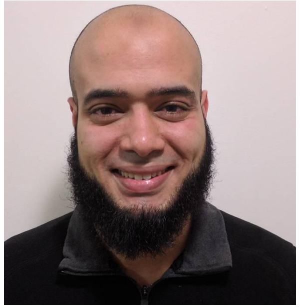
Machine Learning

Evaluation Metrics 6

Mostafa S. Ibrahim
Teaching, Training and Coaching for more than a decade!

Artificial Intelligence & Computer Vision Researcher
PhD from Simon Fraser University - Canada
Bachelor / MSc from Cairo University - Egypt
Ex-(Software Engineer / ICPC World Finalist)

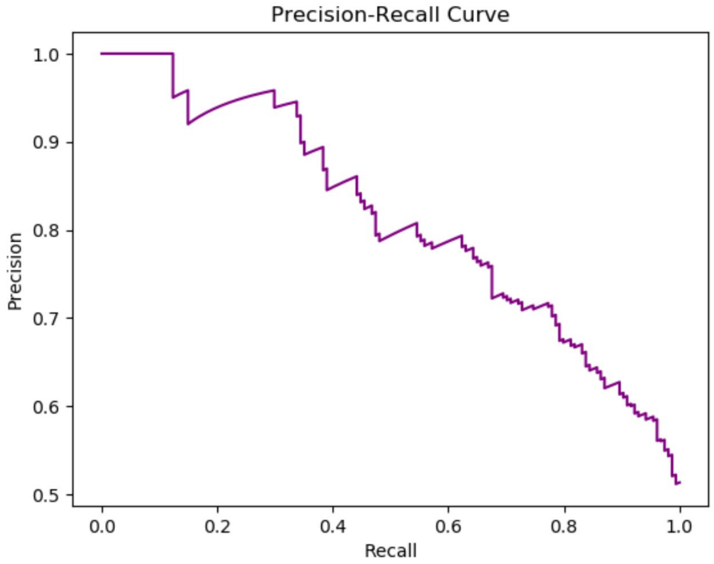


Open Questions

- What if we want the threshold of the best precision/recall?
- What if we are confused / hard to handle all such thresholds?

Precision-Recall curve

- The precision-recall curve plots the precision values against different levels of recall (*threshold*), where we plot the **recall directly against** the precision
    - Each point (recall, precision) on the curve represents a **specific threshold** value
    - plt.plot(recall, precision, **'r--'**)

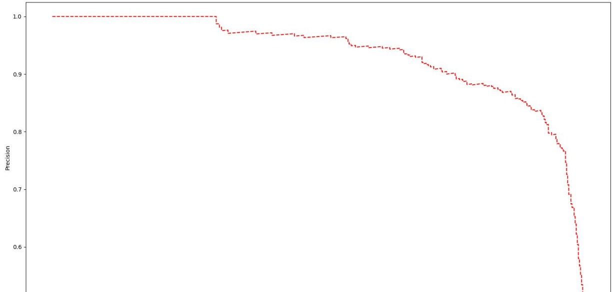


SKlearn: precision_recall_curve

- precision, recall, threshold = precision_recall_curve(y_gt, y_prop)
- Internally, for the thresholds, just use the **distinct probabilities** (in y_prop) as thresholds!

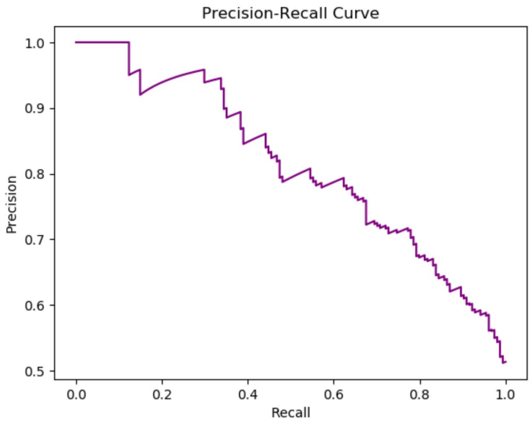
How to find the best threshold?

1. If there is a specific recall or precision, find it directly, as we coded
2. The best performance is at point (1, 1). Find the euclidean distance of each point (r, p) to (1, 1) and use the minimum distance one!
3. Or compute F-score for each point, and find the maximum F-score
4. Youden's J Statistic Maximization
    a. a measure of the classifier's ability to simultaneously maximize both true positive rate (recall) and true negative rate (specificity).
        i. It is defined as J = sensitivity + specificity - 1, see wiki
    b. Compute for each point (r, p) and find the maximum



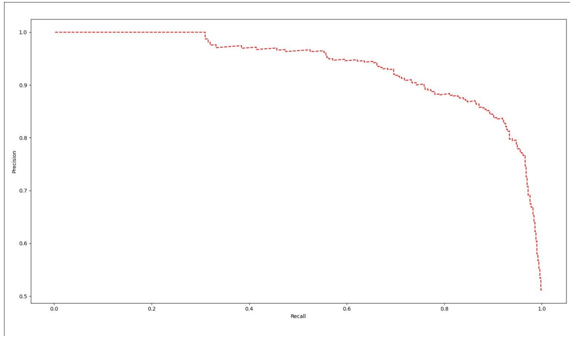
How to find the best threshold?

1. Previous ways are very **systematic**. Another observation-based way:
2. Understand your task and its precision-recall requirements
3. Examine the Precision-Recall curve to understand the **trade-off** between precision and recall
4. Determine the desired precision-recall trade-off
5. Find the threshold of **this point** of interest and apply it
6. Tip: use the previous methods to give you some initial points to visually start from



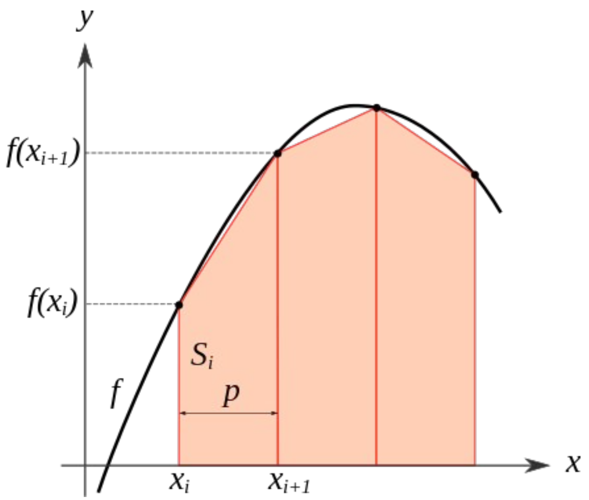
AUC-PR

- **Area under the curve** of precision-recall is another interesting metric
- It summarizes the **whole curve in a single** value (no thresholds)
    - Threshold-independent method
    - The higher the area, the better the classifier
- A high area under the curve represents both high recall (low false negative rate) and high precision (low false positive rate)



Area under the curve: Implementation

- How can we compute such area under a curve?!
- One generic way is the trapezoidal rule
- To keep it simple and informal, divide the function into many little trapezoids
  - We have the trapezoid base: differences in xs
  - And we have the 2 heights, we take their average!
- Now approximate their area and sum them
- We can evaluate AUC-PR using it
  - **from** sklearn.metrics **import** roc_auc_score
  - roc_auc_score(y_gt, y_prop)
    - roc is another curve metric
  - can be used with binary and multiclass classification



Area under the curve: Implementation

- We can also make another implementation, customized to AUC-PR
- It is based on average precision (next lecture)
- Code
  - **from** sklearn.metrics **import** average_precision_score
  - average_precision_score(y_gt, y_prop)
  - We typically use it for auc-pr, for binary classifiers

$$\mathrm{AP} = \sum_n (R_n - R_{n-1}) P_n$$

When to use AUC-PR?

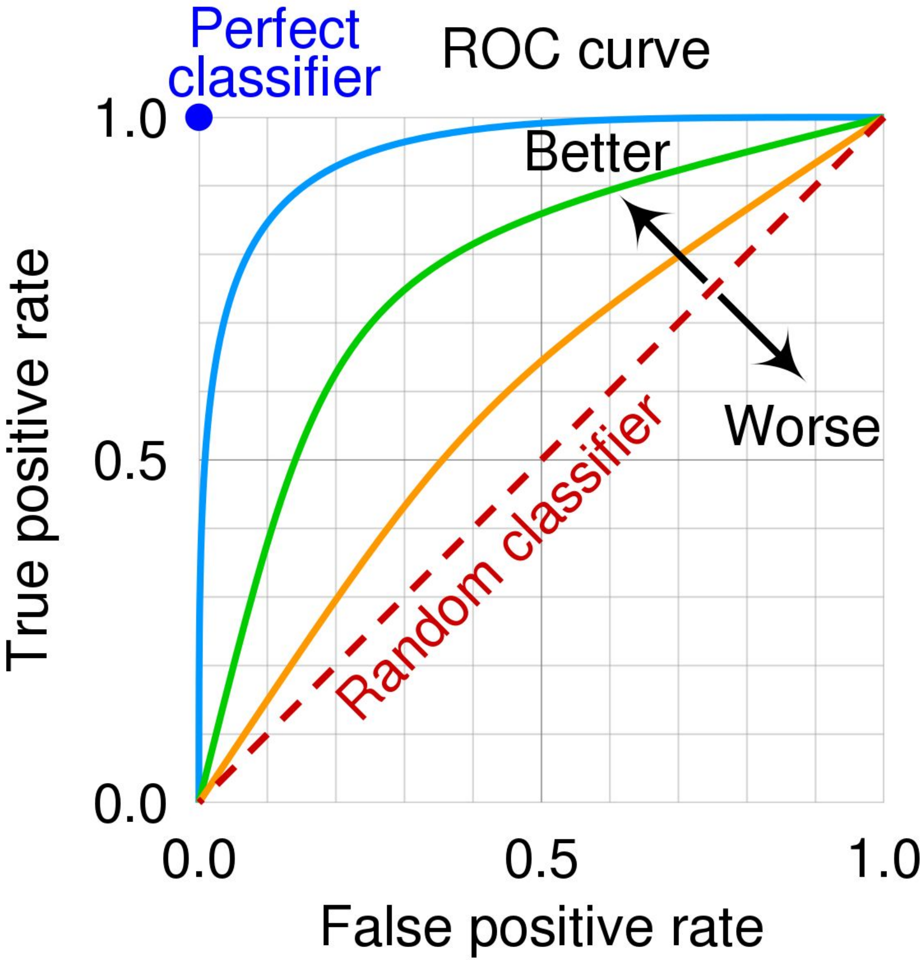
- AUC-PR is based on precision and recall. So, when we need them!
  - Your model predicts the positive examples correctly
- **Imbalanced** Datasets: AUC-PR is well-suited for imbalanced datasets where the number of negative samples significantly outweighs the positive samples
  - It also works on balance datasets
- When you are overwhelmed by the confusion matrices on different thresholds, it is better to focus for a while on a threshold-independent metric
- Tip: when comparing with papers, it is important to know how exactly the value is computed (e.g. Interpolated average precision, 11-point interpolated average precision, etc, like in PASCAL contest)

ROC Curve

- ROC stands for Receiver Operating Characteristic (from signal theory)
- It shows the the trade-off between True Positive Rate (TPR) and the False Positive Rate (FPR) for every threshold
  - Note: TPR = Recall = TP / (TP + FN)
  - FPR = 1 - Specificity
- We can compute its area (AUC) as a single metric (threshold-independent)
- This metric is used with **only balanced datasets**
- We can use it when we emphasize on the false positive rate
  - Remember, FPR by itself is based on a single threshold
- Like AUC-PR (AUPRC), We can use AUC-ROC to visually compare the performance of multiple classifiers

ROC Curve

- Oserve, recall on y-axis
- The baseline: Random classifier has area = 0.5 (on diagonal)
  - TRP = FPR
- Visually, we can compare different models
- roc_auc_score(y_gt, y_prop)



Random Classifier

- A random classifier is basic baseline where the classifier makes predictions randomly regardless of the given input
  - The score depends on the metric and the ground truth distribution
- A random classifier will score
  - 0.5 for AUC-ROC
    - You shouldn’t use this metric for imbalance dataset.
    - It can be misleading (e.g. higher than AUC-PR)
  - For AUC-PR, it scores equal to the fraction of positives samples
    - A horizontal line at P/N
  - *I skipped why such values for these 2 metrics*
- In practice, we should notice if our model is behaving like a random classifier or **significantly** outperforms it

Random Classifier

- Sample of 10000
- Precision and AUC-PR are at value = percent of positive fractions
- All other values are *approximately* at 0.5
- Play with attached code

```
Positive examples are 0.5%
Accuracy: 0.50 - Precision: 0.50 - Recall: 0.50 - AUC-PR: 0.50 - AUC-ROC: 0.50

Positive examples are 0.25%
Accuracy: 0.50 - Precision: 0.25 - Recall: 0.50 - AUC-PR: 0.25 - AUC-ROC: 0.50

Positive examples are 0.75%
Accuracy: 0.50 - Precision: 0.75 - Recall: 0.50 - AUC-PR: 0.75 - AUC-ROC: 0.50

Positive examples are 0.98%
Accuracy: 0.50 - Precision: 0.98 - Recall: 0.50 - AUC-PR: 0.98 - AUC-ROC: 0.50
```

"Acquire knowledge and impart it to the people."

"Seek knowledge from the Cradle to the Grave."