Machine Learning : 06048203

Classification Metrics

Classification Performance Metrics

Part One: Confusion Matrix Basics

Classification Metrics

- You’ve probably heard of terms such as “false positive” or “false negative”. As well as metrics like “accuracy”.
- But what do these terms actually mean mathematically?

Classification Metrics

- Imagine we've developed a test or model to detect presence of a virus infection in a person based on some biological feature.
- We could treat this as a Logistic Regression, predicting:
    - 0 - Not Infected (Tests Negative)
    - 1 - Infected (Tests Positive)

Classification Metrics

- It is unlikely our model will perform perfectly. This means there 4 possible outcomes:
    - Infected person tests positive.
    - Healthy person tests negative.

Classification Metrics

- It is unlikely our model will perform perfectly. This means there 4 possible outcomes:
  - Infected person tests positive.
  - Healthy person tests negative.
    - *Note, these are the outcomes we want! But it is unlikely our test is perfect...*

Classification Metrics

- It is unlikely our model will perform perfectly. This means there 4 possible outcomes:
  - Infected person tests positive.
  - Healthy person tests negative.
  - Infected person tests negative.
  - Healthy person tests positive.

Classification Metrics

- Based off these 4 possibilities, there are many error metrics we can calculate.
- First, let’s start by visualizing these four possibilities as a matrix.

Classification Metrics

- Confusion Matrix

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED		
	HEALTHY		

Classification Metrics

- Confusion Matrix

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	TRUE POSITIVE	
	HEALTHY		

Classification Metrics

- Confusion Matrix

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	TRUE POSITIVE	
	HEALTHY		TRUE NEGATIVE

Classification Metrics

- Confusion Matrix

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	TRUE POSITIVE	FALSE POSITIVE
	HEALTHY		TRUE NEGATIVE

Classification Metrics

- Confusion Matrix

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	TRUE POSITIVE	FALSE POSITIVE
	HEALTHY	FALSE NEGATIVE	TRUE NEGATIVE

Classification Metrics

- What is accuracy?

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

- Accuracy:
  - How often is the model correct?

Acc = (TP+TN)/Total

Classification Metrics

- Calculating accuracy:

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

(4+93)/100 = 97% Accuracy

- Accuracy:
  - How often is the model correct?

Acc = (TP+TN)/Total

Classification Metrics

- Is this a good value for accuracy?

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

(4+93)/100 = 97% Accuracy

- Accuracy:
  - How often is the model correct?

Acc = (TP+TN)/Total

Classification Metrics

- The accuracy paradox...

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

(4+93)/100 = 97% Accuracy

- Accuracy:
  - How often is the model correct?

Acc = (TP+TN)/Total

Classification Metrics

- Imagine we **always** report back “healthy”

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

Classification Metrics

- Imagine we **always** report back “healthy”

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	0	0
	HEALTHY	5	95

Classification Metrics

- Imagine we **always** report back “healthy”

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	0	0
	HEALTHY	5	95

(0+95)/100 = 95% Accuracy

- Accuracy:
  - How often is the model correct?

95% accuracy for a model that always returns “healthy”!

Classification Metrics

- You may be thinking, “*The numbers here are arbitrary, we just happen to get good accuracy in this made up case. Real world data would reflect poor accuracy if a model always returned the same result*”.

Classification Metrics

- This is the accuracy paradox!
  - Any classifier dealing with **imbalanced** classes has to confront the issue of the accuracy paradox.
  - **Imbalanced** classes will always result in a distorted accuracy reflecting better performance than what is truly warranted.

Classification Metrics

- **Imbalanced** classes are often found in real world data sets.
  - Medical conditions can affect small portions of the population.
  - Fraud is not common (e.g. Real vs. Fraud credit card usage).

Classification Metrics

- If a class is only a small percentage (**n%**), then a classifier that always predicts the majority class will always have an accuracy of (1-n).
- In our previous example we saw infected were only 5% of the data.
- Allowing the accuracy to be 95%.

Classification Metrics

- This means we shouldn’t solely rely on accuracy as a metric!
- This is where precision,recall, and f1-score will come in.
- Let’s explore these other metrics in the next lecture.

Classification Performance Metrics

Part Two: Precision and Recall

Classification Metrics

- We already know how to calculate accuracy and its associated paradox.
- Let’s explore three more metrics that can help give a clearer picture of performance:
  - Recall (a.k.a. sensitivity)
  - Precision
  - F1-Score

Classification Metrics

- Let’s begin with recall.

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

- Recall:
  - When it actually is a positive case, how often is it correct?

(TP)/Total Actual Positives

Classification Metrics

- Let’s begin with recall.

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

Recall =
(TP)/Total Actual Positives

- Recall:
  - When it actually is a positive case, how often is it correct?

(TP)/Total Actual Positives

Classification Metrics

- Let’s begin with recall.

ACTUAL

PREDICTED	INFECTED	HEALTHY
INFECTED	4	2
HEALTHY	1	93

Recall = (TP)/5

- Recall:
  - When it actually is a positive case, how often is it correct?

(TP)/Total Actual Positives

Classification Metrics

- Let’s begin with recall.

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

Recall = (4)/5

- Recall:
  - When it actually is a positive case, how often is it correct?

(TP)/Total Actual Positives

Classification Metrics

- Let’s begin with recall.

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

Recall = 0.8

- Recall:
  - How many relevant cases are found?

(TP)/Total Actual Positives

Classification Metrics

- What's the recall if we always classify as "healthy"?

ACTUAL

PREDICTED	INFECTED	HEALTHY
INFECTED	0	0
HEALTHY	5	95

Recall =
(TP)/Total Actual Positives

- Recall:
  - How many relevant cases are found?

(TP)/Total Actual Positives

Classification Metrics

- What's the recall if we always classify as "healthy"?

ACTUAL

PREDICTED	INFECTED	HEALTHY
INFECTED	0	0
HEALTHY	5	95

Recall = (0)/5 !

- Recall:
  - How many relevant cases are found?

(TP)/Total Actual Positives

Classification Metrics

- A recall of 0 alerts you the model isn't catching cases!

ACTUAL

PREDICTED	INFECTED	HEALTHY
INFECTED	0	0
HEALTHY	5	95

Recall = (0)/5 !

- Recall:
  - How many relevant cases are found?

(TP)/Total Actual Positives

Classification Metrics

- Now let's explore **precision**.

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

Precision =
(TP)/Total Predicted Positives

- Precision:
  - When prediction is positive, how often is it correct?

(TP)/Total Predicted Positives

Classification Metrics

- Now let's explore **precision**.

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

Precision = (TP)/Total Predicted Positives

- Precision:
  - When prediction is positive, how often is it correct?

(TP)/Total Predicted Positives

Classification Metrics

- Now let's explore **precision**.

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

Precision = (TP)/6

- Precision:
  - When prediction is positive, how often is it correct?

(TP)/Total Predicted Positives

Classification Metrics

- Now let's explore **precision**.

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

Precision = (TP)/6

- Precision:
  - When prediction is positive, how often is it correct?

(TP)/Total Predicted Positives

Classification Metrics

- Now let's explore **precision**.

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

Precision = (4)/6

- Precision:
  - When prediction is positive, how often is it correct?

(TP)/Total Predicted Positives

Classification Metrics

- Now let's explore **precision**.

		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

Precision = 0.666

- Precision:
  - When prediction is positive, how often is it correct?

(TP)/Total Predicted Positives

Classification Metrics

- What's the **precision** if we always classify as "healthy"?

ACTUAL

PREDICTED	INFECTED	HEALTHY
INFECTED	0	0
HEALTHY	5	95

Precision = (TP)/Total Predicted Positives

- Precision:
  - When prediction is positive, how often is it correct?

(TP)/Total Predicted Positives

Classification Metrics

- What's the **precision** if we always classify as "healthy"?

ACTUAL

PREDICTED	INFECTED	HEALTHY
INFECTED	0	0
HEALTHY	5	95

Precision = 0/0

- Precision:
  - When prediction is positive, how often is it correct?

(TP)/Total Predicted Positives

🔹 Precision (ความแม่นยำ)

- คือสัดส่วนของ **สิ่งที่เราทำนายว่าเป็น Positive แล้วถูกต้องจริง ๆ**
- คิดเป็นสูตรว่า:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

- พูดง่าย ๆ คือ "ทำนายว่าใช่ แล้วมันใช่จริงกี่ %"
- เหมาะเวลาเราอยาก **ลด False Positive** (ไม่อยากบอกว่ามีโรค ทั้งที่จริง ๆ ไม่ได้เป็น)

👉 ตัวอย่าง: ตรวจโรค

- ระบบบอกว่า 100 คนเป็นโรค
- จริง ๆ เป็นจริง 90 คน อีก 10 คนไม่เป็น
- Precision = 90 / (90+10) = 0.9

🔹 Recall (ความครอบคลุม)

- คือสัดส่วนของ **สิ่งที่เป็น Positive จริง ๆ ที่เราสามารถทำนายเจอ**
- คิดเป็นสูตรว่า:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

- พูดง่าย ๆ คือ "จากคนที่เป็นจริง ๆ ทั้งหมด เราหาเจอกี่ %"
- เหมาะเวลาเราอยาก **ลด False Negative** (ไม่อยากพลาดสิ่งสำคัญ เช่น พลาดตรวจเจอโรคร้าย)

👉 ตัวอย่าง: ตรวจโรค

- มีคนป่วยจริง ๆ 100 คน
- ระบบเจอ 80 คน แต่พลาดไป 20 คน
- Recall = 80 / (80+20) = 0.8

🔹 สรุปแบบเปรียบเทียบ

- Precision → สนใจ "ความถูกต้อง" ของสิ่งที่บอกว่าใช่
- Recall → สนใจ "ความครบถ้วน" ของการเจอสิ่งที่ใช่

📌 ถ้าจะให้ชัด ๆ ลองคิดว่า:

- **นักสืบสายระแวง (Precision สูง)** → ถ้าไม่แน่จริงจะไม่บอกว่าคือคนร้าย (ลดการใส่ร้ายผิดคน)
- **นักสืบสายกวาดหมด (Recall สูง)** → ใครน่าสงสัยนิดหน่อยก็จับมาหมด (ลดการพลาดจับตัวจริง แต่เสี่ยงจับผิด)

🔹 ตัวอย่าง 1: การตรวจโรคร้ายแรง

- **โจทย์**: ตรวจหามะเร็งจากผล X-ray
- **สิ่งสำคัญ**: ห้ามพลาดผู้ป่วยจริง (False Negative ต้องน้อยที่สุด)
- ดังนั้นเราจะเน้น **Recall สูง**
- ยอมให้มี False Positive บ้าง (คนที่ระบบบอกว่าเป็น แต่จริง ๆ ไม่เป็น) เพราะตรวจซ้ำก็ได้

👉 **สรุป**: กรณีนี้ **Recall สำคัญกว่า Precision**

🔹 ตัวอย่าง 2: การกรองอีเมลขยะ (Spam Filter)

- **โจทย์**: แยกว่าอีเมลไหนเป็น Spam
- **สิ่งสำคัญ**: อย่าพลาดอีเมลงานสำคัญไปอยู่ในถัง Spam (False Positive ต้องน้อย)
- ดังนั้นเราจะเน้น **Precision สูง**
- ยอมให้ Spam บางฉบับหลุดเข้ามา (False Negative) เพราะเราลบเองได้

👉 **สรุป**: กรณีนี้ **Precision สำคัญกว่า Recall**

🔹 ตัวอย่าง 3: ระบบตรวจจับทุจริตบัตรเครดิต

- **โจทย์**: หาธุรกรรมที่น่าสงสัย
- ถ้าเน้น **Precision สูง** → แจ้งเตือนน้อย แต่แม่นยำ (ไม่รบกวนลูกค้าปกติมาก)
- ถ้าเน้น **Recall สูง** → เจอแทบทุกรายการน่าสงสัย แต่ลูกค้าอาจโดนบล็อกธุรกรรมที่ไม่ผิดจริงบ่อย ๆ

👉 ตรงนี้ขึ้นกับ **ธุรกิจ** ว่าอยากป้องกันการพลาด (Recall) หรืออยากลดการแจ้งเตือนเกินจริง (Precision)

📌 **สรุปง่าย ๆ**

- ถ้า **ห้ามพลาดสิ่งสำคัญ** → เน้น Recall
- ถ้า **ห้ามใส่ร้ายผิดคน** → เน้น Precision
- ถ้าอยากบาลานซ์ทั้งคู่ → ใช้ **F1-Score**

Classification Metrics

- Recall and Precision can help illuminate our performance specifically in regards to the relevant or positive case.
- Depending on the model, there is typically a trade-off between precision and recall, which we will explore later on with the ROC curve.

Classification Metrics

- Since precision and recall are related to each other through the numerator (TP), we often also report the F1-Score, which is the harmonic mean of precision and recall.

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

Classification Metrics

- The harmonic mean (instead of the normal mean) allows the entire harmonic mean to go to zero if **either** precision or recall ends up being zero.

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

📊 Evaluation Metrics: F1 & ROC

🔹 F1 Score Variants

1. Macro F1

- คำนวณ F1 ของแต่ละคลาส แล้วนำมาหาค่าเฉลี่ยแบบ **เท่ากันทุกคลาส**
- เหมาะกับ class imbalance ที่ต้องการให้ความสำคัญกับทุกคลาสเท่าๆ กัน

📌 สูตร:

$$F1_{macro} = \frac{1}{N} \sum_{i=1}^{N} F1_i$$

---

2. Micro F1

- รวม True Positive, False Positive, False Negative ของทุกคลาสก่อน แล้วคำนวณ F1
- เหมาะกับกรณีที่คลาสไม่สมดุล และต้องการดู performance โดยรวม

📌 สูตร:

$$F1_{micro} = \frac{2 \cdot TP_{total}}{2 \cdot TP_{total} + FP_{total} + FN_{total}}$$

---

3. Weighted F1

- คำนวณ F1 ของแต่ละคลาส แล้วเฉลี่ยแบบ ถ่วงน้ำหนักตามจำนวนตัวอย่างในแต่ละคลาส

📌 สูตร:

$$F1_{weighted} = \sum_{i=1}^{N} w_i \cdot F1_i \quad , \quad w_i = \frac{n_i}{\sum_j n_j}$$

Classification Performance Metrics

Part Three: ROC Curves (Receiver Operating Characteristic))

Classification Metrics

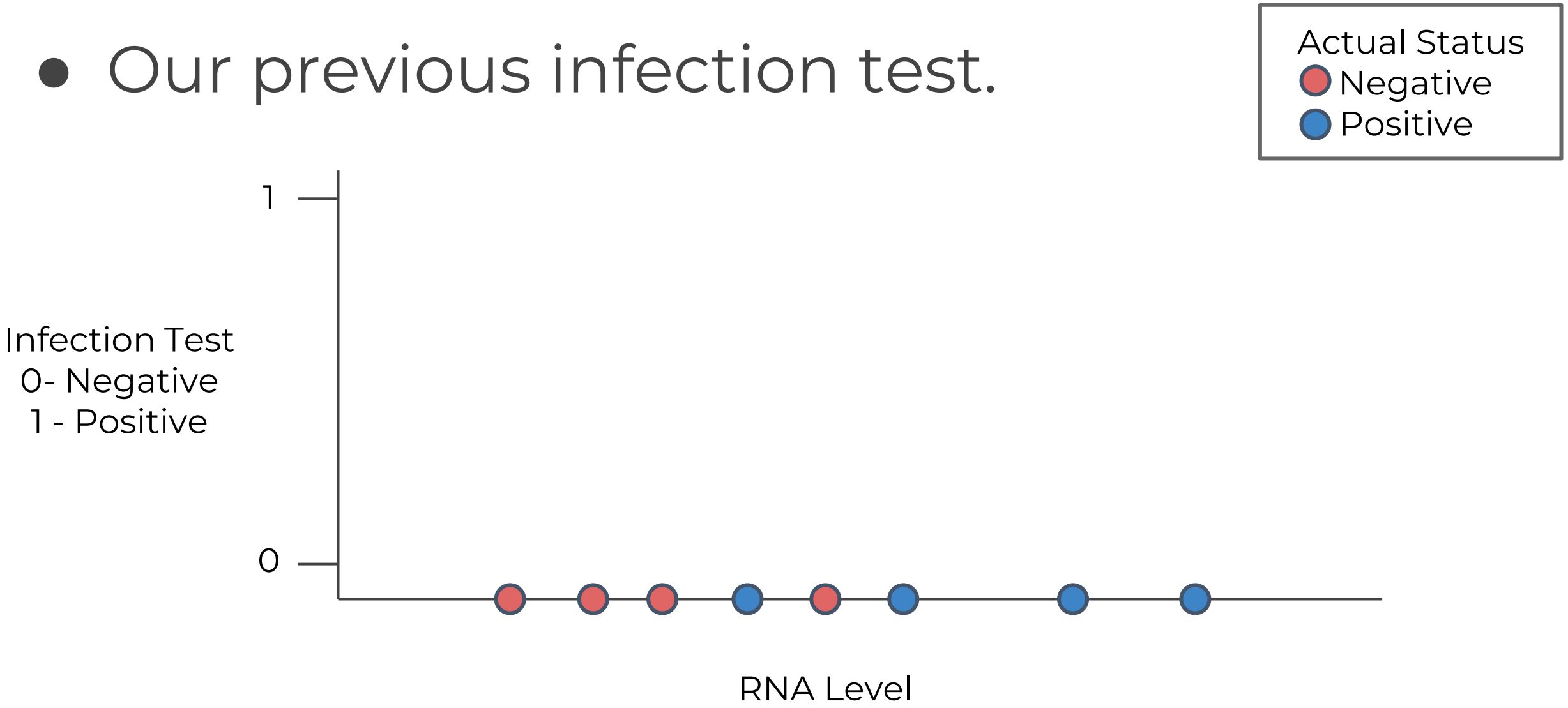
- During World War 2, Radar technology was developed to help detect incoming enemy aircraft.

Classification Metrics

- The technology was so new, the US Army wanted to develop a methodology to evaluate radar operator performance.

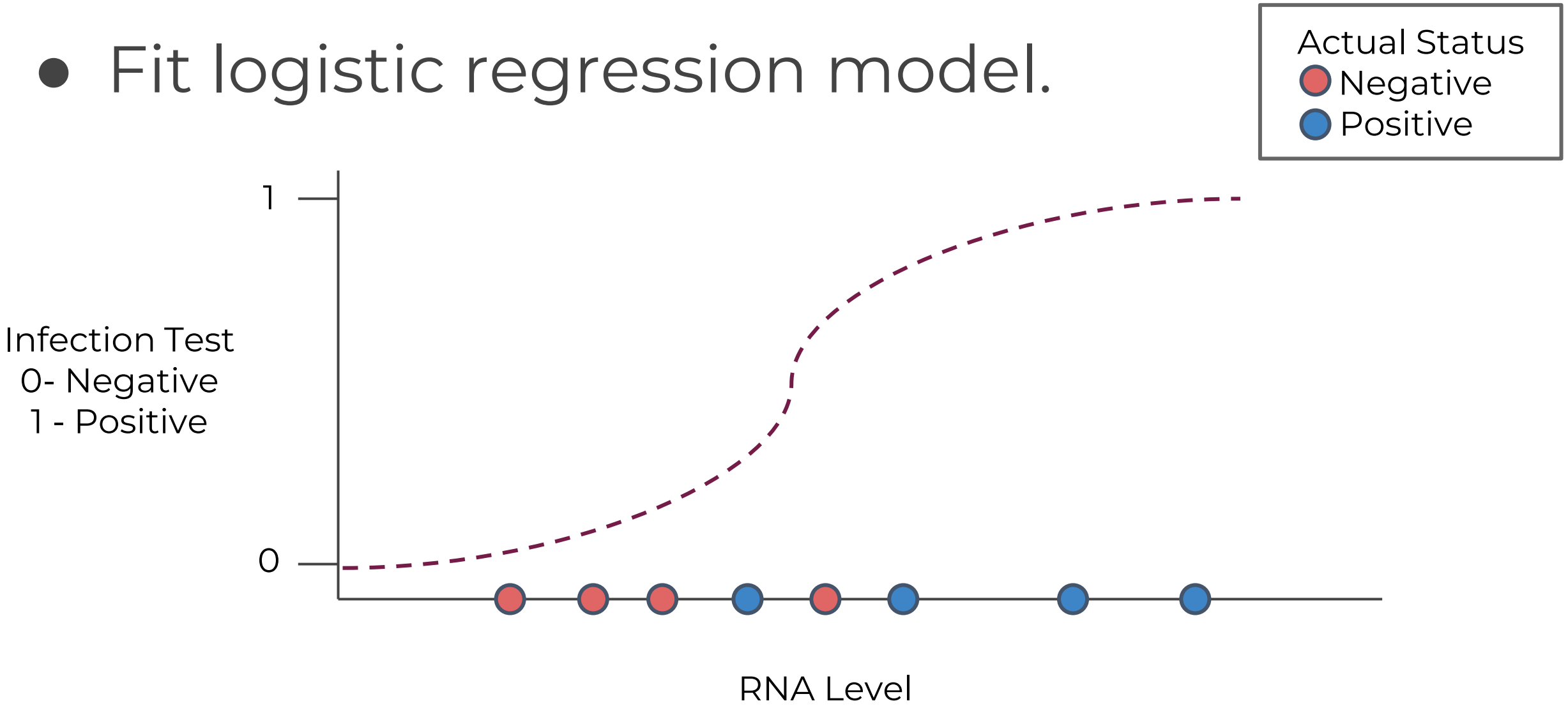
Classification Metrics

- Our previous infection test.



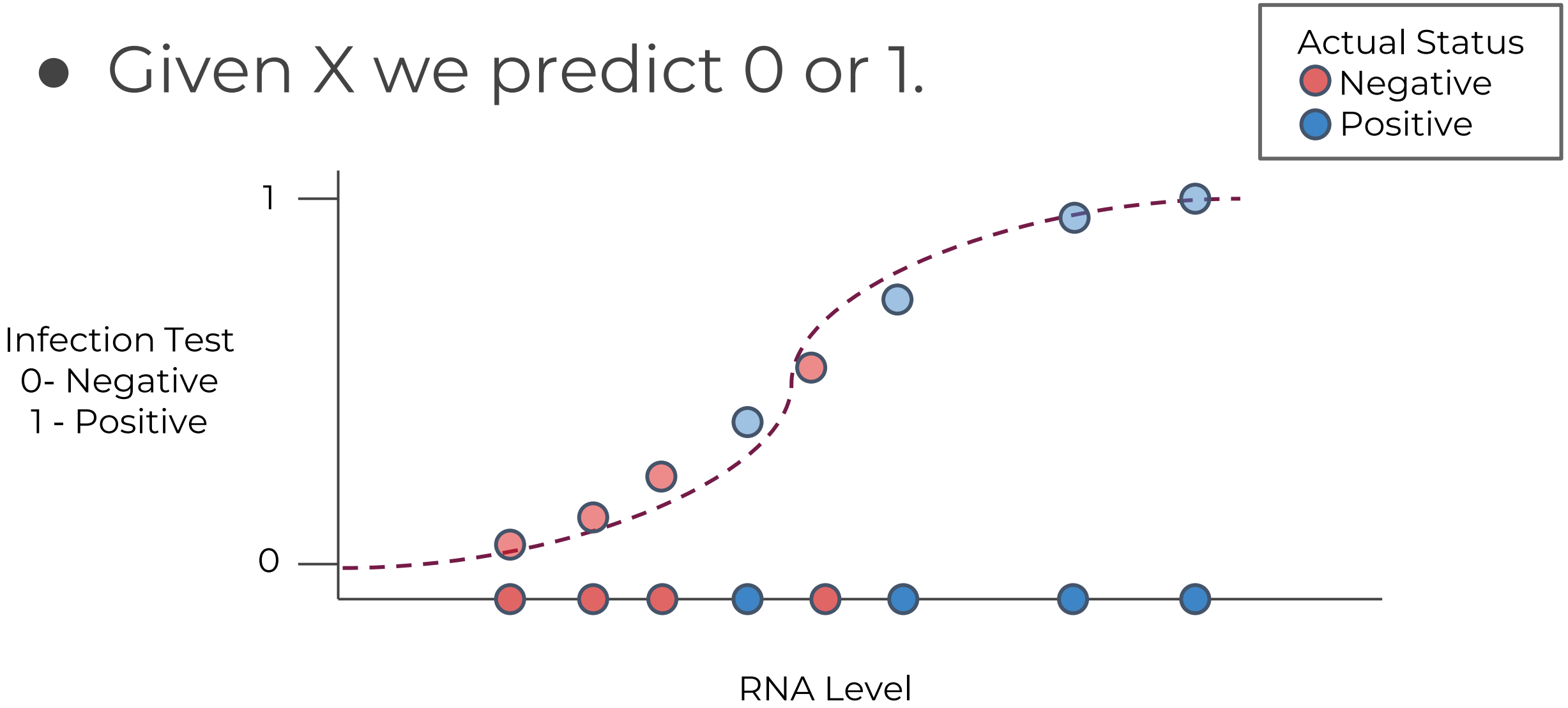
Classification Metrics

- Fit logistic regression model.



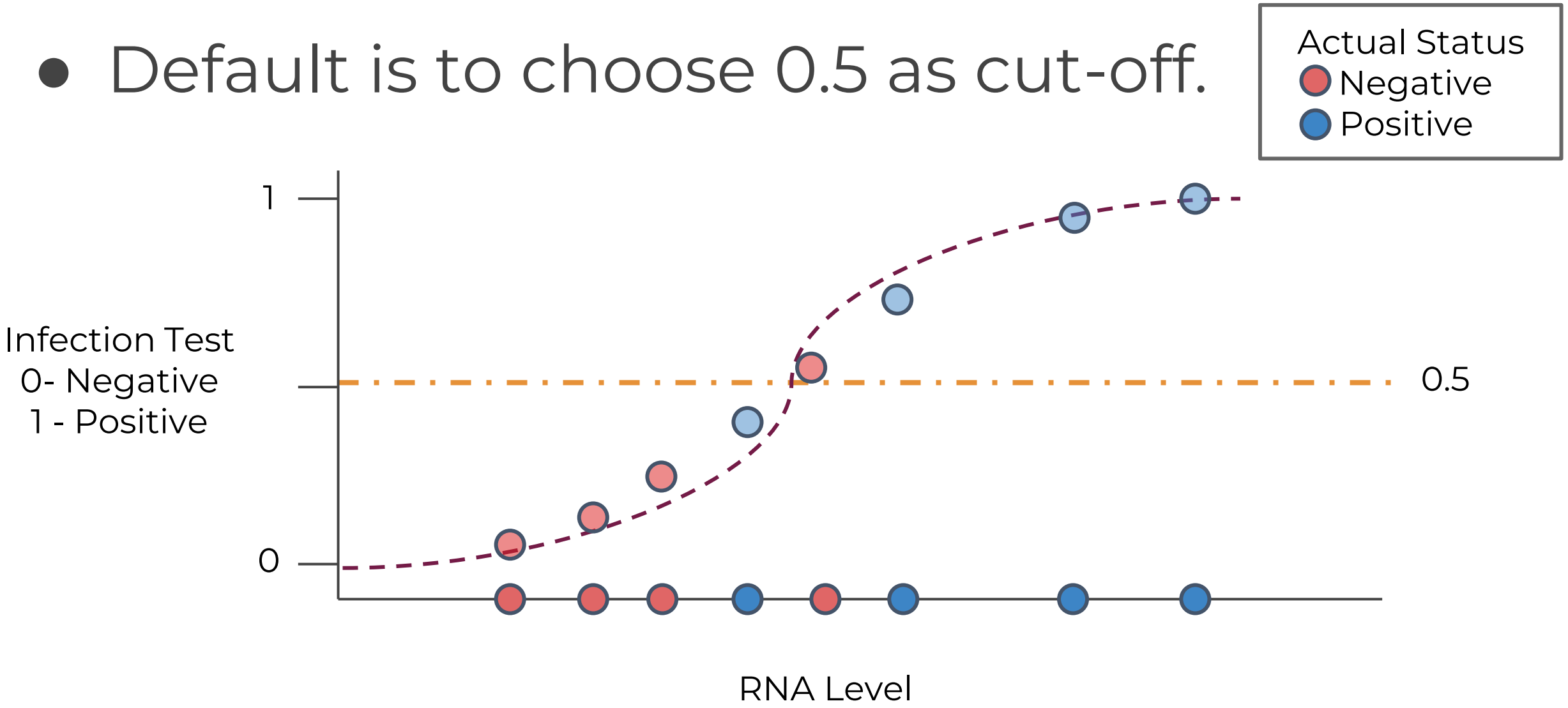
Classification Metrics

- Given X we predict 0 or 1.



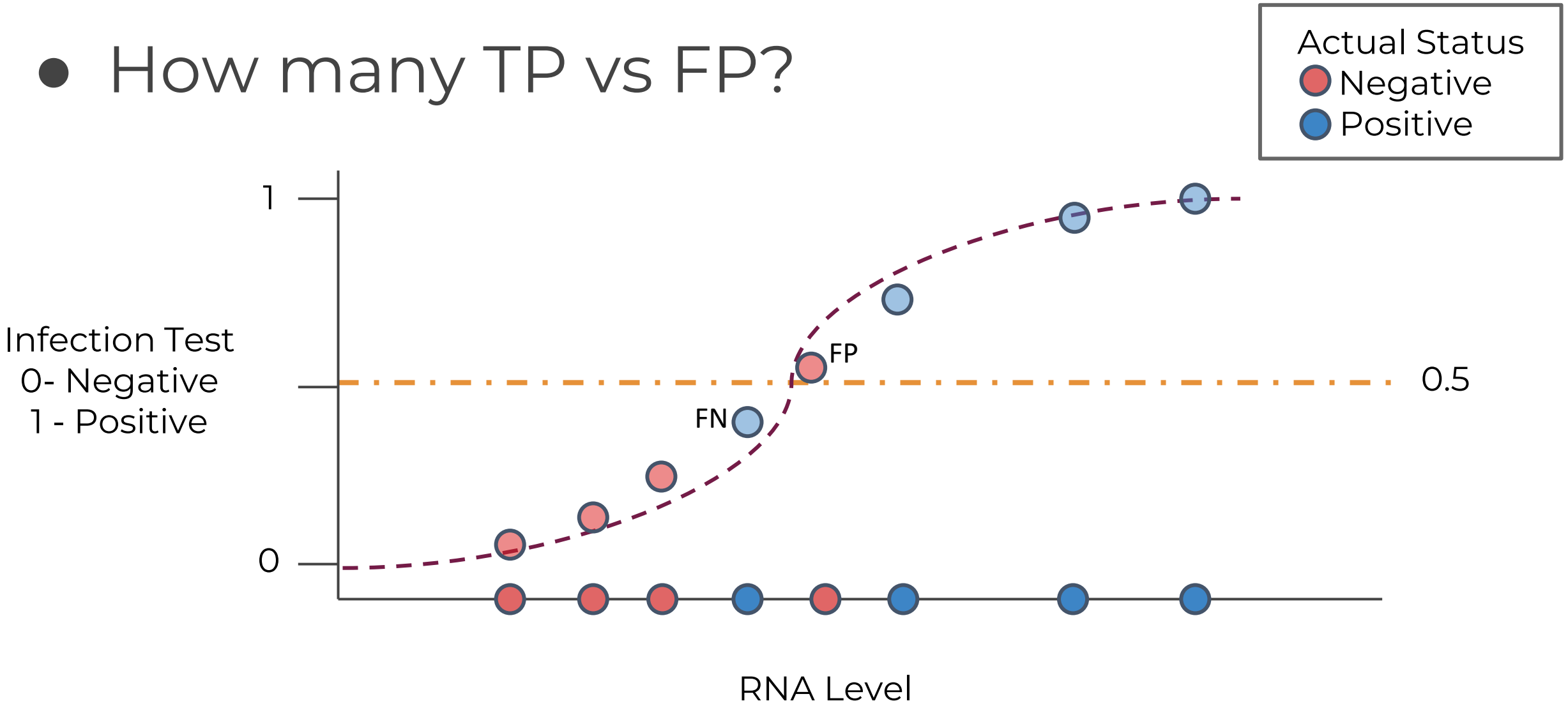
Classification Metrics

- Default is to choose 0.5 as cut-off.



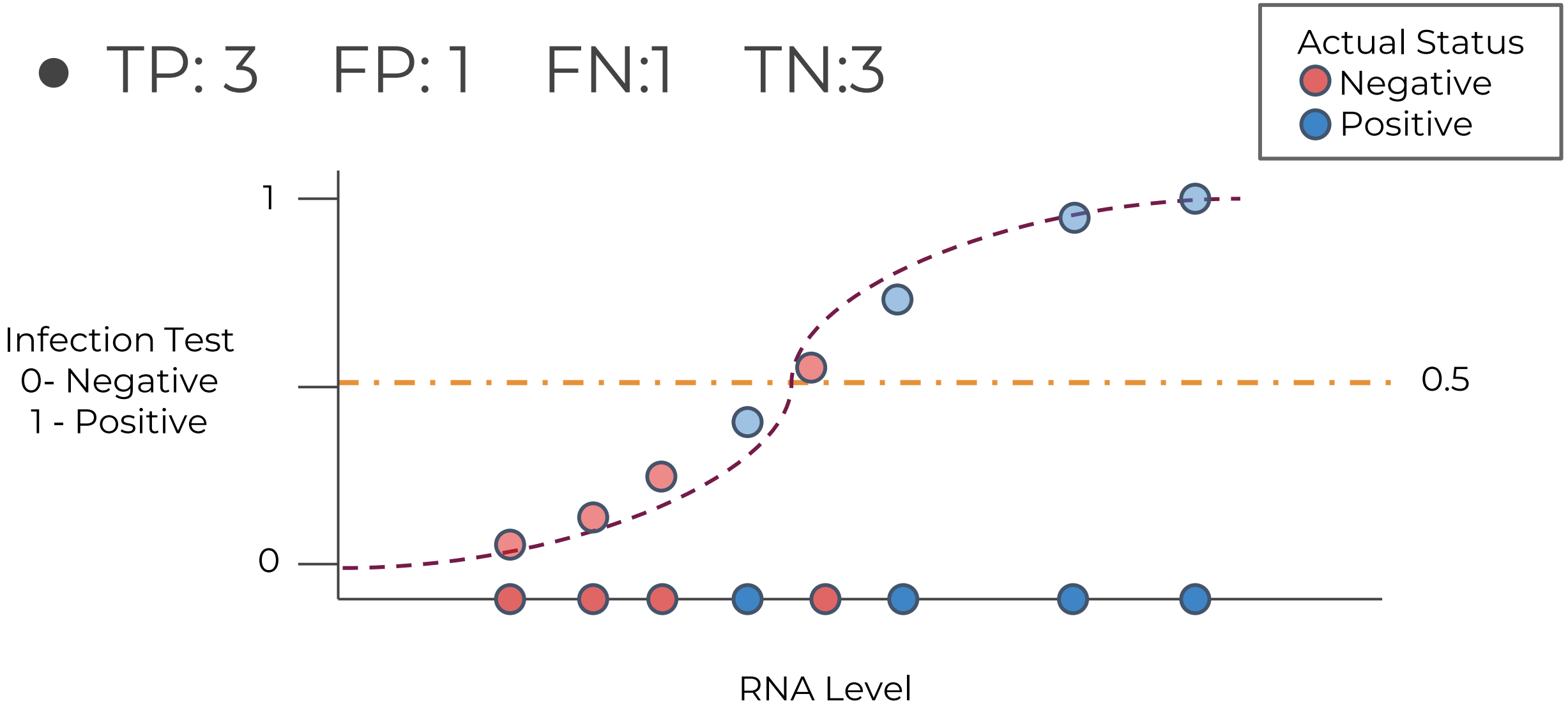
Classification Metrics

- How many TP vs FP?



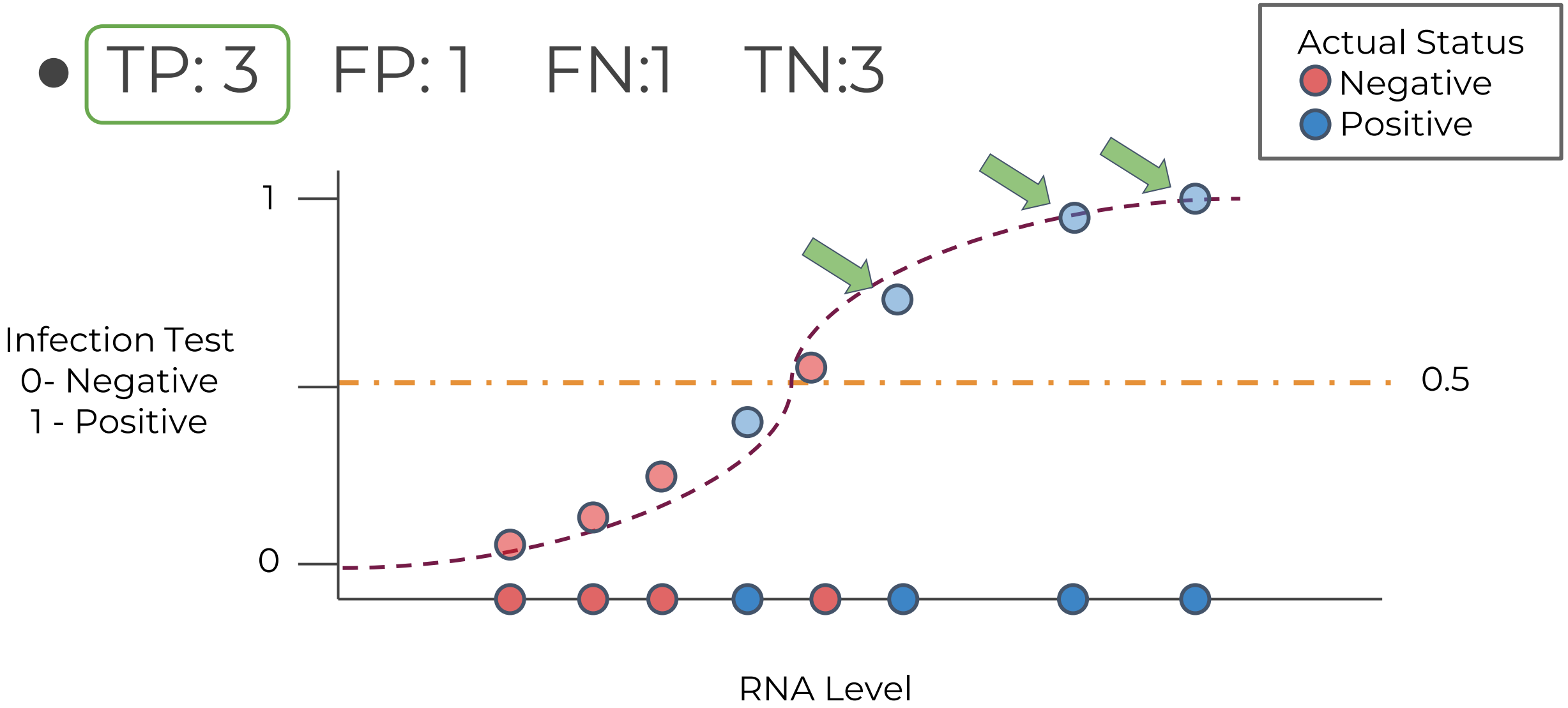
Classification Metrics

- TP: 3 FP: 1 FN:1 TN:3



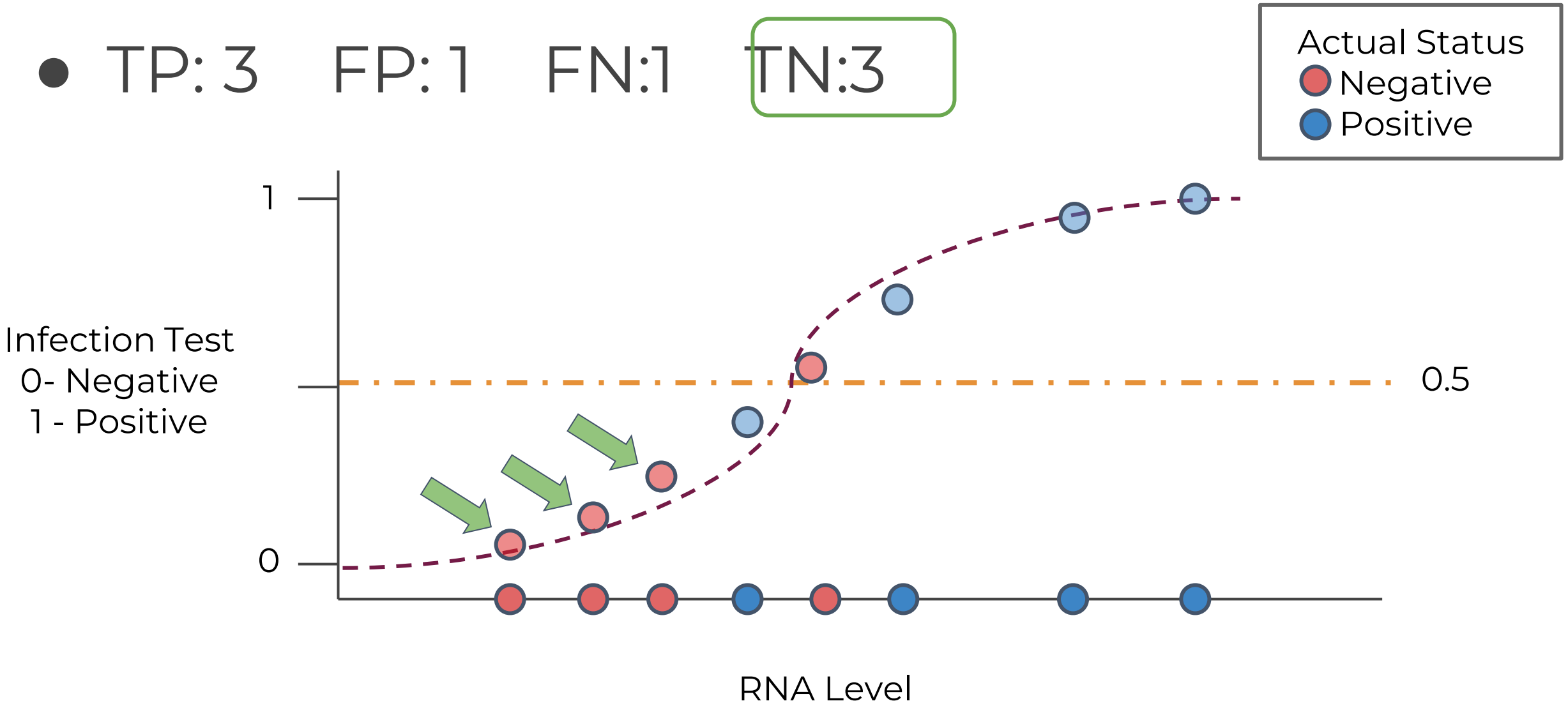
Classification Metrics

- TP: 3 FP: 1 FN:1 TN:3



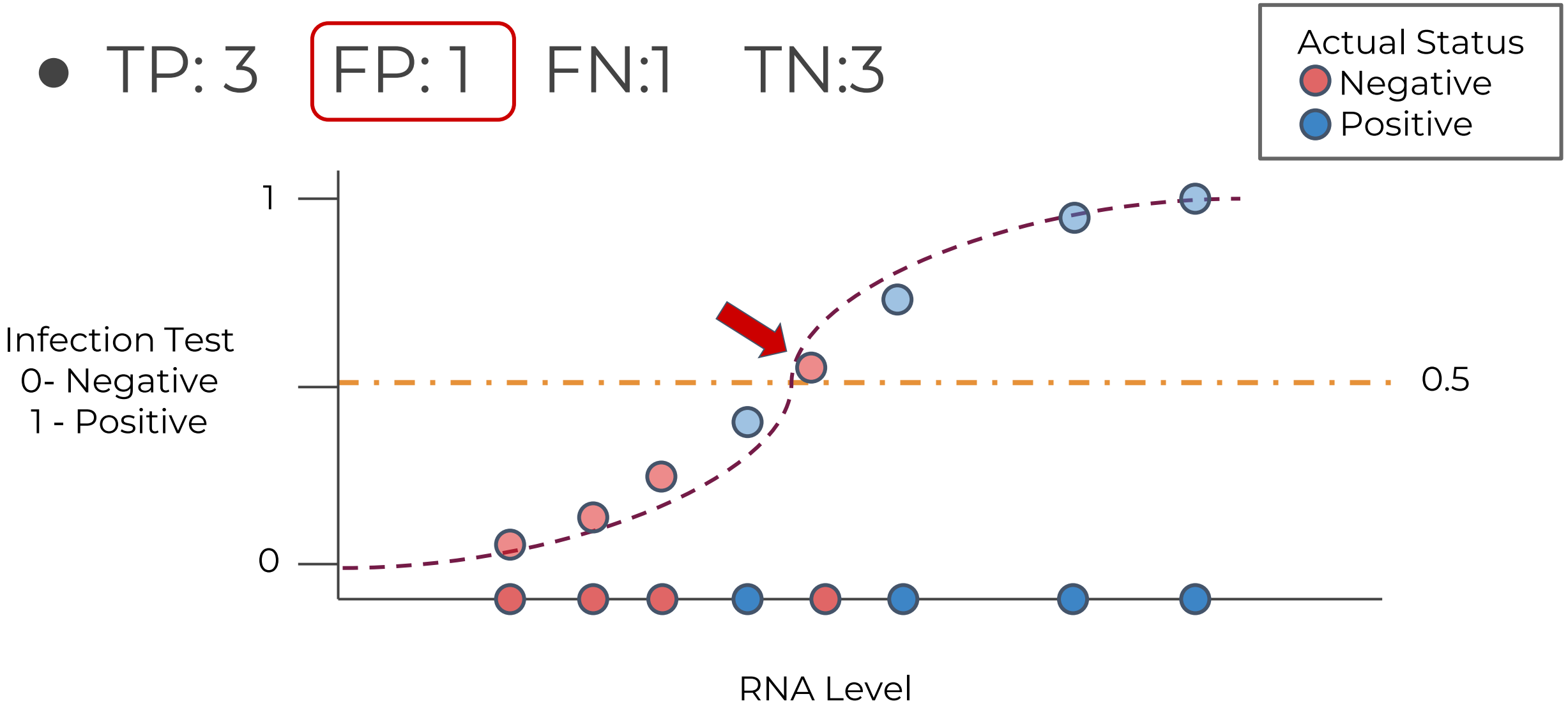
Classification Metrics

- TP: 3 FP: 1 FN:1 TN:3



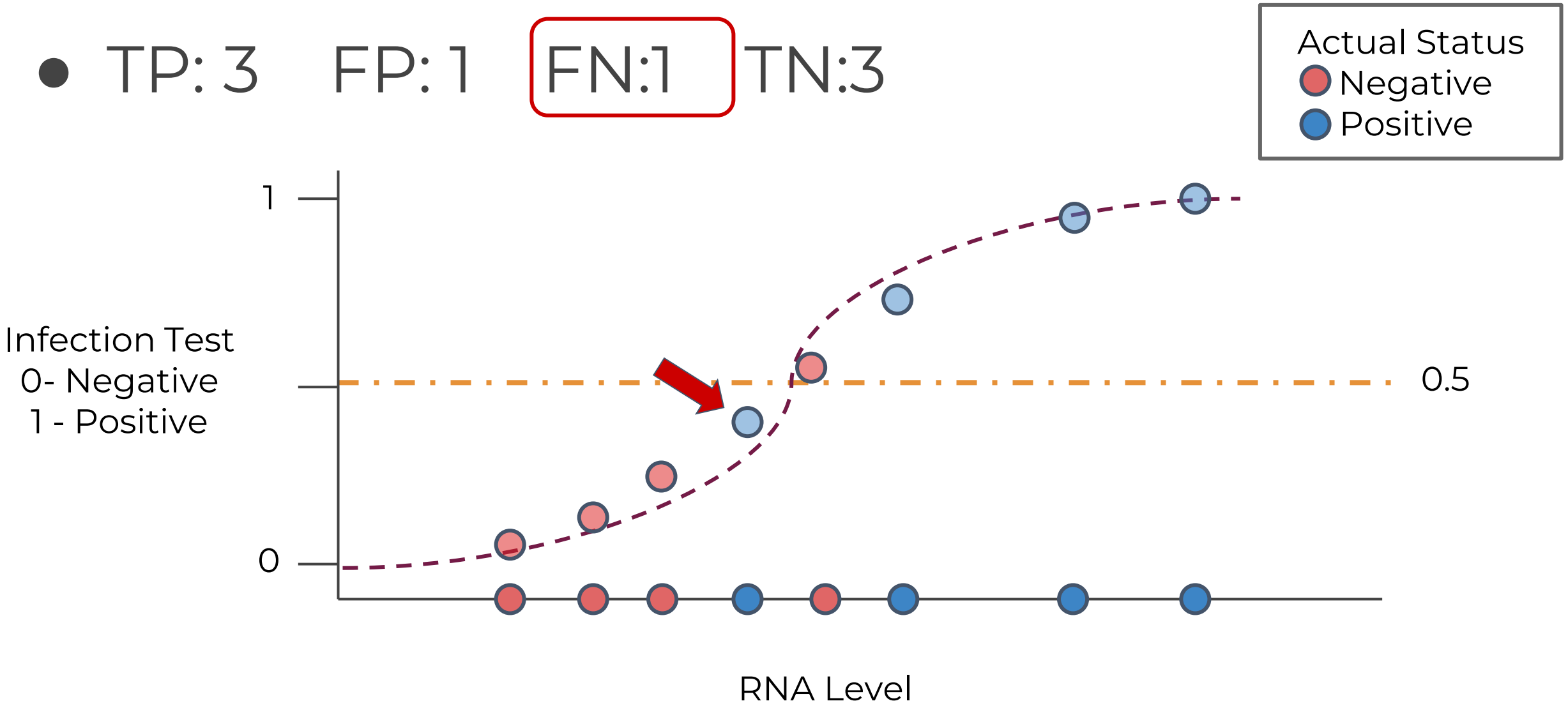
Classification Metrics

- TP: 3 FP: 1 FN:1 TN:3



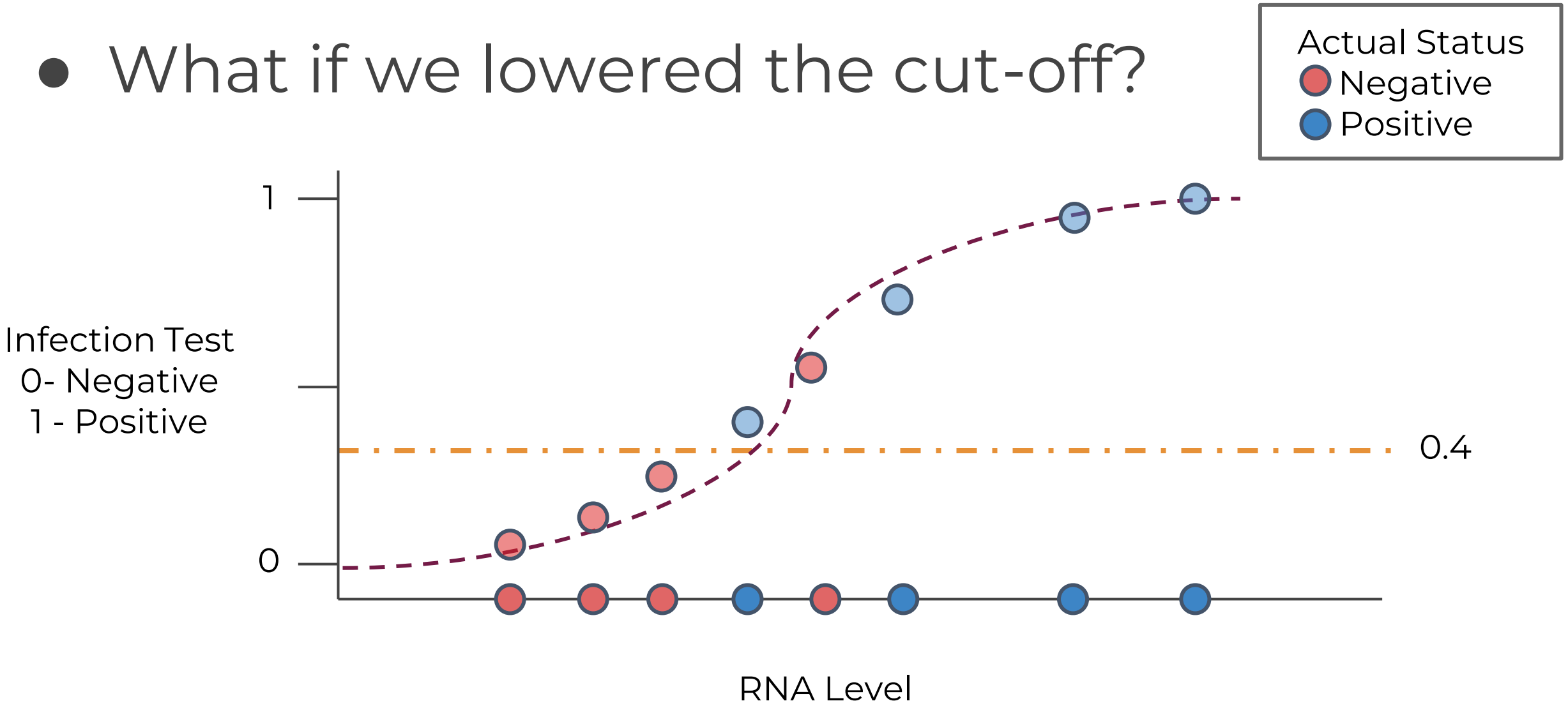
Classification Metrics

- TP: 3 FP: 1 FN:1 TN:3



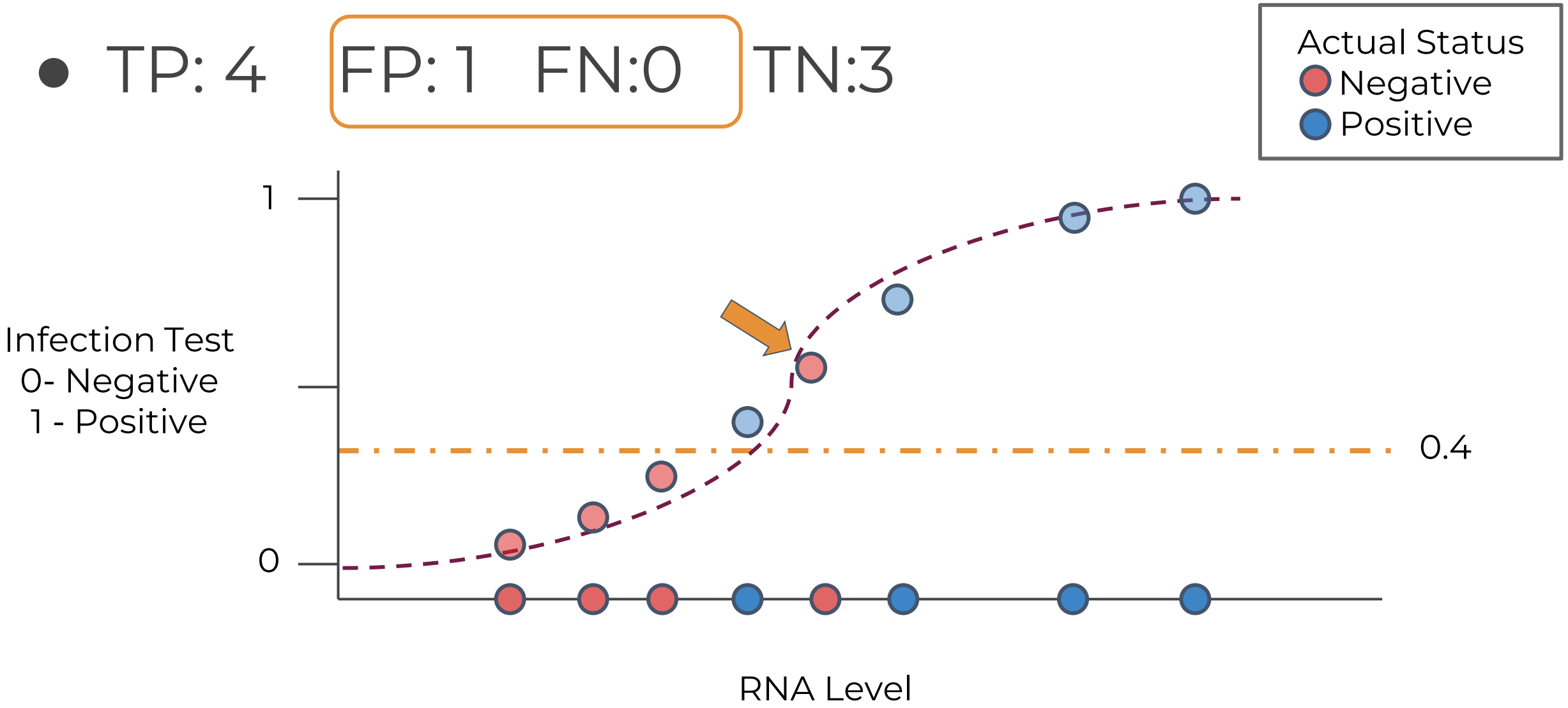
Classification Metrics

- What if we lowered the cut-off?



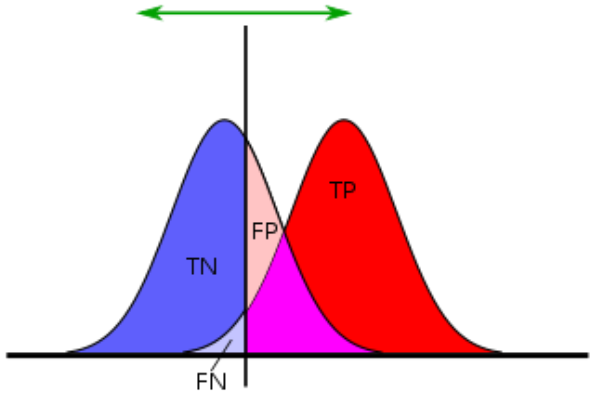
Classification Metrics

- TP: 4 FP: 1 FN:0 TN:3

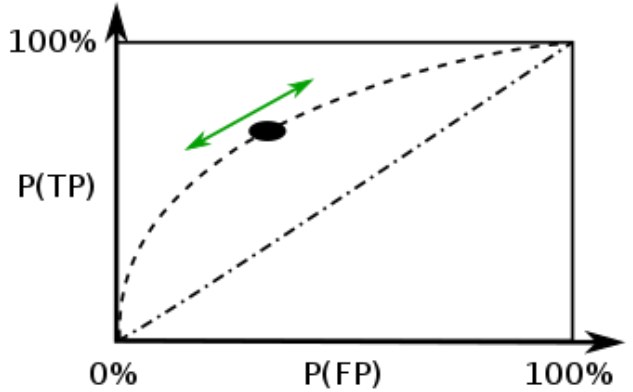


Classification Metrics

- There can be a trade-off between True Positives and False Positives.



TP	FP
FN	TN



Classification Metrics

- They developed the Receiver Operator Characteristic curve.

$$\text{True Positive Rate (TPR)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

$$\text{False Positive Rate (FPR)} = \frac{\text{False Positives (FP)}}{\text{False Positives (FP)} + \text{True Negatives (TN)}}$$

ปัญหาที่ ROC พยายามแก้ (Why ROC?)

- โมเดลทำนายแบบ probability/score ต้องเลือก threshold เพื่อแบ่งชั้น (class)
- Threshold ต่างกัน → Confusion matrix เปลี่ยน → Precision/Recall/TPR/FPR เปลี่ยน
- ROC แสดง performance ทุก threshold พร้อมกัน → เห็นภาพรวมความสามารถแยกแยะ (discrimination)

Notes: ROC ไม่ขึ้นกับ class prevalence เท่ากับ PR (จะพูดเทียบภายหลัง)

🔹 ความหมายของ TPR และ FPR

1. True Positive Rate (TPR) หรือที่มักเรียกว่า Recall / Sensitivity

- นิยาม:

$$TPR = \frac{TP}{TP + FN}$$

- อธิบาย: เป็นสัดส่วนของ **คนที่เป็น Positive จริง ๆ (เช่น เป็นโรค)** ที่โมเดลสามารถทำนายถูกว่าเป็น Positive ได้
- พูดง่าย ๆ = โมเดลจับ "ของจริง" ได้ครบแค่ไหน

---

2. False Positive Rate (FPR)

- นิยาม:

$$FPR = \frac{FP}{FP + TN}$$

- อธิบาย: เป็นสัดส่วนของ **คนที่เป็น Negative จริง ๆ (เช่น ไม่เป็นโรค)** แต่โมเดลดันทำนายผิดว่าเป็น Positive
- พูดง่าย ๆ = โมเดลเผลอ "ใส่ร้าย" ว่ามีโรค ทั้ง ๆ ที่เขาไม่เป็น

🔹 ตัวอย่างเข้าใจง่าย

สมมติเรามีโมเดลตรวจโรคมะเร็ง

- Positive = มีมะเร็ง
- Negative = ไม่มีมะเร็ง

ผลลัพธ์จากโมเดลตรวจ 100 คน (ความจริงอยู่ในวงเล็บ):

ทำนาย/จริง	Positive (มีโรค)	Negative (ไม่มีโรค)
Positive	40 (TP)	10 (FP)
Negative	10 (FN)	40 (TN)

- TPR = 40 / (40+10) = 0.80 → 80%
  หมายถึง โมเดลเจอผู้ป่วยจริง 80%
- FPR = 10 / (10+40) = 0.20 → 20%
  หมายถึง คนที่ไม่เป็นโรค 20% โดนโมเดลใส่ร้าย

🔹 ความสำคัญ

- TPR สูง → ดีต่อการ "ไม่พลาด" ผู้ป่วย (sensitive)
- FPR ต่ำ → ดีต่อการ "ไม่ตกใจเกินเหตุ" สำหรับคนสุขภาพปกติ

การเลือกจุดสมดุลขึ้นอยู่กับงาน เช่น

- งานแพทย์ → เน้น TPR สูง (อย่าให้พลาดผู้ป่วย)
- งานสแปมเมล → เน้น FPR ต่ำ (อย่าให้เมลดี ๆ โดนบล็อก)

Classification Metrics

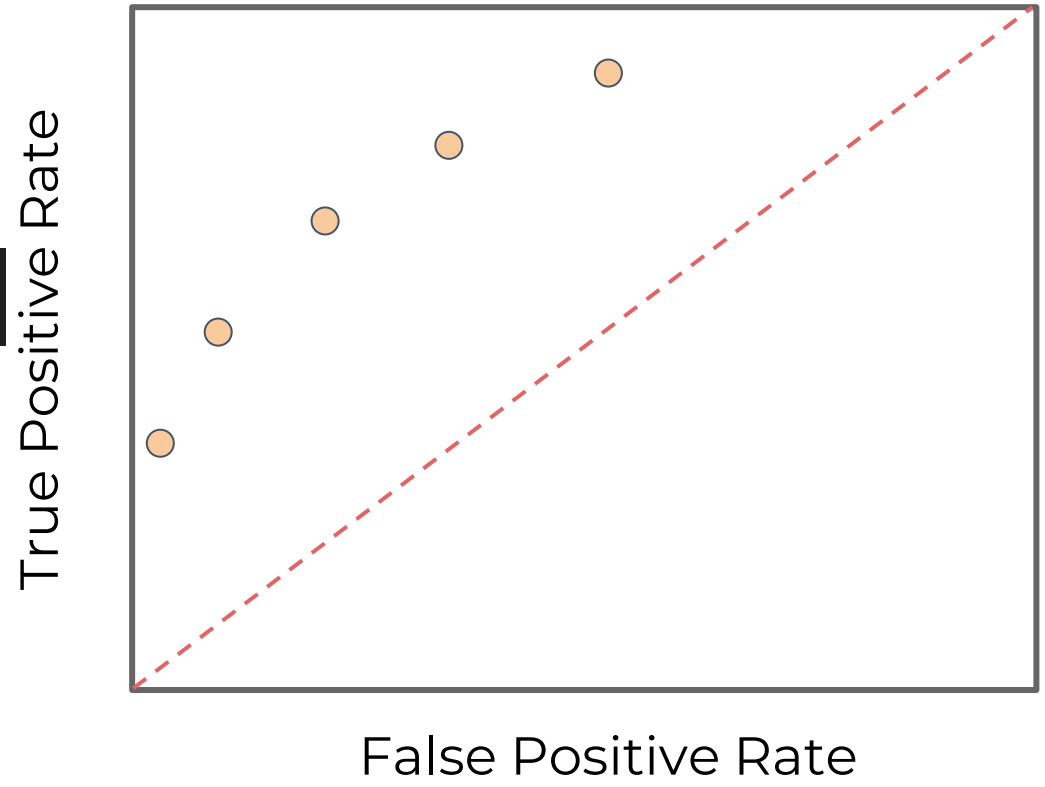
- They developed the Receiver Operator Characteristic curve.

Classification Metrics

- Chart the True vs. False positives for various cut-offs for the ROC curve.

$$\text{True Positive Rate (TPR)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

$$\text{False Positive Rate (FPR)} = \frac{\text{False Positives (FP)}}{\text{False Positives (FP)} + \text{True Negatives (TN)}}$$



Classification Metrics

- They developed the Receiver Operator Characteristic curve.

$$\text{True Positive Rate (TPR)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

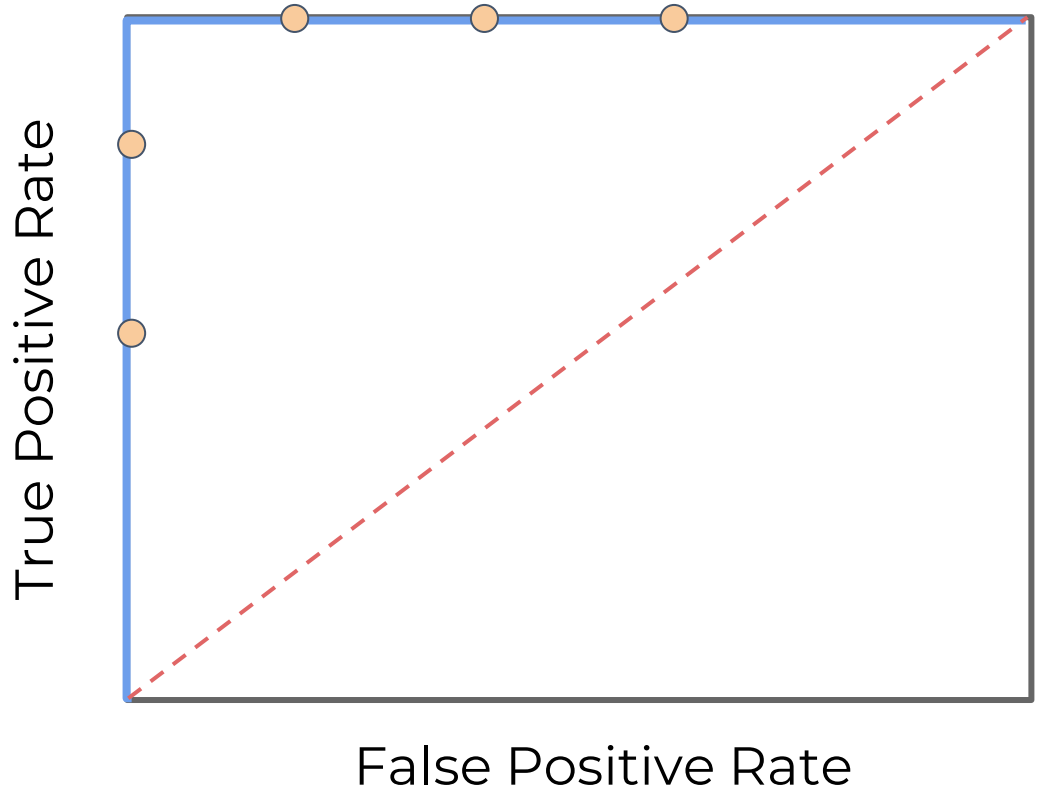
$$\text{False Positive Rate (FPR)} = \frac{\text{False Positives (FP)}}{\text{False Positives (FP)} + \text{True Negatives (TN)}}$$

Classification Metrics

- There can be a trade-off between True Positives and False Positives.

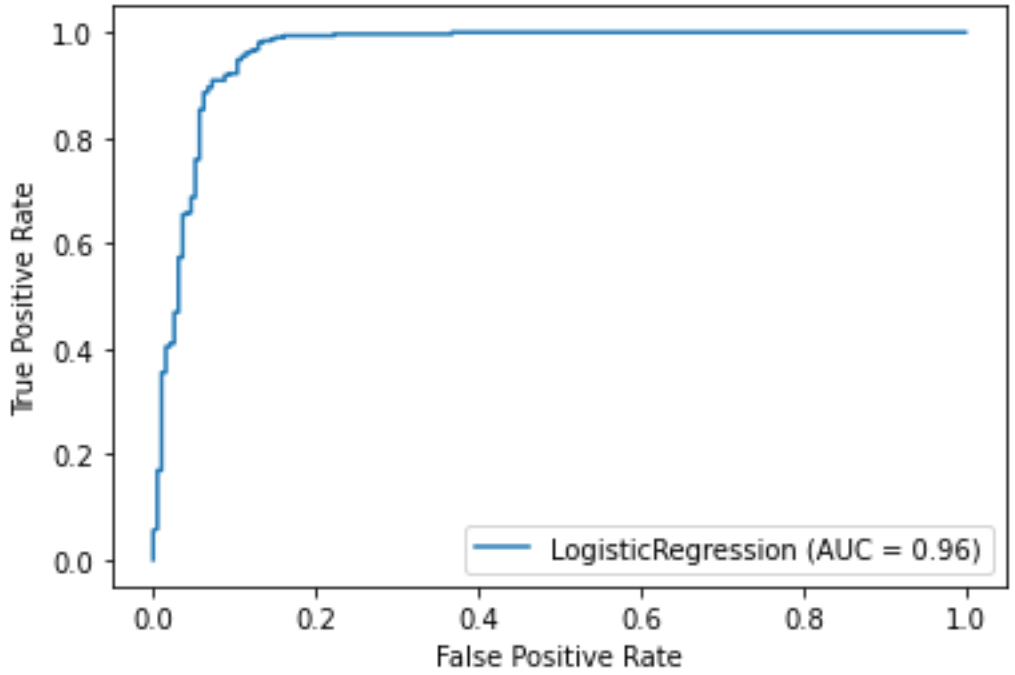
Classification Metrics

- A perfect model would have a zero FPR.
- Random guessing is the red line.



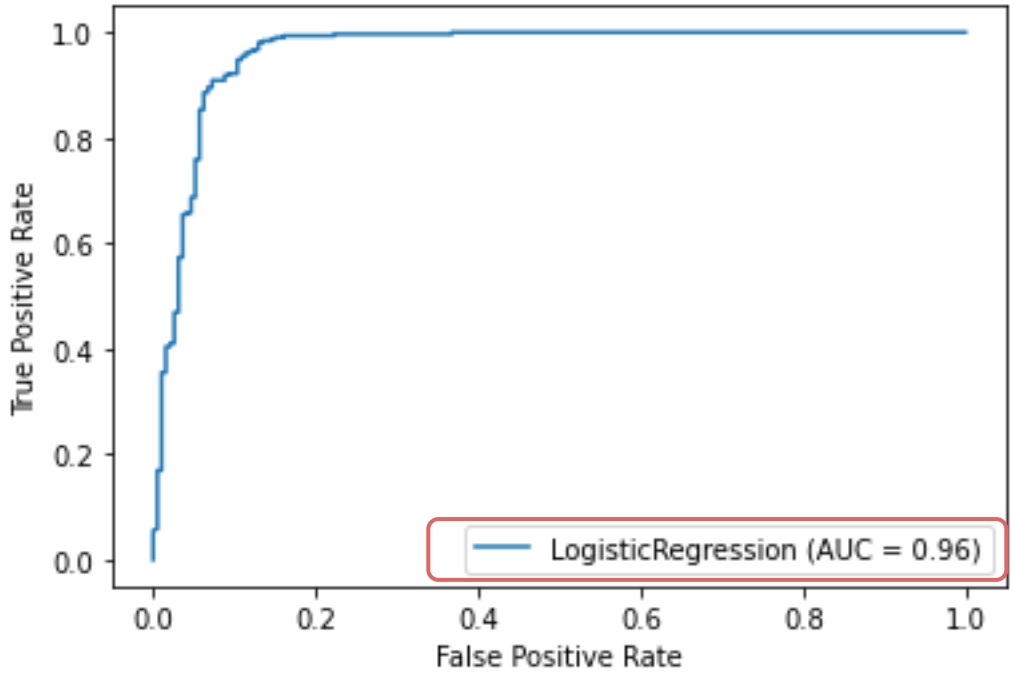
Classification Metrics

- Realistically with smaller data sets the ROC curves are not as smooth.



Classification Metrics

- AUC - Area Under the Curve , allows us to compare ROCs for different models.



Classification Metrics

- Can also create precision vs. recall curves: