Probability and Statistics II

Changmin Yu

February 2020

1 Maximum Likelihood Estimation

Likelihood is a concept useful for comparing alternative models with each other. It describes the probability of the observed data being generated by an underlying model (of a random process) having certain parameters. In other words, it is a measure of goodness of fit between the data and the proposed model with given parameter values. More formally:

- Given data $y \in \mathbb{R}^n$ as realisations of a random variable Y, specify its density $f(y;\theta)$ up to some unknown vector of parameters $\theta \in \Theta \subset \mathbb{R}^d$, where Θ is the parameter space.
- Define the **likelihood function** as a function of the parameters θ:

  $$L(\theta) = L(\theta; y) = c(y)f(y;\theta) \tag{1}$$

  where $c(y)$ is some unknown constant for normalisation
- The maximum likelihood estimator (MLE) of θ, $\hat{\theta}$, is the value of parameters such that $\hat{\theta}$ maximises $L(\theta)$.

  $$\hat{\theta} = \arg\max_{\theta} L(\theta; y) \tag{2}$$
- It is common to work with the log-likelihood function.

2 Maximum A Posteriori Estimation

An alternative way of evaluating the quality of models or selecting between multiple alternative parameter values is to calculate the opposite: the probability of a model's parameters given the data observed. Then the most probable values can be selected. This probability can be calculated from the likelihood using Bayes' theorem:

- Bayes' theorem:

  $$f(x|y) = \frac{f(y|x)f(x)}{f(y)} = \frac{f(y|x)f(x)}{\int f(y|x)f(x)dx} \tag{3}$$

- Given data y and parametric density function $f(y;\theta)$, the maximum a posteriori (MAP) estimator of θ given its prior belief $g(\theta)$ is

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \frac{f(y;\theta)g(\theta)}{\int f(y;\theta)g(\theta)d\theta} \tag{4}$$

Note that in order to calculate the posterior probability $f(\theta;y)$, we need to have a notion of how likely the parameter values are to occur in general in the wild i.e. the probability distribution of parameter values - $g(\theta)$. This is called the **prior**. What is happening above is that this prior belief, $g(\theta)$, is being adjusted in light of the data observed, y, to yield new probability values, $f(\theta;y)$; then the value with maximum probability is selected.

3 Ordinary Least Squares

Linear regression is a simple statistical model where the output variable(s) Y are modeled by / predicted to be a linear combination of input variable(s) X, plus some unknown error. In other words:

- $Y = X\beta + \epsilon$

where β is a matrix of coefficients. These coefficients are typically selected by minimizing the sum of squared differences between the predicted and actual values, i.e. via ordinary least squares (OLS):

- OLS estimator

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} ||Y - X\beta||^2 = (X^T X)^{-1} X^T Y \tag{5}$$

  Note that this is also an orthogonal projection onto the column space of X.

  - Recall from the Linear Algebra II notes that the column space of matrix X is a subspace of $\mathbb{R}^m$ that contains all vectors reachable via a linear combination of X's column vectors. Let's call this subspace $\text{Col}(X) = W$.
  - The orthogonal projection involves finding the point in a subspace closest to the original point. More specifically, it involves decomposing a vector (let's call it Y) into a component that is contained within the subspace (let's call it Y_W) and a component that is completely **orthogonal** to the subspace (let's call it $Y_{W^\perp}$, where $Y = Y_W + Y_{W^\perp}$). The orthogonal projection is then the **set of weights c** that must be applied to the subspace basis vectors in order to yield Y_W (i.e. $Y_W = X\boldsymbol{c}$).

  The proof of the above equivalences is left as an exercise.