Categorical Features

Categorical Features

- **Real-valued feature** – number, a quantitative property of an object

- **Categorical feature** – element of an unordered set, a qualitative property of an object

Categorical Features:

- UserID, ItemID, ShopID,
- Category, Region, City, Color, IP, etc.

City ∈ {Beijing, New York, Moscow, …}
UserID ∈ {69384, 34567, 159094, …}
Category ∈ {books, movies, clothing, …}

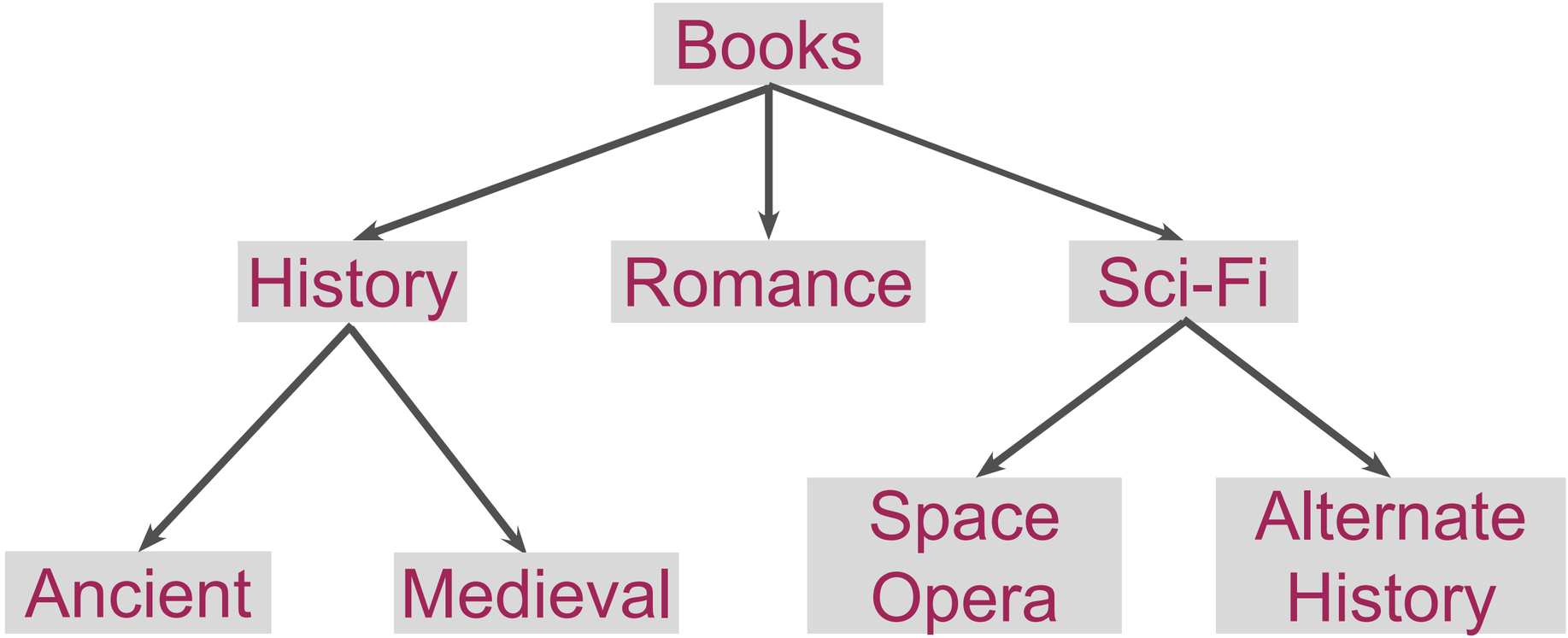
Elements of these sets are called **levels** of a categorical feature

Categorical Features:
UserID, ItemID, ShopID

UserID ∈ {69384, 34567, 159094, …}

It makes no sense to **add**, **multiply**, and **compare** categorical features!

Hierarchy in Categorical Features



One-hot encoding

City ∈ {Beijing, New York, Moscow, Paris}

City	X
Beijing	[1,0,0,0]
New York	[0,1,0,0]
Moscow	[0,0,1,0]
Paris	[0,0,0,1]

Length of ***X*** equals the number of levels

One-hot encoding

Category ∈ {books, movies, clothing, electronics}

Category	X
books	[1,0,0,0]
movies	[0,1,0,0]
clothing	[0,0,1,0]
electronics	[0,0,0,1]

Length of ***X*** equals the number of levels

City ∈ {Beijing, New York, Moscow, Paris}
Category ∈ {book, movies, clothing, electronics}

City,Category	X
Beijing, electronics	[1,0,0,0,0,0,0,1]
New York, movies	[0,1,0,0,0,1,0,0]
Moscow, books	[0,0,1,0,1,0,0,0]
Paris, clothing	[0,0,0,1,0,0,1,0]

Concatenation of vectors of “City” and “Category”

Summary

- **Categorical feature** – element of an unordered set, a qualitative property of an object

- It makes no sense to add, multiply, and compare **categorical features**!

- **Categorical features** can be transformed to a vector of real-valued features using **one-hot-encoding**