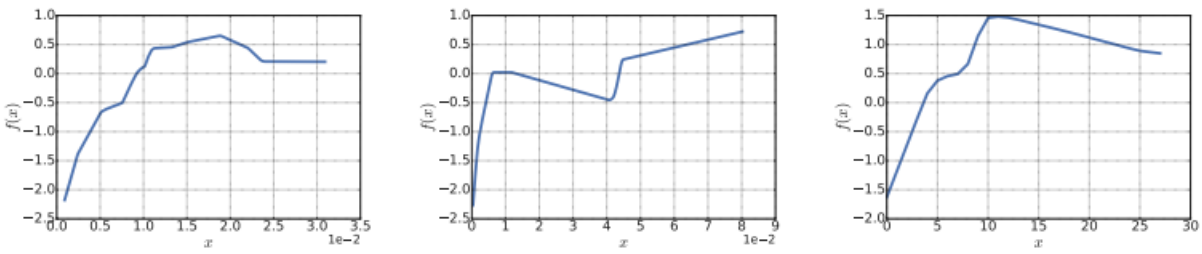
Interpretable Learning-to-Rank with Generalized Additive Models

本文试图获得模型自身的可解释性(**ante-hoc**)，而不是从一个black-box中去获得**post-hoc**可解释性。这是因为对于一些high-stake的问题，例如医疗、法律等对人们生活有着重大影响的问题，使用神经网络的可解释性太差，所以以往大家都只能用传统的机器学习模型，至少这样可解释性能够得到保障，可以获得**每个feature**的贡献程度。而神经网络由于做了隐式的特征交叉，我们并不能看到每个神经元都发生了什么，所以可解释性很差。（不是因为我们不会用深度模型，而是深度模型是个黑盒！）

那么，有没有办法利用好神经网络的非线性性，同时保持较好的可解释性呢?

广义加法模型(GAM)是由多个子模型(sub-model)构成，每个子模型**只**以一个feature x_j作为输入，输出sub-score $f_j(x_j)$. 由于每个子模型的输入没有交互，所以该模型是完全不考虑特征交叉的（当然我们可以手动构造交叉特征来缓解这个问题）。这是GAM模型的缺点，也是带来其良好可解释性的关键。所以，GAM比传统的线性模型要更灵活，也引入了非线性；但是比更复杂的网络可解释性要好得多。

这样，每个子模型由于只包含一个feature，那么这个feature的变化对于sub-score的影响是可解释的 -- 我们可以把这个DNN蒸馏成一个简单的线性模型；或者直接画出图来，表示feature的变化对sub-score的影响，即f_j函数。like this:



(d) Feature 7 (e) Feature 11 (f) Feature 15

Ranking GAM

业务场景：根据query和context，对item进行排序。

- query: 如"hotel"
- item features: 如relevance score, distance
- context features: 如device type

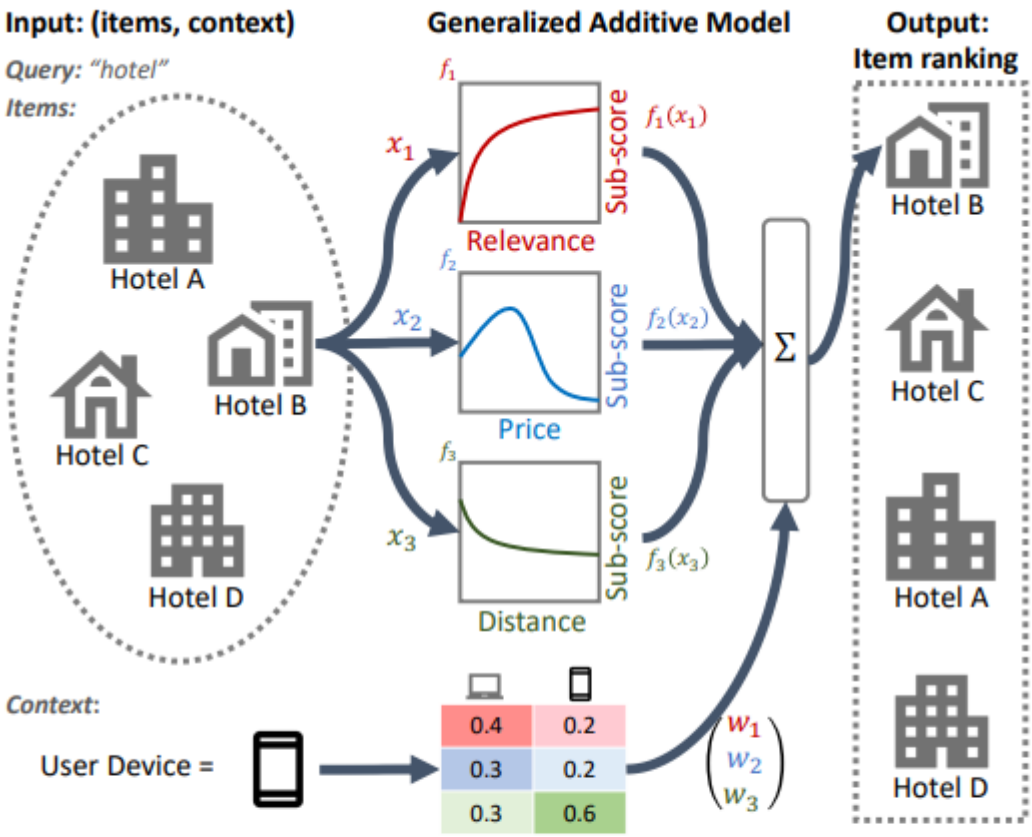


Figure 1: An example of a ranking GAM for local search. For each input feature x_j (*e.g.* price, distance), a sub-model produces a sub-score $f_j(x_j)$. Context features (*e.g.* user device type) can be utilized to derive importance weights of sub-models. The ranking score of each item is a weighted sum of sub-scores. The output is a ranked list of items sorted by their ranking scores.

对item 的每个feature，都用一个sub-model来得到sub-score，如$f_1(x_1), \ldots f_n(x_n)$. 这些item feature虽然没有和context做交互，但是可以是**context dependent**的。比如用BM25作为relevance score，就是依赖于context(query)和document(item)的相关性的。

同时，我们把context feature作为注意力机制的"裁判"（重要特征做裁判的思想），用它得到分配给每个sub-score的权重。例如，当用户搜索"hotel"的时候，可能distance特征更重要；搜索"convention center"的时候，可能relevance更重要 -- 所以想到用context作为注意力机制的selector。那么，为什么不能把context特征也像item特征一样，每个feature输入一个sub-model，然后输出sub-score，最后把这些sub-score加起来呢？这是因为在Learning-to-Rank问题中，我们关心的是两个item之间的**偏序关系**，而context信息对于一个query中的所有item都是**相同的**，如果只是简单的把context特征得分加起来的话，这个context得分对于所有item都是一样的！那么加不加这个context得分都不会影响偏序关系。

这个注意力机制也带来了良好的可解释性 -- 我们知道了哪个特征更加重要。如下面这个热力图表示以context feature "region"作为裁判，分配给不同item feature的注意力权重：

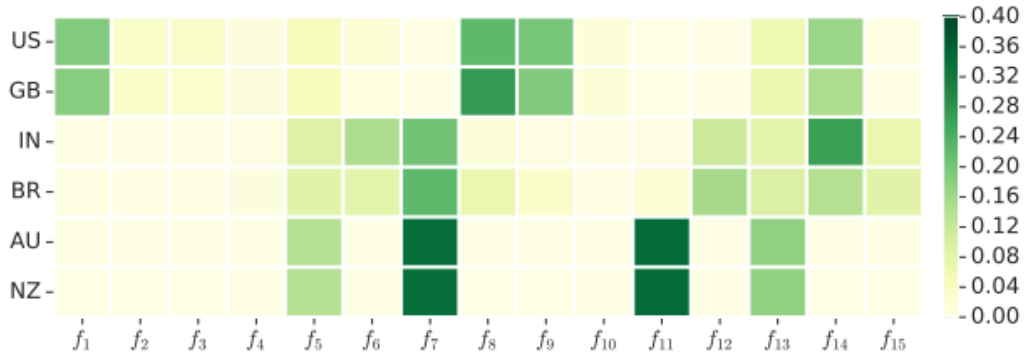


Figure 6: A case study of context feature "region". Each row corresponds to a specific context feature value, where the j-th column corresponds to the derived importance weights on item feature sub-model $f_j(x_j)$.

可以看到，类似的国家（美国&英国，印度&巴西，澳大利亚&新西兰）分配给item feature的权重也是类似的。

整个模型的结构图如下：

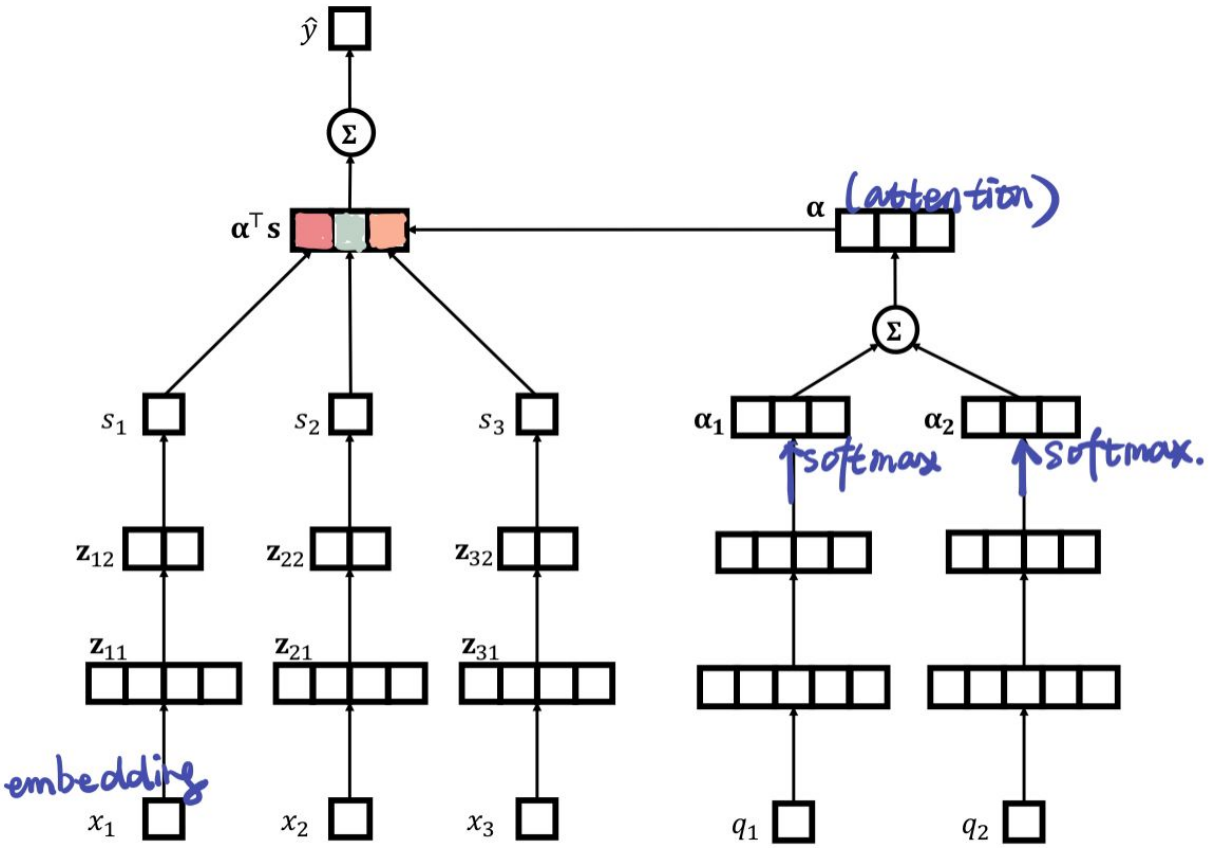


Figure 3: A graphical illustration of a context-present neural ranking GAM.

x_1, x_2, x_3是item feature；q_1, q_2是context feature，作为attention的selector。最后按照$\hat{y}$得分排序。