum a posteriori (MAP) estimation.

Notation

- Let $p(\boldsymbol{y};\boldsymbol{\theta})$ denote a probability mass function or density function with associated parameter vector $\boldsymbol{\theta}$.
- Y_1, Y_2, $\ldots, Y_N$ is random sample from this distribution. The random variables are independent.
- $\mathcal{D} = \{y_1, \ldots, y_N\}$ are the actual observed values.
- $\widehat{\boldsymbol{\theta}}(\mathcal{D})$ is an estimator of $\boldsymbol{\theta}$.
- $\widehat{\boldsymbol{\theta}}$ an estimator (as above) or estimate of $\boldsymbol{\theta}$ according to the context.

Empirical risk and regularised risk minimisation

Empirical risk minimisation (key concept)

Let $\mathcal{D} = \{(y_i, \boldsymbol{x}_i)\}_{i=1}^N$ be training data and $f(\cdot; \boldsymbol{\theta})$ a prediction function that depends on the parameter vector $\boldsymbol{\theta}$. The **empirical risk minimisation** estimation principle estimates $\boldsymbol{\theta}$ as

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(\boldsymbol{x}_i; \boldsymbol{\theta}))$$

Example: least squares estimation whenever the loss is the squared error loss.

Regularised risk minimisation (key concept)

Minimising the empirical risk will typically lead to overfitting. In **regularised risk minimisation**, we estimate the model by solving the optimisation problem

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \; \left[\frac{1}{N}\sum_{i=1}^{N} L(y_i, f(\boldsymbol{x}_i;\boldsymbol{\theta}))\right] + \lambda\, C(\boldsymbol{\theta}),$$

where $C(\boldsymbol{\theta})$ measures the complexity of the prediction function and λ is a complexity penalty. Measuring complexity and selecting the penalty are core issues that we need to address.

Maximum likelihood

Maximum likelihood

Maximum likelihood estimation (MLE), which we have discussed in the context of linear regression is one of the most important concepts in statistics. We now present it more generally discuss inference.

Note that the nonparametric bootstrap method from the last module also applies to ML estimators.

ML for discrete distributions (key concept)

Let $p(y;\theta)$ be a discrete probability distribution. The likelihood function is

$$\begin{aligned}\ell(\theta) &= P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_N = y_N) \\ &= P(Y_1 = y_1)\, P(Y_2 = y_2)\, \ldots\, P(Y_N = y_N) \\ &= \prod_{i=1}^{N} p(y_i;\boldsymbol{\theta})\end{aligned}$$

The maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ is the value of $\boldsymbol{\theta}$ that maximises $\ell(\boldsymbol{\theta})$.

ML for continuous distributions (key concept)

Let $p(y;\theta)$ be a density function. The likelihood function is

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= p(y_1;\boldsymbol{\theta})\,p(y_2;\boldsymbol{\theta})\;\ldots\;p(y_N;\boldsymbol{\theta})\\ &= \prod_{i=1}^{N} p(y_i;\boldsymbol{\theta})\end{aligned}$$

The maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ is the value of $\boldsymbol{\theta}$ that maximises $\ell(\boldsymbol{\theta})$.

Maximum likelihood

- Even though $\ell(\boldsymbol{\theta})$ equals an expression that involves $p(y_i; \boldsymbol{\theta})$, we think of these functions in different ways.

- When considering a probability mass function or density $p(y; \boldsymbol{\theta})$, we consider y to be a variable, and $\boldsymbol{\theta}$ to be fixed.

- In the likelihood, $\boldsymbol{\theta}$ is a variable, and y is fixed.

Log-likelihood (key concept)

The log-likelihood is

$$\begin{aligned} L(\boldsymbol{\theta}) &= \log\left(\prod_{i=1}^{N} p(y_i;\boldsymbol{\theta})\right) \\ &= \sum_{i=1}^{N} \log p(y_i;\boldsymbol{\theta}) \end{aligned}$$

Because the log-likelihood is a monotonic transformation of the likelihood, maximising it is the same as maximising the likelihood.

Example: Bernoulli distribution

Suppose that $Y_1, \ldots, Y_N$ follow the Bernoulli distribution with parameter θ (the probability of a success).

$$p(y_i; \theta) = \theta^{y_i}(1-\theta)^{(1-y_i)}$$

$$\ell(\theta) = \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{(1-y_i)}$$

$$\begin{aligned} L(\theta) &= \sum_{i=1}^{N} \left[y_i \log(\theta) + (1-y_i)\log(1-\theta) \right] \\ &= \left(\sum y_i \right) \log(\theta) + (N - \sum y_i) \log(1-\theta) \end{aligned}$$

Example: Bernoulli distribution

Derivative of the log-likelihood with respect to θ:

$$\frac{dL(\theta)}{d\theta} = \frac{\sum y_i}{\theta} - \frac{N - \sum y_i}{1 - \theta}$$

The ML estimate therefore satisfies

$$\frac{\sum y_i}{\widehat{\theta}} = \frac{N - \sum y_i}{1 - \widehat{\theta}}.$$

The solution is the sample proportion:

$$\widehat{\theta} = \frac{\sum_{i=1}^{N} y_i}{N}.$$

Properties of the ML estimator

- The MLE is consistent.
- The MLE is asymptotically unbiased (if there is a bias, it goes to zero as $N \to \infty$).
- The MLE is **asymptotically optimal**: it has the smallest asymptotic variance of any asymptotically unbiased estimator.

Inference for the ML estimator

The **score function** is

$$\boldsymbol{s}(\widehat{\boldsymbol{\theta}}) = \nabla L(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$$

For example, when the parameter is a scalar

$$s(\widehat{\theta}) = \sum_{i=1}^{N} \frac{d \log p(y_i; \theta)}{d\theta}\bigg|_{\theta=\widehat{\theta}}$$

Inference for the ML estimator

The **observed information matrix** is the negative of the gradient function, or the negative Hessian of the log-likelihood

$$\boldsymbol{J}(\widehat{\boldsymbol{\theta}}(\mathcal{D})) = -\nabla^2_{\boldsymbol{\theta}}\, L(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$$

When the parameter is a scalar,

$$J(\widehat{\theta}) = \sum_{i=1}^{N} \left.\frac{d^2 \log p(y_i)}{d\theta^2}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}.$$

Inference for the ML estimator

We define the **Fisher information matrix** as the expected value of the observed information matrix

$$\boldsymbol{I}_N(\widehat{\boldsymbol{\theta}}) = E_{\boldsymbol{\theta}}\left[\boldsymbol{J}(\widehat{\boldsymbol{\theta}}(\mathcal{D}))\right]$$

Inference for the ML estimator

A standard result shows that the sampling distribution of the ML converges to the normal distribution

$$\widehat{\boldsymbol{\theta}} \to N(\theta, \boldsymbol{I}_N^{-1}(\boldsymbol{\theta}))$$

as $N \to \infty$.

That suggests the large sample approximations

$$N(\boldsymbol{\theta}, \boldsymbol{I}_N(\widehat{\boldsymbol{\theta}})^{-1}) \text{ or } N(\boldsymbol{\theta}, \boldsymbol{J}(\widehat{\boldsymbol{\theta}})^{-1})$$

Example: Bernoulli distribution

Continuing the example, the observed information matrix is

$$-\frac{d^2L(\theta)}{d\theta^2} = \frac{\sum y_i}{\theta^2} + \frac{N - \sum y_i}{(1-\theta)^2}$$

Since $E(Y) = \theta$,

$$E(J(\theta)) = \frac{N}{\theta(1-\theta)},$$

so that

$$I_N^{-1} = \frac{\theta(1-\theta)}{N},$$

which is familiar as the variance of a sample proportion from basic statistics.

Inference for the ML estimator

The corresponding estimates for the standard errors of individual parameters are

$$\mathsf{SE}(\widehat{\theta}_j) = \sqrt{\boldsymbol{I}_N(\widehat{\boldsymbol{\theta}})^{-1}_{jj}} \text{ or } \mathsf{SE}(\widehat{\theta}_j) = \sqrt{\boldsymbol{J}(\widehat{\boldsymbol{\theta}})^{-1}_{jj}}$$

A large sample $100 \times (1-\alpha)\%$ confidence interval is

$$\widehat{\theta}_j \pm z_{\alpha/2} \times \mathsf{SE}(\widehat{\theta}_j)$$

Inference for the ML estimator

The following large sample approximation leads to accurate confidence intervals and hypothesis tests

$$2\left(L(\widehat{\boldsymbol{\theta}}) - L(\boldsymbol{\theta})\right) \sim \chi^2_d,$$

where d is the number of parameters in $\boldsymbol{\theta}$.

Parametric Bootstrap

Suppose that we estimate a model $p(\boldsymbol{y};\boldsymbol{\theta})$ by maximum likelihood to obtain $\widehat{\boldsymbol{\theta}}$.

In the **parametric bootstrap**, we draw many different samples $\boldsymbol{y}_s^*$ from the estimated model $p(\boldsymbol{y};\widehat{\theta})$ for $s = 1,\ldots,S$. For each bootstrap sample, we obtain a ML estimate $\boldsymbol{\theta}_s^* = \widehat{\theta}(\boldsymbol{y}_s^*)$.

The bootstrap distribution is the empirical distribution of $\{\boldsymbol{\theta}_s^*\}_{s=1}^S$. As before, using the parametric bootstrap has the advantage of not relying on large sample results.

Bayesian statistics

Bayesian inference (key concept)

In **Bayesian inference**, we specify a sampling model $p(\boldsymbol{y}|\boldsymbol{\theta})$ (likelihood) and a **prior distribution** $p(\boldsymbol{\theta})$, which represents the uncertainty about $\boldsymbol{\theta}$ before analysing the data.

The core of Bayesian statistics is to compute the **posterior distribution** $p(\boldsymbol{\theta}|\boldsymbol{y})$, which summarises everything that we know about $\boldsymbol{\theta}$.

Posterior distribution (key concept)

It follows from **Bayes' theorem** that

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

where the denominator is called the **marginal likelihood** or **model evidence**, which we also write as $p(\boldsymbol{y})$.

Bayesian inference

In Bayesian statistics, the parameter $\boldsymbol{\theta}$ is random and we make inference conditional on the data.

In classical (frequentist) inference, the parameter is fixed and the data are random.

Predictive distribution

The posterior distribution leads to the predictive distribution

$$p(y_0|\boldsymbol{y}) = \int p(y_0|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}$$

Bayesian decision theory

In the Bayesian approach to decision theory, we take the action that minimises the **posterior expected loss**. In the supervised learning setting, the **Bayes estimator**, also called the **Bayes decision rule**, is

$$\delta(\boldsymbol{x}) = \underset{a}{\operatorname{argmin}} \int L(y_0, a)\, p(y_0|\boldsymbol{x}_0, \boldsymbol{\theta})\, dy_0,$$

where $L(\cdot, a)$ is the loss function, a is the action, and $p(y_0|\boldsymbol{x}_0, \boldsymbol{\theta})$ is the predictive distribution of the response for a new observation with input value $\boldsymbol{x}_0$.

Bayesian decision theory

$$\delta(\boldsymbol{x}) = \operatorname*{argmin}_{a} \int L(y_0, a)\, p(y_0 | \boldsymbol{x}_0, \boldsymbol{\theta})\, dy_0,$$

Unlike in classical statistics, where the expectation that defines the risk is with respect to an unknown population, the posterior expected loss is computable.

The cogency of the Bayesian argument therefore makes it an appealing approach for many problems.

Example: Bayesian linear regression

The Gaussian linear regression model specifies

$$p(y|\boldsymbol{x},\, \boldsymbol{\beta}) = N(\boldsymbol{\beta}^T\boldsymbol{x}, \sigma^2).$$

The sampling model is therefore multivariate normal,

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}) = N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_N).$$

Example: Bayesian linear regression

In the Bayesian approach, we need to choose the prior. We assume that

$$p(\boldsymbol{\beta}) = N(\boldsymbol{\beta}_0, \boldsymbol{V}_0),$$

where $\boldsymbol{\beta}_0$ and $\boldsymbol{V}_0$ are hyperparameters.

Example: Bayesian linear regression

The posterior is (the symbol $\propto$ means “proportional to”)

$$p(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta})p(\boldsymbol{\beta})$$

The formula for the multivariate normal density gives

$$p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right)$$

$$p(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\boldsymbol{V}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right)$$

We can work with these densities directly, but it is easier to use standard results for the multivariate normal distribution.

Linear Gaussian systems

Let Y and X be two random vectors such that

$$p(\boldsymbol{x}) = N(\boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$$

$$p(\boldsymbol{y}|\boldsymbol{x}) = N(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{\Sigma}_{\boldsymbol{y}})$$

In several applications of interest, X is a hidden variable, Y is a noisy observation observation, and we interested in recovering the conditional $p(\boldsymbol{x}|\boldsymbol{y})$.

Linear Gaussian systems

We can show that

$$p(\boldsymbol{x}|\boldsymbol{y}) = N(\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}}, \boldsymbol{\Sigma}_{\boldsymbol{x}|\boldsymbol{y}}),$$

where

$$\boldsymbol{\Sigma}_{\boldsymbol{x}|\boldsymbol{y}}^{-1} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} + \boldsymbol{A}^T \boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1} \boldsymbol{A}$$

and

$$\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}} = \boldsymbol{\Sigma}_{\boldsymbol{x}|\boldsymbol{y}} \left[\boldsymbol{A}^T \boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1} (\boldsymbol{y} - \boldsymbol{b}) + \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\mu}_{\boldsymbol{x}}\right]$$

Bayesian linear regression

Using the previous result, we can show that the posterior for β in the linear regression model is

$$p(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{y}) = N(\boldsymbol{\beta}_N, \boldsymbol{V}_N)$$

where

$$\boldsymbol{V}_N^{-1} = \boldsymbol{V}_0^{-1} + \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X},$$

and

$$\boldsymbol{\beta}_N = \boldsymbol{V}_N\boldsymbol{V}_0^{-1}\beta_0 + \frac{1}{\sigma^2}\boldsymbol{V}_N\boldsymbol{X}^T\boldsymbol{y}.$$

Bayesian linear regression

Posterior mean and variance:

$$\boldsymbol{\beta}_N = \boldsymbol{V}_N \boldsymbol{V}_0^{-1} \boldsymbol{\beta}_0 + \frac{1}{\sigma^2} \boldsymbol{V}_N \boldsymbol{X}^T \boldsymbol{y}$$

$$\boldsymbol{V}_N^{-1} = \boldsymbol{V}_0^{-1} + \frac{1}{\sigma^2} \boldsymbol{X}^T \boldsymbol{X}$$

Let $\boldsymbol{V}_0$ be a diagonal matrix $\tau^2 \boldsymbol{I}$. If we set $\tau^2 \to \infty$, the mean of the posterior converges to $(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$, the classical OLS estimator.

Maximum a posteriori (MAP) estimation

The **maximum a posteriori (MAP) estimator** is the mode of the posterior distribution

$$\widehat{\boldsymbol{\theta}}_{\mathsf{MAP}} = \underset{\boldsymbol{\theta}}{\mathsf{argmax}}\ \log(p(\boldsymbol{y}|\boldsymbol{\theta})) + \log(p(\boldsymbol{\theta}))$$

Many regularised risk minimisation methods have an interpretation as MAP estimation, without necessarily being fully Bayesian. In a Bayesian approach to supervised learning, the prior plays the role of incorporating information to reduce overfitting.

Review questions

- What is regularised risk minimisation?
- What is maximum likelihood estimation?
- What is Bayesian statistics?
- What is a posterior distribution?