Formative Activity: ROC Curve

**ROC Example:**

We have This simple example of ten observations, and we also have a classifier, it could be any classifier. This classifier gives us what is the probability of the observation belong to class 1 P(1|A) in Table 1 the observations are sorted from highest probability to lowest. **True class** is the actual class, which is from training set.

We are going to apply this rule:

If P(1|A) is greater than *t* then Ŷ = 1 else Ŷ = 0

*t* is a threshold.

**Table 1: Classifier probabilities and target variable values in training dataset**

|  |  |  |
| --- | --- | --- |
| Instance | P(1|A) | True class |
| 1 | 0.95 | 1 |
| 2 | 0.93 | 1 |
| 3 | 0.87 | 0 |
| 4 | 0.84 | 1 |
| 5 | 0.84 | 0 |
| 6 | 0.84 | 0 |
| 7 | 0.75 | 0 |
| 8 | 0.52 | 1 |
| 9 | 0.44 | 0 |
| 10 | 0.26 | 1 |

**Table 2: Threshold for all the possible trade-offs**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Class | + | - | + | - | - | - | + | - | + | + |  |
| *t* | 0.26 | 0.44 | 0.52 | 0.75 | 0.84 | 0.84 | 0.84 | 0.87 | 0.93 | 0.95 | 1.00 |
| TP | 5 | 4 | 4 | 3 | 3 | 3 | 3 | 2 | 2 | 1 | 0 |
| FP | 5 | 5 | 4 | 4 | 3 | 2 | 1 | 1 | 0 | 0 | 0 |
| TN | 0 | 0 | 1 | 1 | 2 | 3 | 4 | 4 | 5 | 5 | 5 |
| FN | 0 | 1 | 1 | 2 | 2 | 2 | 2 | 3 | 3 | 4 | 5 |
| TPR | 1 | 0.8 | 0.8 | 0.6 | 0.6 | 0.6 | 0.6 | 0.4 | 0.4 | 0.2 | 0 |
| FPR | 1 | 1 | 0.8 | 0.8 | 0.8 | 0.4 | 0.2 | 0.2 | 0 | 0 | 0 |

**This table will generate Figure 1.**

Et bilde som inneholder tekst, diagram, line, nummer

KI-generert innhold kan være feil.

**Figure 1: ROC curve**

This is how ROC curve could be construct.

**Question:**

**What is the best threshold for this example?**

**Solution**

ROC Curve (Receiver Operating Characteristic Curve):

* It plots the True Positive Rate (TPR, or sensitivity) agains the False Positive Rate (FPR, or 1-specificity) at various threshold settings
* It helps visualize the trade-off between sensitivity and specificity

Best threshold:

* The ideal threshold maximizes the TPR while minimizing the FPR
* Visually, this corresponds to the point on the ROC curve that is closest to the top-left corner (where TPR=1 And FPR=0)
* Also the point that maximizes the area under the curve (AUC)
* In a more practical sense, the "best" threshold depends on the specific costs associated with false positives and false negatives in your application

Analyzing table 2:

* How to analyze the table to find the best threshold:
  1. Look for the "Elbow" or Steepest Rise:
     + The best threshold often lies in the region where the ROC curve rises most steeply. This indicates a significant increase in TPR with a relatively small increase in FPR.
     + Looking at the table, we want to find the place where the TPR increases the most while the FPR increases the least.
  2. Consider the Trade-off:
     + If false positives are very costly, you might prefer a lower FPR, even if it means a slightly lower TPR.
     + If false negatives are very costly, you'd prioritize a high TPR, even if it means a higher FPR.
  3. Specific Analysis of the Table:
     + Notice the jump from FPR=0.2 and TPR=0.6 to FPR=0 and TPR=0.4. This is a bad jump, because the TPR decreased, and the FPR decreased.
     + The largest increase of TPR while having the smallest increase of FPR happens when the treshold goes from 0.84 to 0.75.
     + When treshold goes from 0.84 to 0.75 TPR increases from 0.6 to 0.6 and the FPR increases from 0.4 to 0.8. this is not a good jump.
     + The best jump is from treshold 0.87 to 0.84, where the TPR increases from 0.4 to 0.6, and the FPR increases from 0.2 to 0.4.
     + When the treshold is 0.84, we have a TPR of 0.6 and a FPR of 0.4.

Conclusion:

* Based on the analysis, a threshold of **0.84** appears to be a reasonable choice. It offers a good balance between TPR and FPR.
  + TPR = 0.6
  + FPR = 0.4

* Therefore, if P(1|A) is greater than 0.84, then Ŷ = 1; otherwise, Ŷ = 0.