# Session 4

# One Hot Encoding

بیشتر مدل های یادگیری ماشین با داده های عددی کار می کنند.

؟ اگر بخواهیم بر روی داده های categorical کار کنیم، چه طور باید این مشکل را حل کنیم ؟

خب ساده است باید داده هایمان را به عدد تبدیل کنیم، اما چه طور ؟

1. Integer Encoding

2. One-Hot Encoding

"Red" : 1

"Yellow" : 2

"Green" : 3

| Color |
|---|
| Red |
| Red |
| Yellow |
| Green |
| Yellow |

Integer Encoding →

| Color |
|---|
| 1 |
| 1 |
| 2 |
| 3 |
| 2 |

ordered relationship ظاهراً که همه چیز به نظر خوب می‌رسه، اما واقعاً اینجوریه؟

اگر ما به هر Category یک عدد بدهیم و حدل ما ارزش آن عدد را در حسابات دخیل کند درحالی که عدد بزرگترشان دهنده‌ی برتری یک Category نسبت به دیگری نیست.

؟: برای رفع این مشکل چه کنیم؟ می‌توانیم از One-Hot Encoding استفاده کنیم.

| Color |
|---|
| Red |
| Red |
| Yellow |
| Green |
| Yellow |

One-Hot Encoding →

| Red | Yellow | Green |
|---|---|---|
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 0 | 1 |
| 0 | 1 | 0 |

## مشکلات One-Hot Encoding چیست؟

1. افزایش تعداد ویژگی‌ها

2. Dummy Variable Trap

| Red | Yellow | Green |
|---|---|---|
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 0 | 1 |
| 0 | 1 | 0 |

| Red | Yellow |
|---|---|
| 1 | 0 |
| 1 | 0 |
| 0 | 1 |
| 0 | 0 |
| 0 | 1 |