Thinking1 CTR数据中的类别数据处理，编码方式有哪些，区别是什么

答： "onehot encode，将feature编码成01的组合方式，代价是特征里类别多的话会比较长。好处是没有对值的比较。

Label encode，将feature编码成数字，特征类别有多少就编码多少种数字，好处是比较节约空间。"

Thinking2 对于时间类型数据，处理方法有哪些

答： "时间序列模型：机器学习的方法主要是ARIMA。深度的方式是LSTM。

对于数据的处理我们从时间的角度可以分为年，季度，月。必要的情况下可以再细分。对特殊时间做特殊处理，比如节假日等。"

Thinking3 你是如何理解CTR预估中的特征组合的，请举例说明

答： 人工特征组合需要对行业有比较深的洞察，DeepFM，wide&deep，可以帮助我们解决部分特征组合问题，DCN和xDeepFM更进一步。

Thinking4： DCN和xDeepFM都可以进行自动特征组合，有何区别

答： "DCN的原理：

将特征向量embedding后作为原始输入层x0，DCN有两部分组成，Cross Network和Deep Network。在Cross Network里，计算公式为y=X0\*XlT\*W+B+Xl,其中XIT是上一层的输出的转置，加上了Bias和上一次的输出XI，也就是说每一次的计算结果都是带有之前所有的计算组合。

xDeepFM原理：

将特征向量embedding后作为原始输入层X0，xDeepFM有两部分组成，CIN和Deep Network。在CIN里，学习高阶的显示feature。CIN每层的计算是一个D\*m的矩阵，每次计算的D不变，每次计算X0和上次的输出做组合，在组合后m会增加。

DCN和xDeepFM的区别：

DCN是原子组合，每个feature在onehot encode，embedding后的原始特征向量x0与x'的组合是混在一起组合的，ex：x0=[0110001….],x'=[10001100….]，他们不对原始的特征做区分，直接相乘。每次计算都是带着上次的组合结果。

xDeepFM是VectorWise，每次计算是按照feature field去组合计算，不会破坏feature的结构。每次计算不带着上次组合的结果，最后要做sum pooling来将之前的结果concat。"

Thinking5： 今天讲解的特征组合只是特征工程中的一部分，你理解的特征工程都包括哪些，不防做个思维导图

think5见附件