QBUS6810: Statistical Learning and Data Mining

Lecture 5: Estimation Methods

Semester 1, 2019

Discipline of Business Analytics, The University of Sydney Business School

Lecture 5: Estimation Methods

1. Empirical risk minimisation
2. Maximum Likelihood Estimation (MLE)
3. Bayesian approach

经验

Empirical risk minimisation

A parametric setting

Let $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ be the training data and let $f(\boldsymbol{x}; \boldsymbol{\theta})$ denote the candidate prediction functions, which depend on the parameter vector $\boldsymbol{\theta}$.

Estimating f comes down to estimating the "true" value of the parameter vector $\boldsymbol{\theta}$:

$$\widehat{f}(\boldsymbol{x}) = f(\boldsymbol{x}; \widehat{\boldsymbol{\theta}})$$

In linear regression the parameter is the vector of regression coefficients, which is traditionally denoted by $\boldsymbol{\beta}$:

$$f(\boldsymbol{x}; \boldsymbol{\beta}) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

Empirical risk minimisation

The **empirical risk minimisation** estimation approach solves the following optimisation problem, in which L is the loss function:

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

The $\underset{\boldsymbol{\theta}}{\operatorname{argmin}}$ operation identifies the value of θ that minimises the function on the right-hand side.

Note that in the case of linear regression and the squared error loss this approach produces the OLS estimator.

Regularised empirical risk minimisation

Minimising the empirical risk will typically lead to overfitting. In **regularised** empirical risk minimisation we solve:

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

where $C(\boldsymbol{\theta})$ is some measure of the complexity of the prediction function, and λ is a non-negative weight in the complexity penalty $\lambda\, C(\boldsymbol{\theta})$.

Next week we will discuss a number of linear regression approaches that implement regularised empirical risk minimisation

Maximum Likelihood Estimation (MLE)

Before formally defining MLE, we will consider a brief introductory example.

Probability vs. Statistics

Probability Question: X counts the number of successes in 20 independent random trials with probability of success θ (i.e. X has Binomial distribution with parameters 20 and θ).

Assume $\theta = 0.3$. What is the probability of observing $X = 4$?

$$P(X = 4) = \binom{20}{4}(0.3)^4(1 - 0.3)^{20-4} = 0.1304$$

Statistics Question: X has Binomial distribution with parameters 20 and θ. We observed $X = 4$. How do we estimate θ?

Statistics question

X has Binomial distribution with parameters 20 and θ. We observed $X = 4$.
How do we estimate θ?

A simple special case:

Suppose we know that θ is either 0.3 or 0.6. Which parameter value should we choose based on the observed data, $X = 4$?

$$P(X = 4) = \binom{20}{4}\theta^4(1-\theta)^{20-4}$$

$$P(X = 4\,;\,\theta = 0.3) = \binom{20}{4}(0.3)^4(0.7)^{20-4} = 0.1304$$

$$P(X = 4\,;\,\theta = 0.6) = \binom{20}{4}(0.6)^4(0.4)^{20-4} = 0.0003$$

Statistics question (special case)

Suppose we know that θ is either 0.3 or 0.6. Which parameter value should we choose based on the observed data, $X = 4$?

$$P(X = 4\,;\, \theta = 0.3) = 0.1304$$

$$P(X = 4\,;\, \theta = 0.6) = 0.0003$$

Under the choice $\theta = 0.3$, the observed data, $X = 4$, is much more *likely*. Thus, it makes sense to pick $\theta = 0.3$ from the available two options.

This is the main idea of the maximum likelihood approach.

Statistics question

X has Binomial distribution with parameters 20 and θ. We observed $X = 4$.

How do we estimate θ?

Given the observed data, the *Likelihood* is a function of the unknown parameter:

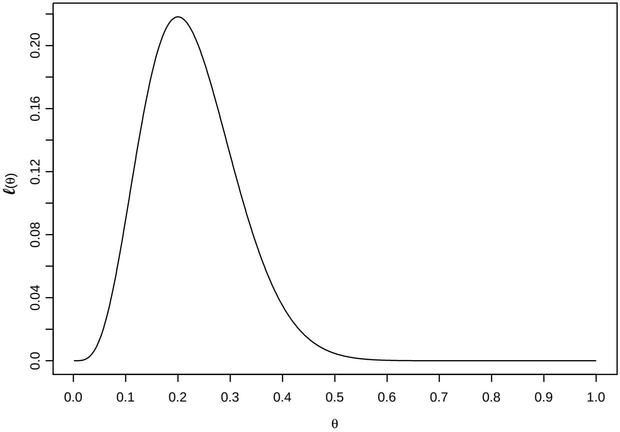
$$\ell(\theta) = P(X = 4\,;\,\theta) = \binom{20}{4}\theta^4(1-\theta)^{20-4}$$

$\ell(\theta)$ gives the probability of observing the data at hand for each value of the parameter θ.

MLE: choose the value of θ that maximizes $\ell(\theta)$

In other words, choose θ that corresponds to the maximum probability of observing the data at hand

Likelihood function, $\ell(\theta)$, in the binomial example



MLE: $\widehat{\theta} = 0.2$

Notation

- Let $p(\boldsymbol{y}; \boldsymbol{\theta})$ denote a probability mass function or a density function (for a random variable Y), which depends on the parameter (vector) $\boldsymbol{\theta}$.

- Y_1, Y_2, $\ldots, Y_n$ is a random sample from the above distribution; we think of Y_i as independent identically distributed random variables.

- $y_1, \ldots, y_n$ are the actual observed values (the observed sample); these are non-random.

- $\widehat{\boldsymbol{\theta}}$ is an estimator of $\boldsymbol{\theta}$ constructed from the sample.

Maximum likelihood for discrete distributions

Let $p(y;\boldsymbol{\theta})$ be a discrete probability distribution that depends on parameter $\boldsymbol{\theta}$. Given $\boldsymbol{\theta}$, the **likelihood function**, $\ell(\boldsymbol{\theta})$ equals the corresponding probability of the observed data:

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

Here $\boldsymbol{\theta}$ is the argument of the likelihood function and y_i are fixed.

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

Maximum likelihood for continuous distributions

Let $p(y; \boldsymbol{\theta})$ be a density function. Given $\boldsymbol{\theta}$, the likelihood equals the corresponding density function, evaluated at the observed data:

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

Again, $\boldsymbol{\theta}$ is the argument of the likelihood function and y_i are fixed.

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

Log-likelihood

The log-likelihood is

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

单调

Because $L(\boldsymbol{\theta})$ is a monotonic transformation of $\ell(\boldsymbol{\theta})$, maximising the log-likelihood leads to the same solution, $\widehat{\boldsymbol{\theta}}$, as when maximising the likelihood.

Log-likelihood is often easier to work with than likelihood.

Example: Bernoulli distribution

Suppose that $Y_1, \ldots, Y_n$ come from the Bernoulli distribution with parameter θ (i.e. $Y_i = 1$ with probability θ and $Y_i = 0$ with prob. $1-\theta$).

Note that we can write:

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

Thus,

$$\ell(\theta) = \prod_{i=1}^{n} p(y_i; \theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{(1-y_i)}$$

We now take the $\log$ to get the log-likelihood:

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

Example: Bernoulli distribution

Derivative of the log-likelihood with respect to θ:

$$\frac{dL(\theta)}{d\theta} = \frac{\sum_{i=1}^n y_i}{\theta} - \frac{n - \sum_{i=1}^n y_i}{1-\theta}$$

Setting the derivative to zero, the MLE, $\widehat{\theta}$ must satisfy:

$$\frac{\sum_{i=1}^n y_i}{\widehat{\theta}} = \frac{n - \sum_{i=1}^n y_i}{1-\widehat{\theta}}$$

The solution is the sample proportion:

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

Example: Gaussian MLR

We treat the x values as fixed (non-random), and focus on the
Multiple Linear Regression
estimation of the β. Recall that under the Gaussian MLR random
每条记录误差都是独立同分布的:
variables Y_i are independent $N\Big(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}\,,\ \sigma^2\Big)$.

成比例的

Using $\propto$ to denote "proportional to" and leaving out positive multiplicative constants we can write the likelihood as follows:

$$
\begin{aligned}
\ell(\boldsymbol{\beta}) &= \prod_{i=1}^{n} p(y_i; \boldsymbol{\beta}) \\
&\propto \prod_{i=1}^{n} \exp\left\{ -\frac{1}{2\sigma^2} (y_i - \beta_0 - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 \right\} \\
&= \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 \right\}.
\end{aligned}
$$

Example: Gaussian MLR

$$\ell(\boldsymbol{\beta}) \;\propto\; \exp\Big\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_p x_{ip})^2\Big\}.$$

Maximizing the above expression over $\boldsymbol{\beta}$ is equivalent to maximizing the part in the exponent:

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\Big(y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_p x_{ip}\Big)^2$$

which is equivalent to *minimizing* the residual sum of squares.

Thus: under the Gaussian multiple linear regression model, the MLE and the OLS estimators of the regression coefficients are identical.

Large sample properties of the ML estimator

- The MLE converges to the true parameter value as $n \to \infty$.

渐近

- 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 variance (as $n \to \infty$) of any asymptotically unbiased estimator.

Bayesian approach

Bayesian inference

Recall that in the classical statistics the true parameter is fixed (nonrandom).

In Bayesian statistics, however, the parameter $\boldsymbol{\theta}$ can be treated as random, and we make inference about it conditional on the data.

们以数据为条件推断它

Bayesian inference

In Bayesian inference, in addition to a sampling model $p(\boldsymbol{y}|\boldsymbol{\theta})$ we specify a **prior distribution** $p(\boldsymbol{\theta})$, which represents our beliefs about the parameter $\boldsymbol{\theta}$ before we see any data.

后验分布

The Bayesian approach computes the **posterior distribution** $p(\boldsymbol{\theta}|\boldsymbol{y})$, which represents our updated beliefs about $\boldsymbol{\theta}$ after we observe the data $\boldsymbol{y}$.

As before, $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|\boldsymbol{y})$ denote either probability mass functions or probability densities, depending on the context.

Posterior distribution

It follows from **Bayes' theorem** that the posterior distribution satisfies:

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

Again using the $\propto$ notation and leaving out multiplicative constants that do not depend on $\boldsymbol{\theta}$ we can write the posterior as follows:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

We can restate the above relationship in words:

Posterior is proportional to Likelihood times Prior

Example: Gaussian MLR

We continue with the earlier example. Recall that under the Gaussian MLR random variables Y_i are independent $N\Big(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}\,,\, \sigma^2\Big)$.

We have shown that the likelihood has the following form:

$$p(\boldsymbol{y}|\boldsymbol{\beta}) \;\propto\; \exp\Big\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} + ... - \beta_p x_{ip})^2\Big\}.$$

Example: prior

In the Bayesian approach, we need to choose a prior. Under our prior, the true regression coefficients β_j, for $j = 1, ..., p$ are independent $N(0, \tau^2)$, for some $\tau^2 > 0$.

We will not put an informative prior on the intercept β_0 (this is equivalent to using a flat prior density for β_0, i.e., a density that is proportional to the constant 1).

我们不会在截距β0上提供信息先验
（这相当于使用β0的平坦先验密度，即与常数1成比例的密度）

Note that this prior satisfies

$$p(\boldsymbol{\beta}) \propto \prod_{j=1}^{p} \exp\left\{-\frac{\beta_j^2}{2\tau^2}\right\}$$

$$= \exp\left\{-\frac{1}{2\tau^2}(\beta_1^2 + ... + \beta_p^2)\right\}$$

Example: posterior

Consequently,

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

$$\propto \exp\Big\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_p x_{ip})^2\Big\}$$

$$\times \exp\Big\{-\frac{1}{2\tau^2}(\beta_1^2 + ... + \beta_p^2)\Big\}$$

$$\propto \exp\Big\{-\frac{1}{2\sigma^2}\Big[\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_p x_{ip})^2 + \frac{\sigma^2}{\tau^2}(\beta_1^2 + ... + \beta_p^2)\Big]\Big\}$$

Example: posterior

$$p(\boldsymbol{\beta}|\boldsymbol{y}) \quad \propto \quad \exp\left\{ -\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \frac{\sigma^2}{\tau^2}\sum_{j=1}^{p}\beta_j^2\right]\right\}.$$

In Tutorial 6 you will show that the mode of this posterior density corresponds to the Ridge regression estimator, which we will cover next week.

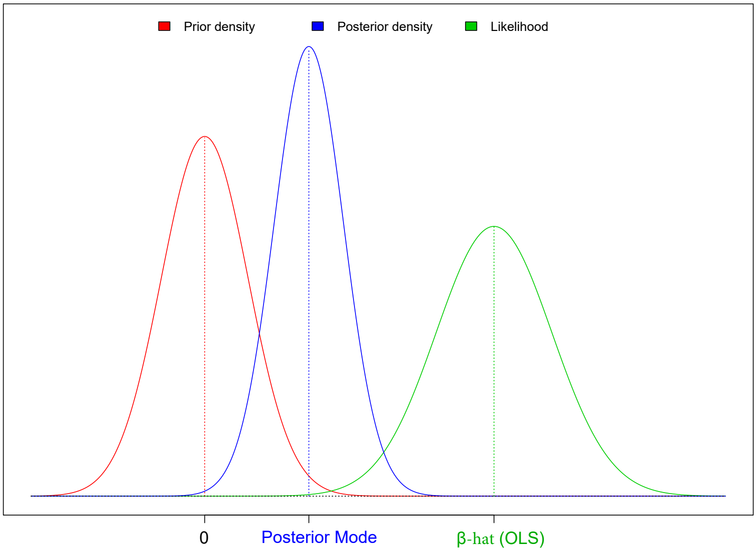
The function in the exponent is quadratic in $\boldsymbol{\beta}$. In fact, it can be shown that the posterior distribution of $\boldsymbol{\beta}$ is Gaussian (we will not worry about the proof of this fact).

指数函数在β中是二次的。事实上，可以证明β的后验分布是高斯分布（我们不
心这个事实的证明）。
二次
该分布的模式（并且因此均值）收敛于OLS估计量β，
因为先验的方差τ2变为无穷大（即，先验变得越来越少，信息量越来越少）。
当先验的方差变为零时（即，先验变得越来越集中在β= 0附近），模式收敛到

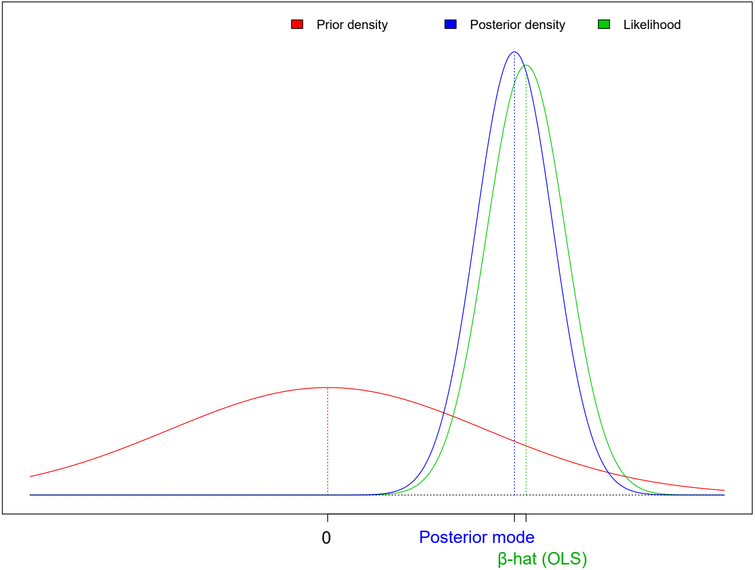
The mode (and, thus, the mean) of this distribution converges to the OLS estimator $\widehat{\boldsymbol{\beta}}$ as the variance of the prior, τ^2 goes to infinity (i.e. the prior becomes less and less informative).

The mode converges to zero as the variance of the prior goes to zero (i.e. the prior becomes more and more concentrated around $\boldsymbol{\beta} = 0$).

Informal Illustration



Informal Illustration



Maximum a posteriori (MAP) estimation

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

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}_{\mathsf{MAP}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ p(\boldsymbol{\theta}|\boldsymbol{y}) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ \log(p(\boldsymbol{\theta}|\boldsymbol{y})) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ \Big[\log(p(\boldsymbol{y}|\boldsymbol{\theta})) + \log(p(\boldsymbol{\theta}))\Big]
\end{aligned}$$

Inside the square brackets we have the log-likelihood plus the log-prior. Incorporating the prior can be thought of as regularisation, and may reduce overfitting.

结合先验可以被认为是正则化，并且可以减少过度拟合。

Many regularised risk minimisation methods have an interpretation as MAP estimation, without necessarily being fully Bayesian.

Review questions

- What is regularised risk minimisation?
- What is maximum likelihood estimation?
- What is the MLE of the regression coefficients under the Gaussian MLR model?
- What are prior and posterior distributions, and what is the relationship between likelihood, prior and posterior?
- What is a MAP estimator?