特征工程需要干什么?

原创 秋枫学习笔记 秋枫学习笔记 2022-05-06 08:34

收录于合集
#特征工程
2个

关注我们，一起学习~

秋枫学习笔记
分享数据挖掘，机器学习，推荐系统等知识和学习笔记
141篇原创内容
公众号

最近阅读了《特征工程入门和实践》这本书，对其中的内容结合笔者的现有知识进行一定的总结，看完之后给我的感觉是这本书比较适合入门，里面的内容相对简单，并且有代码讲解，适合新人入门。对于一些进阶的方法介绍较少，并且有些内容不是特别全面。ok，接下来就进入正题吧。

特征理解

首先分析得到的数据是**结构化数据还是非结构化数据**，通常我们分析的是结构化数据，即表格形式的；对于非结构化数据，需要对其进行清洗和组织。

得到结构化数据后，分析哪些是**定量特征**，哪些是**定性特征**。所谓定量特征，通常也称之为数值型特征，就是可以用数值衡量的，比如身高，体重，温度等；定性特征通常也称之为类别型特征，如性别，职级等。需要注意的是，并不是带有数字的就一定是定量特征，比如性别，可以表示为男：0，女：1，这并不代表他就是定量特征。分析好定量还是定性后，对特征进行简单处理，例如身高对应的数据可能是172cm，这需要将该特征中的“cm”都去掉。

对既有的定量和定性的特征可以继续进行划分，对不同等级的数据可以采用不同的操作：

- **定类等级**：类别型的特征，如血型，性别等，可以对不同值进行计数，找众数等，但无法计算均值等统计信息
- **定序等级**：是类别型特征，但是包含顺序含义在里面，比如职级，评分1星，2星等，这种可以计算中位数，百分位数等
- **定距等级**：数值型特征，可以进行加减法，求均值，标准差，如考试分数，平均分之类的
- **定比等级**：数值型特征，可以进行乘除法，这里和定距等级的区别在于，这里的乘除是有意义的，比如分数的乘除，一个人的分数是另一个人分数的两倍貌似没啥意义。但是比如像工资，浓度等可能就有意义了。

特征增强

特征增强部分主要是对特征中缺失值的识别和处理，对特征进行归一化标准化以除去量纲的影响。

缺失值

实际问题中通常会遇到一些特征的值的缺失的，而这些缺失值可能对模型产生一定的影响，因此需要做出相应的处理。**首先**识别缺失值，最简单直白的就是看特征有没有为空的或者None，NULL的；**其次**是看有没有特别奇怪的，不符合常理的，比如身高特征列中有很多0，人不可能身高为0，因此这里可能是在前置工作中，已经被处理过的缺失值，开发人员将确实的身高默认填充为0，这种需要认为识别，观察数据是否存在一些特殊的值。

缺失值处理的**最直接的方法就是删除含有缺失值的数据**，这种简单暴利的方法通常无法达到很好的效果，

- 一方面，一条数据有很多特征，只要存在缺失值就删除，那么可能会删除很多数据，导致可用于训练的数据很少，没有足够的数据进行训练，导致模型性能欠佳；
- 另一方面，这种方式无法用应对线上传来含有缺失值的数据进行预测的情况；

另一种方式是对缺失值进行填充，填充的方式有很多种，

- 比如前文所述的用一个特殊值填充，如身高0cm；
- 对于定量类型的，也可以用统计量，整个训练集的均值，中位数进行填充，这种方法要用训练集的均值填充训练集和测试集的缺失值，方式数据泄露和穿越；
- 对于定性类型的，即类别型，可以采用出现最多次的进行填充，
- 或者用另一个模型结合已有数据进行预测得到；
- 用一些能处理缺失值的模型，比如神经网络，决策树，xgboost等这些模型本身对缺失值是鲁棒的，可以较好的应对还有缺失值的数据；

note：不同的方法适应不同的场景，比如缺失值较少的情况下，可能删除数据会比较划算；而对于缺失值较多时，采用填充或者用更鲁棒的模型更划算。对于以上没有注明定量还是定性的就都可以使用。

标准化和归一化

标准化和归一化主要是用于消除量纲的影响，比如同一条数据中，由于单位/量纲的不同，有的数据可能很小（身高1.72m，这个值是1.72），而有的数值可能很大（体重63000g，这个值是63000）。有些模型是受量纲影响较大的，比如逻辑回归，因此通常需要进行归一化或标准化。

- z-score标准化：$z=\frac{x-\mu}{\sigma}$ μ为特征列的均值，σ为标准差，这种方式对正太分布的数据比较友好；
- min-max归一化：$m=\frac{x-x_{min}}{x_{max}-x_{min}}$，这种方式对均匀分布的数据比较友好，容易受异常值或离群值的影响；
- 均值归一化：$m=\frac{x-\mu}{x_{max}-x_{min}}$，优缺点和min-max类似。
- 行归一化：前面都是对特征列进行标准化或归一化，这里是对一条数据包含的特征行进行约束，使得每一行特征的L2范数都是1，可以理解为向量长度相同，都在同一个超球上。$||x||=\sqrt{(x_1^2+...+x_n^2)}$

实际使用时可以都试试

特征构建

对于定性类型的特征，即类别型的特征，我们需要对其进行编码，比如考试登记是优、良、及格和不及格，这个无法直接送入模型，因此需要进行编码。

这部分可以参考特征为桥梁 | 特征工程中你了解的和不了解的都在这了中的**特征构造**部分。除了上面分享文章中总结的特征构造方法，还可以采用一些方法来构造新的特征，比如多项式特征，可以采用sklearn中的polynomial-features类通过两个特征a,b构造多项式特征[1, a, b, ab, a^2, b^2]。

对于文本特征可以参考特征为桥梁 | 特征工程中你了解的和不了解的都在这了中的**文本的特征工程**部分。

特征选择

给定数据后，不是所有特征都是有用的特征，有些特征是冗余的甚至是有害的。进行特征选择，筛选去除冗余特征和噪声，得到一个更好的特征子集能有效提高模型的性能，并减少训练和预测时间。

之前的总结特征为桥梁 | 特征工程中你了解的和不了解的都在这了中基本涵盖了本书所提到的特征选择方法，这里不在赘述，书中提到了如何选择正确的特征选择方法，具有一定的参考意义，这里总结记录一下。

- 对于类别型特征，使用卡方检验或树模型会比较不错；
- 对于定量特征，使用线性模型（L1，L2正则化）或相关性计算会比较好；
- 对于二元分类，会用模型选择和SVC会比较好；
- 手动选择特征前，进行探索性数据分析，使用邻域知识。

当然这些建议只是在遇到问题是提供一些参考，不可生搬硬套。

特征转换与生成

这部分我写到一起，这两块内容作者主要是介绍降维以对特征进行转换，以及如何生成新的特征。降维部分无外乎PCA，LDA这里不做过多介绍，采用这些降维的方式可以过滤掉一些噪声，从而使得转换后的特征对模型更友好。

对于特征生成，书中介绍了**受限玻尔兹曼机RBM以及伯努利受限玻尔兹曼机（可以直接调用sklearn的方法）**，RBM有点类似于自编码器autoencoder和GAN这类特征生成的方法。RBM是无监督的，由两层神经网络构成，先前向传播得到输出，然后以得到的输出作为输入“反向传播”，即从右往左还原输入，这里的反向传播和梯度反向传播是两种概念。

书中还对文本特征的构建与生成方式（词向量）进行了简单介绍，部分内容一起一些降维方法也都总结在了特征为桥梁 | 特征工程中你了解的和不了解的都在这了

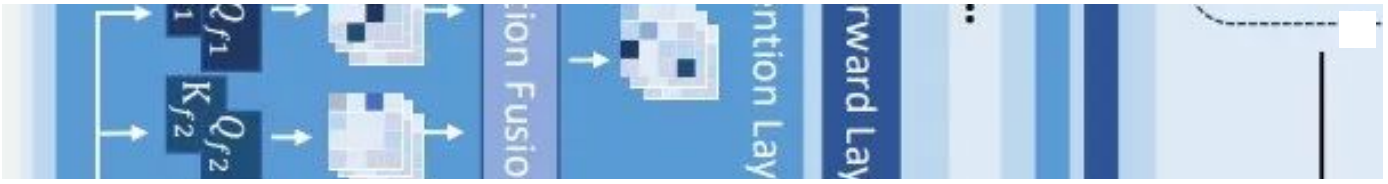
总结

特征工程中不能忽视人的专家知识，要善于利用工具如seaborn和matplotlib库进行可视化和相关性计算，根据不同的阶段，不同的类型采用上面不同的方式。

交流群：点击“联系作者”--备注“研究方向-公司或学校”

欢迎|论文宣传|合作交流

往期推荐

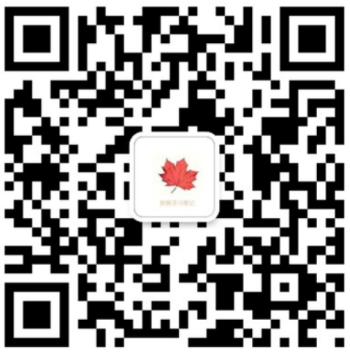


SIGIR'22 序列推荐：对辅助信息解耦后再融合

编程语言大乱斗

SIGIR'22「蚂蚁」CORE：会话推荐中会话表征和商品表征的一致性建模

长按关注，更多精彩



点个 在看 你最好看

发表于上海

喜欢此内容的人还喜欢

一起看会书?

秋枫学习笔记