Deterministic model and Probabilistic model

Yuwei Zheng

1 Introduction

This short theoretical report aims to help students to understand what are deterministic and probabilistic models, what are their differences, and how to use them for learning.

2 Deterministic Model

The deterministic model is a mathematical model in which outcomes are precisely determined through known relationships among states and events, without any room for random variation. In simple words, we can only observe one y value at a given value of x. This is also called a model without noisy. For example, the model for predicting stock prices is a deterministic model; at every 15 seconds, there is only 1 stock price generated.

2.1 Estimation Process

If the model we learn is a deterministic model, we can simply do the following to find the best fit line:

1. Set a target function to predict Y: $f : x \to\ y$
2. Randomly select training samples from the observed data set. $D = \{x1, y1; x2, y2; ...; xn, yn\}$
3. Define a set of hypothesis that you think are likely to represent the true function f:$H = \{h1, h2, \ldots, hn\}$
4. Minimise the loss function to find the hypothesis that fits the training sample best.

$$L2 : E_{in} = \frac{1}{N} \sum_{i=1}^{n} [y_i - h^*(x_i)]^2$$

$$L1 : E_{in} = \frac{1}{N} \sum_{i=1}^{n} |y_i - h^*(x_i)|$$

$$L0 : E_{in} = \left\{ \begin{array}{l} 1\ if\ \mathrm{y}_i \neq h^*(x_i) \\ 0\ if\ \mathrm{y}_i = h^*(x_i) \end{array} \right\}$$

Those loss functions measure how different the observed y values and it's prediction value $g^*(x)$ are by adding up the distance in between. Before minimising the loss function, you should choose the level of loss functions yourself for serving your needs. If you are afraid of the impact of outliers, then do not use the L2 function.

3 Probabilistic Model

Probabilistic models incorporate random variables and probability distributions into the model of an event or phenomenon. While a deterministic model gives a single possible outcome for an event, a probabilistic model gives a probability distribution as a solution. This means that we can observe more than one y value for each given value of x. For example, we can have many incomes at each age; incomes are distributed in a certain way at the value of x.

3.1 Estimation Process

Because the OLS model ignores the fact that probabilistic models are heteroskedastic; if the model we learn is a probabilistic model, we cannot apply the OLS estimation directly to it. Thus, we should understand the target distribution first:

1. Set a target distribution to predict Y: $p(x, y) = p(y|x)p(x)$
2. Randomly select training samples from the observed data set. $D = \{x1, y1; x2, y2; ...; xn, yn\}$
3. Define a set of hypothesis that you think are likely to represent the true distribution f: $H = \{h_1(y_1|x_1), ...h_n(y_n|x_n)\}$
4. Minimise the loss function to find the distribution that fits the training sample best. This process is called minimising KL Divergence.

$$E_{in} = \frac{1}{N}\sum_{i=1}^{n}(log\ p(x_i) - log\ q(x_i))$$

3.2 KL Divergence

KL Divergence can be seen as the loss function of probabilistic models, it equals to cross entropy subtract entropy: $D_{KL}(p||q) = H(p, q) - H(p)$. To understand this, we need to understand what entropy and cross entropy are first.

3.2.1 Entropy

Entropy is a measurement of disorder.

$$H(p) = -\sum_{i=1}^{n} p(x_i) \log\ p(x_i)$$

For example, I have 2 red balls and 2 blue balls on the table but covered by a black sheet. They are arranged as R,R,B,B, but you can not see it. The probability of guessing each ball correctly is $P(correct) = (\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$. Now think about this question: What is the least number of questions you need to ask to find out the colour for each ball? The less the questions you ask the less disordered this arrangement is. In this game, you only need to ask 1 question: is this red? If the answer is yes, it is a red ball, otherwise it is a blue ball.

$$E(Number\ of\ questions) = \frac{1}{4}(1 + 1 + 1 + 1) = 1 = -\frac{1}{4} \times (\log_2(\frac{1}{2})^4) = Entropy$$

For another game that I have 4 red balls on the table; now you have to ask questions to guess how balls are arranged. As we all know that $P(red) = 1$, we can draw the conclusion that those 4 balls are arranged as R,R,R,R without asking any further question. Thus entropy of this game is 0.

3.2.2 CROSS ENTROPY

Cross entropy is the entropy calculated by playing game 2 with game 1's strategy.

$$H(p,q) = -\sum_{i=1}^{n} p(x_i) \log\ q(x_i))$$

In game 2, $P(red) = 1$ and no questions need to be asked. However, game 1's strategy was asking 1 question each. Therefore, the expected number of questions you ask would be different.

$$E(number\ of\ questions) = 1+1+1+1 = 4 = -1 \times (\log_2(\frac{1}{2})^4) = Cross\ Entropy$$

3.2.3 KL Divergence

KL Divergence measures the difference between using the true strategy and false strategy to play the game. Kl Divergence for previous game is $H(p_2, p_1) - H(p_2) = 4 - 0 = 4$. The general model of KL Divergence is:

$$-\sum_{i=1}^{n} (p(x_i) \log\ q(x_i)) + \sum_{i=1}^{n} (p(x_i) \log\ p(x_i))$$

$$= \sum_{i=1}^{n} \left(p(x_i) \Big(\log p(x_i) - \log q(x_i) \Big) \right)$$

Minimising KL Divergence is equal to picking up a value of $q(x_i)$ from hypothesis set which gives the smallest E_{in}. $p(x_i)$ is unaffected during this optimisation; therefore, we can just ignore it. This process of minimising KL Divergence is also called Maximising Likelihood Estimation (MLE).

$$Min\ \frac{1}{N} \sum_{i=1}^{n} \Big(\log p(x_i) - \log q(x_i) \Big)$$

$$= Min\ \ -\sum_{i=1}^{n} \log q(x_i)$$

$$= Max\ \sum_{i=1}^{n} \log q(x_i)$$

References

En.wikipedia.org. (2019). Kullback–Leibler divergence. [online] [Accessed 17 Mar. 2019.]

What is Entropy?. (2017). [video] Directed by J. phillips. Youtube.

Shannon Entropy and Information Gain. (2017). [video] Directed by L. Shannon. Youtube.