Session 05

F1-Score & One-Hot Encoding

Machine Learning | Zahra Amini

Telegram: @zahraamini_ai & Instagram:@zahraamini_ai & LinkedIn: @zahraamini-ai

https://zil.ink/zahraamini

؟: اگر در مسائل classification فقط از accuracy برای ارزیابی مدل استفاده کنیم، آیا می‌توانیم تصمیم بگیریم که عملکرد مدل خوب است؟

Cancer

Acuracy: 95% ⟵ عملکرد مدل؟

$\text{Accuracy} = \frac{\text{تعداد درست‌ها}}{\text{تعداد کل}}$

100 نفر ⟵ 95 سالم، 5 بیمار

مدل ما اگر همیشه "سالم" را پیش‌بینی کند با دقت 95% دارد درست عمل می‌کند. ∴

برای رفع این مشکل باید از Confusion Matrix استفاده کنیم.

y ⟵ True / False

y_Pred ⟵ Postive / Negative

True Positive ⟶ Predict Positive, True

True Negative ⟶ Predict Negative, True

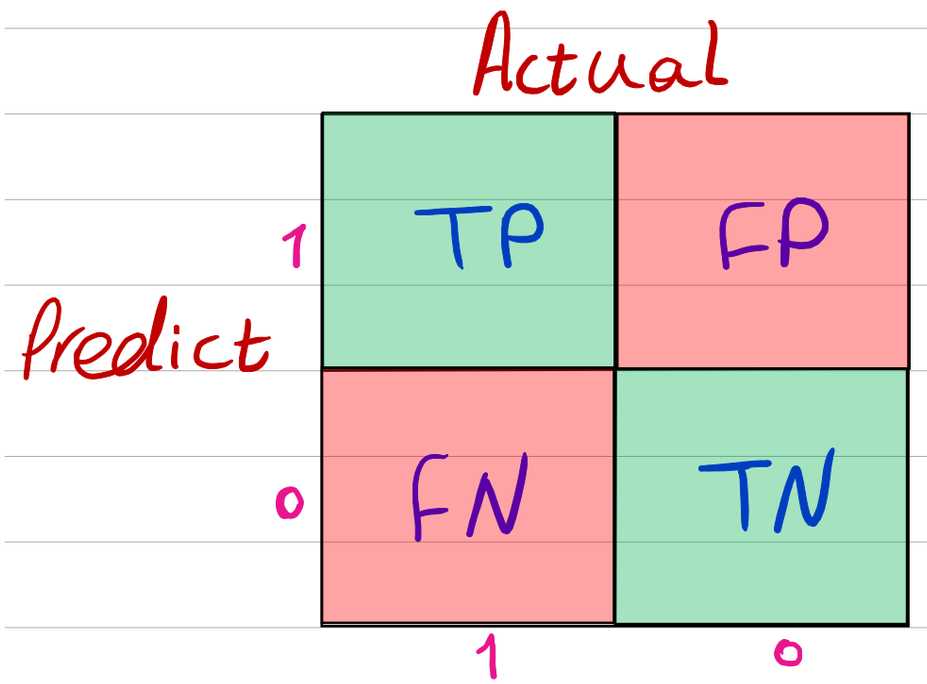
False Positive ⟶ Predict Positive, False

False Negative ⟶ Predict Negative, False

TP TN

FP FN

	y		thr = 0.6	Recall	Precision	Accuracy
TN ←	0	0.5	0			
TP ←	1	0.9	1			
FP ←	0	0.7	1	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{4}{7}$
TP ←	1	0.7	1			
FN ←	1	0.3	0			
TN ←	0	0.4	0			
FN ←	1	0.5	0			



$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Accuracy} = \frac{TP+TN}{\text{Total}}$$

$$\text{F1-Score} = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$$

بیشتر مدل‌های یادگیری ماشین با داده‌های عددی کار می‌کنند.

؟ اگر بخواهیم بر روی داده‌های Categorical کار کنیم، چه طور باید این مشکل را حل کنیم ؟

خب ساده است باید داده‌هایمان 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