In the last class, we discussed the business understanding of the case, while in this class, we focused on more modelling techniques.

The statistical model we discussed today is classification and regression, which reflect discrete and continuous ‘y’, respectively. But it should be case by case, some methods with continuous outcomes may use in the split and make the discrete result (like linear regression), and when tree induction divides the area tiny enough, it can regard as a continuous outcome. When it comes to the real problem, like the case we discussed in class, we not only can produce the number of bikes and whether we should hire trailers or not but also think about calculating the probability of the hire. The objective functions among different methods are different, but the similar opinion is to min the loss. We use SSE to compute the linear regression and mention hinge loss when it comes to linear SVM.

Then there are two types of costs we must face. When we know the cost, under the function and calculation, we can compute a margin probability and use this to compare with our solved one to make the decision. Suppose the message is not known to us. In that case, we can run several methods (with specific variables we choose and translate) and find which suit the most appropriate accuracy/false-positive/false-negative (depending on what we can’t bear most).

Also, the last question the professor asked before class was whether to use the decision tree should pay attention because it may cause an over-fitting problem.