AutoDis: Automatic Discretization for Embedding Numerical Features in CTR Prediction

Abstract

计算广告的ctr预估模型的特征主要分为两类：categorical特征和numerical特征，两种特征都需要embedding化才能供深度模型使用。常用的numerical特征embedding化有两种方法：Normalization和Discretization。

Normalization方法是每类特征共享一个embedding，Discretization是将特征离散化后再转化成embedding。前者表达能力有限，后者由于离散化的规则不能和模型一起优化导致效果没有保证。本文提出了AutoDis方法，以end-to-end的方式优化离散化规则和ctr模型，该方法为每个numerical类特征引入embedding集合来建模特征内部的关系，然后提出了自动化的离散化和聚合方法来捕捉特征与embedding的相关关系。

Motivation

现有numerical特征的embedding化方法

1. **Categorization**：给每个特征取值都分配一个embedding，独立优化，只适用于取之较少的特征
2. **Normalization**：特征共享一个embedding，特征取值和embedding的乘积作为最终embedding
3. **Discretization**:将特征取值离散化，分为多个桶，然后为每个桶分配一个embedding

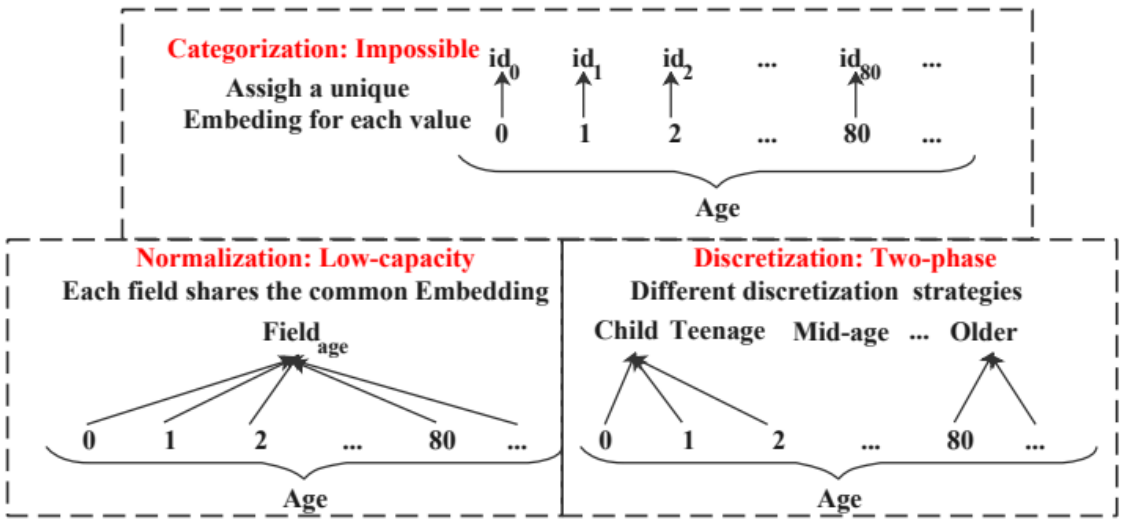

Figure 1: Existing embedding methods for numerical features.

Discretization方法的限制：TPP（Two-Phase Problem）、SBD（Similar value But Dis-similar embedding）、DBS(Dis-similar value But Same embedding)

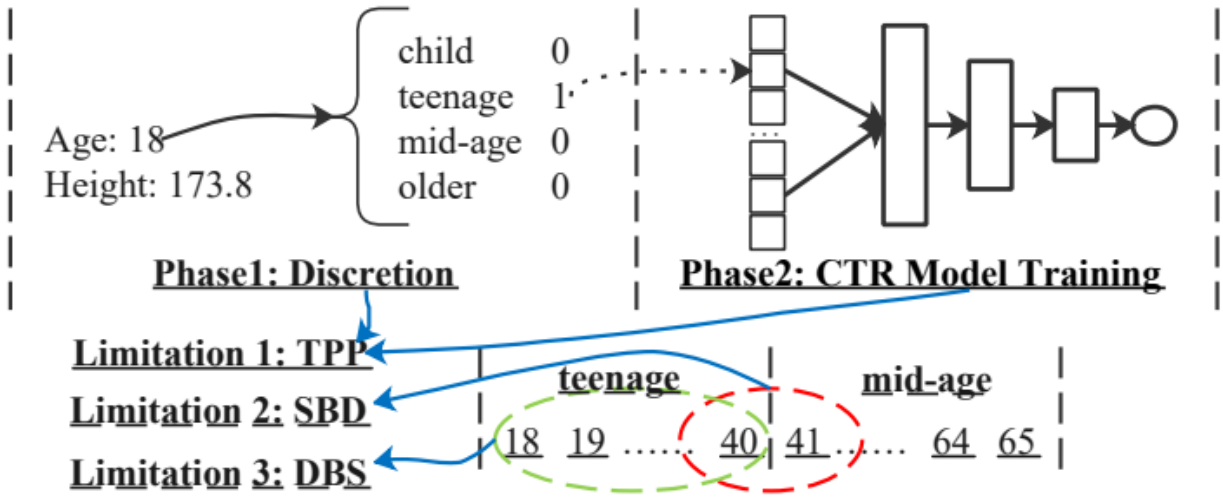

Figure 2: Limitations of existing discretization approaches.

AutoDis

基于现有方法的缺点，本文提出了AutoDis能够end-to-end的学习numerical特征的embedding，能直接应用在现有的深度模型当中。

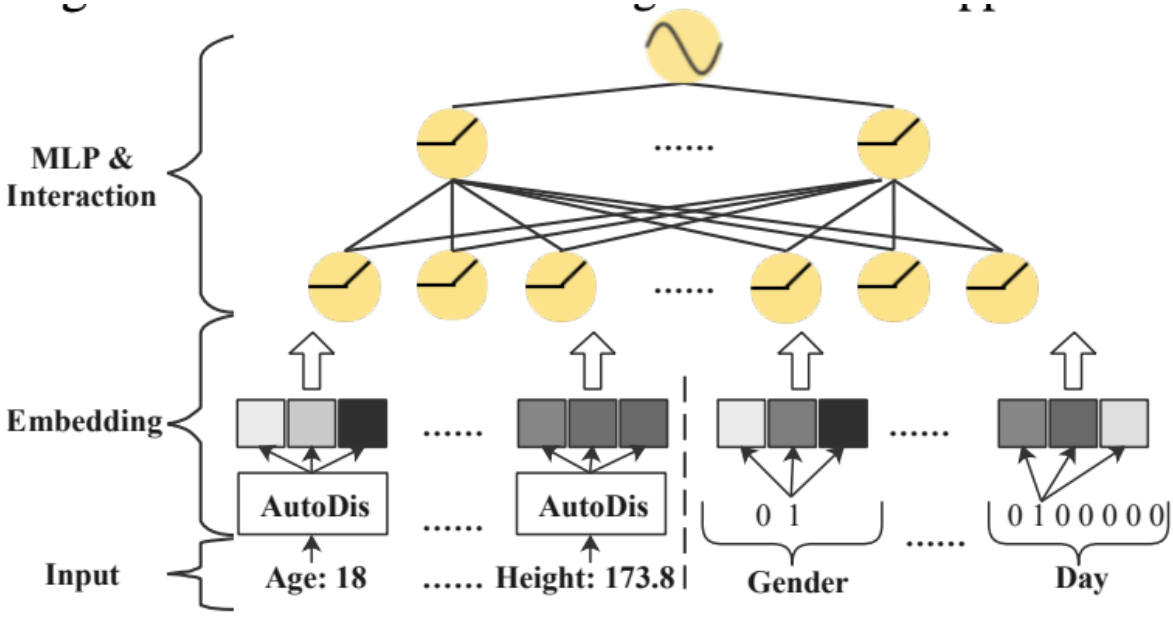

Figure 3: AutoDis works as a component compatible to the existing deep CTR models.

AutoDis为每个numerical特征(如age,gender等)设计了一组 meta embedding $ME_{n_j} \in R^{H_j \times d}$，然后利用一个参数向量$w_{n_j} \in R^{H_j}$ 产生每个特征取值x_{n_j}对应ME_{n_j}的score

$$\hat{x}^h_{n_j} = w^h_{n_j} \cdot x_{n_j}$$

$$h \in [1, H_j]$$

$$p_{n_j}^h = \frac{e^{\frac{1}{\tau}\hat{x}_{n_j}^h}}{\sum_{l=1}^{H_j} e^{\frac{1}{\tau}\hat{x}_{n_j}^l}}$$

然后根据计算出来的分数p聚合ME，例如Max-Pooling, Top-K-Sum, Weighted-Average等

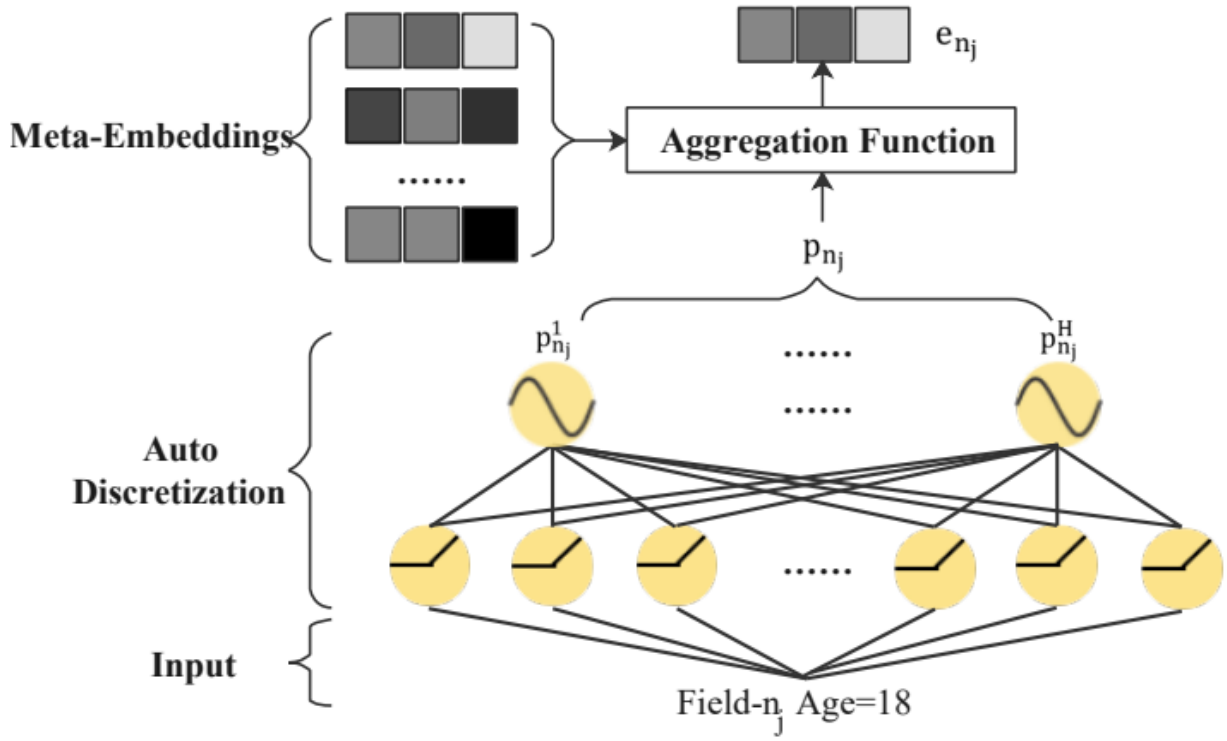

Figure 4: AutoDis Framework.

Experiments

Table 2: The overall performance comparison. Boldface denotes the highest score and underline indicates the best result of the baselines. $\star$ represents significance level p-value < 0.05 of comparing AutoDis with the best baselines.

	Criteo		AutoML		Industrial	
	AUC	LogLoss	AUC	LogLoss	AUC	LogLoss
DeepFM-Norm	0.8107	0.4412	0.7523	0.1899	0.7248	0.1369
DeepFM-EDD	0.8125	0.4399	<u>0.7545</u>	<u>0.1898</u>	0.7251	0.1371
DeepFM-LD	<u>0.8138</u>	<u>0.4388</u>	0.7527	0.1899	<u>0.7265</u>	<u>0.1368</u>
DeepFM-TD	0.8130	0.4392	0.7531	0.1899	0.7262	0.1369
DeepFM-AutoDis	**0.8149***	**0.4372***	**0.7556***	**0.1892***	**0.7277***	**0.1367***
% Improv.	0.14%	0.36%	0.15%	0.32%	0.17%	0.07%

Table 3: Compatibility Study of AutoDis. Normalization method is performed as the baseline to compare with AutoDis.

	Criteo		AutoML		Industrial	
	AUC	LogLoss	AUC	LogLoss	AUC	LogLoss
FNN	0.8059	0.4456	0.7383	0.1926	0.7271	0.1367
FNN-AutoDis	0.8091*	0.4426*	0.7448*	0.1911*	0.7286*	0.1365*
Wide&Deep	0.8097	0.4419	0.7407	0.1923	0.7275	0.1366
Wide&Deep-AutoDis	0.8121*	0.4390*	0.7442*	0.1918*	0.7287*	0.1365*
DeepFM	0.8108	0.4411	0.7525	0.1898	0.7262	0.1369
DeepFM-AutoDis	0.8149*	0.4372*	0.7556*	0.1892*	0.7277*	0.1367*
DCN	0.8091	0.4425	0.7489	0.1909	0.7262	0.1369
DCN-AutoDis	0.8128*	0.4397*	0.7508*	0.1903*	0.7281*	0.1366*
IPNN	0.8101	0.4415	0.7519	0.1896	0.7269	0.1366
IPNN-AutoDis	0.8135*	0.4384*	0.7541*	0.1894*	0.7283*	0.1365*