پیش پردازش داده‌ها

جواد وحدت
vahdatjavad@gmail.com

این مراحل جزء مهمترین گام‌های تحلیل داده محسوب می‌شوند

به طور کلی پیش پردازش داده به دو دسته عمده تقسیم می‌شود که عبارتند از:

- **انتخاب اشیا داده و ویژگی ها (attribute) برای تحلیل**
- **ایجاد کردن یا تغییر دادن ویژگی ها**

۱-یکپارچه سازی داده ها – Data Integration

شامل ترکیب داده ها از چند منبع غیر همگون، انتقال آن ها به یک انبار منسجم و ایجاد یک دید یکتا به داده ها.

برای ترکیب داده‌ها از دو روش کمک میگیریم :

- **اتصال محکم** Tight Coupling – **از این به عنوان مولفه بازیابی اطلاعات استفاده میکنیم.**
- **اتصال سست** Loose Coupling – **در این روش رابطی تهیه می‌گردد که در آن query را از کاربر گرفته و به شیوه‌ای که پایگاه داده مبدا متوجه آن شود، تبدیل می‌کند.**

۲-تجمیع داده – Data Aggregation

به ترکیب متغیرها باهم و ادغام در یک متغیر تجمیع داده گفته می‌شود.

به‌منظور:

- **کاهش داده**
- **تغییر مقیاس (** به عنوان مثال: شهرها می‌توانند به مناطق و استان‌ها و یا کشورها تبدیل شوند و یا روزها به هفته‌ها و ماه‌ها و سال‌ها تبدیل شوند.**)**
- **ثبات بیشتر داده**

۳-نمونه گیری – Sampling

نمونه گیری یک روش معمول برای انتخاب یک زیر مجموعه از داده ها برای تحلیل داده است.

- **نمونه گیری تصادفی عادی** Simple Random Sampling
- **نمونه گیری بدون جایگزینی** Sampling without replacement
- **نمونه گیری با جایگزینی** Sampling with replacement
- **نمونه گیری طبقه بندی شده** Stratified sampling
- **نمونه گیری تصاعدی** Progressive Sampling

۴- کاهش ابعاد – Dimension Reduction

Curse of Dimensionality – **خطای** classify **کردن دیتا افزایش پیدا میکند.**

- **آنالیز مولفه اصلی** – PCA
- **تجزیه مقدارهای منفرد** – SVD

۵-انتخاب زیر مجموعه ای از ویژگی ها – Feature Subset Selection

- ویژگی های غیر مرتبط Irrelevant
- ویژگی های زائد Redundant

سه راه استاندارد برای انتخاب زیرمجموعه ها وجود دارد:

- روش‌های تعبیه شده **Embedded**(به عنوان بخشی از الگوریتم داده کاوی اتفاق می‌افتد ، به طور دقیق خود الگوربتم تصمیم میگرد که از چه ویژگی هایی استفاده کند و چه ویژگی‌هایی استفاده نکند.)
- روش‌های مبتنی بر فیلتر **Filter**(انتخاب ویژگی با استفاده از روش‌هایی مستقل از عملیات داده کاوی قبل از اجرای الگوریتم‌های داده کاوی.)
- روش‌های بسته‌بندی **Wrapper**(از الگوریتم هدف به عنوان یک جعبه سیاه برای پیدا کردن بهترین مجموعه از ویژگی‌ها استفاده می‌کنند.)

۶-خلق ویژگی – Feature Creation

- **استخراج ویژگی**
- **ساخت ویژگی**
- **نگاشت داده ها به فضای جدید**

۷- گسسته سازی و دوگانه سازی – Discretization and Binarization

گسسته سازی Discretization : **گسسته سازی روند تبدیل یک متغیر پیوسته به یک متغیر ترتیبی است.**

دوگانه سازی Binarization : **دوگانه سازی یک ویژگی پیوسته یا طبقه ای** categorical **را به یک یا چند متغیر دودویی تبدیل می‌کند.**

Categorical Value	Integer Value	x_1	x_2	x_3
awful	0	0	0	0
poor	1	0	0	1
OK	2	0	1	0
good	3	0	1	1
great	4	1	0	0

۸-تغییر شکل متغیر ها – Variable Transformation

- **تابعی ایست که تمامی مقادیر ویژگی مورد نظر را به مقادیر جایگزینی تبدیل می‌کند. به نحوی که هر کدام از مقادیر قبلی با یکی از مقادیر جدید مشخص می‌شود.**

- **توابع ساده مثل توان، لگاریتم، قدر مطلق**
- **نرمال‌سازی** – برای تنظیم اختلاف میان ویژگی ها به لحاظ تناوب رخداد، میانگین، واریانس و بازه به کار می‌روند
- **استانداردسازی** – در آمار به اختلاف میانگین‌ها تقسیم بر میانگین واریانس‌ها اشاره دارد.