

Hands-on Machine Learning (Gueron) Ch 3: Classification

DOLOR SIT AMET

MNIST Dataset



A 10x10 grid of handwritten digits from the MNIST dataset. The digits are as follows:

5	0	4	1	9	2	1	3	1	4
3	5	3	6	1	7	2	8	6	9
4	0	9	1	1	2	4	3	2	7
3	8	6	9	0	5	6	0	7	6
1	8	1	9	3	9	8	5	9	3
3	0	7	4	9	8	0	9	4	1
4	4	6	0	4	5	6	1	0	0
1	7	1	6	3	0	2	1	1	7
9	0	2	6	7	8	3	9	0	4
6	7	4	6	8	0	7	8	3	1

- ❖ A dataset comprised of 70,000 hand-written, 28 x 28 (784 features) pixelated images of digits 0-9 written by high school students and US Census Bureau
- ❖ Classify the handwritten digits from 0 – 9
- ❖ Gueron simplifies the analysis by limiting to classification to the digit 5 using a binary classifier

Training a Binary Classifier

1. Split into training and test datasets and explores training dataset
 1. Heterogeneity – different portion of digits per category
 2. Homogeneity – same/similar portion of digits per category
 3. Shuffle data prior to minimize heterogeneity
2. Fits SGD Classifier
3. Predicts class
4. Evaluate classifier

Accuracy – 10% Digits 5

Classifier	Accuracy
Stochastic Gradient Descent	0.95
Never5/Guessing	0.90

Confusion Matrix

		Predicted		
		Negative	Positive	
Actual	Negative	8 3 9	6	Precision (e.g., 3 out of 4)
	Positive	5 5	5 5 5	
		Recall (e.g., 3 out of 5)		

Diagram illustrating a Confusion Matrix for a classification task. The matrix is divided into four quadrants based on Actual (rows) and Predicted (columns) outcomes.

Quadrants and Counts:

- TN (True Negative):** 8 (Actual Negative, Predicted Negative)
- FP (False Positive):** 6 (Actual Negative, Predicted Positive)
- FN (False Negative):** 5 (Actual Positive, Predicted Negative)
- TP (True Positive):** 5 (Actual Positive, Predicted Positive)

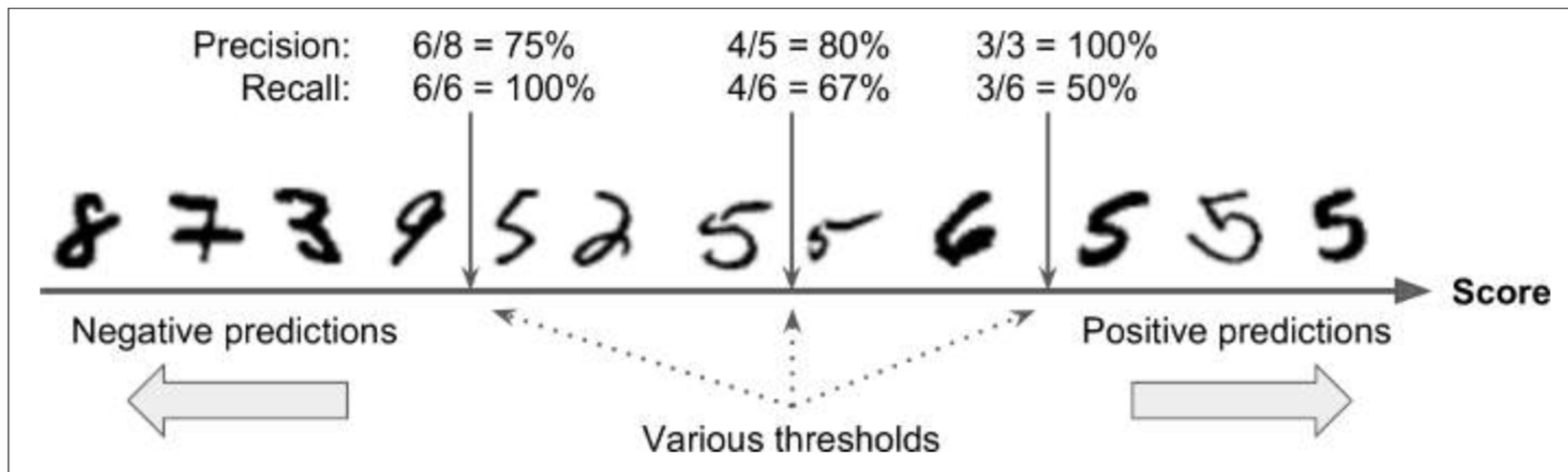
Metrics:

- Precision:** (e.g., 3 out of 4) - Calculated as TP / (TP + FP)
- Recall:** (e.g., 3 out of 5) - Calculated as TP / (TP + FN)

Confusion Matrix Terms

Cell Terms	True positive	Predict +, Actual +
	True negatives	Predict -, Actual -
	False positives	Predict +, Actual -
	False negatives	Predict -, Actual +
Marginal Terms	Precision	$TP : (TP + FP)$
	Recall	$TP : (TP + FN)$
	Accuracy	$(TP + TP) : N$
Composite	F1-score	Harmonic mean of precision and recall

Precision/Recall Tradeoff

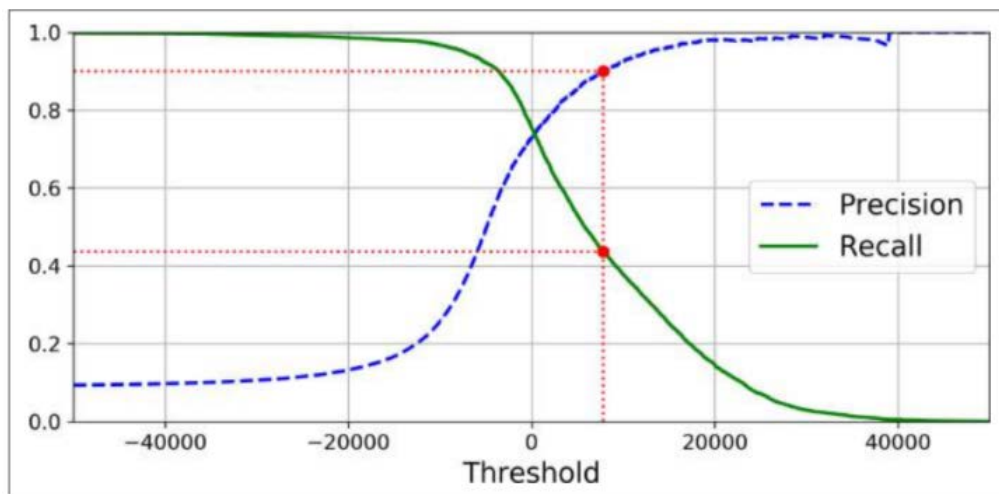


Effect Precision-Recall on F1

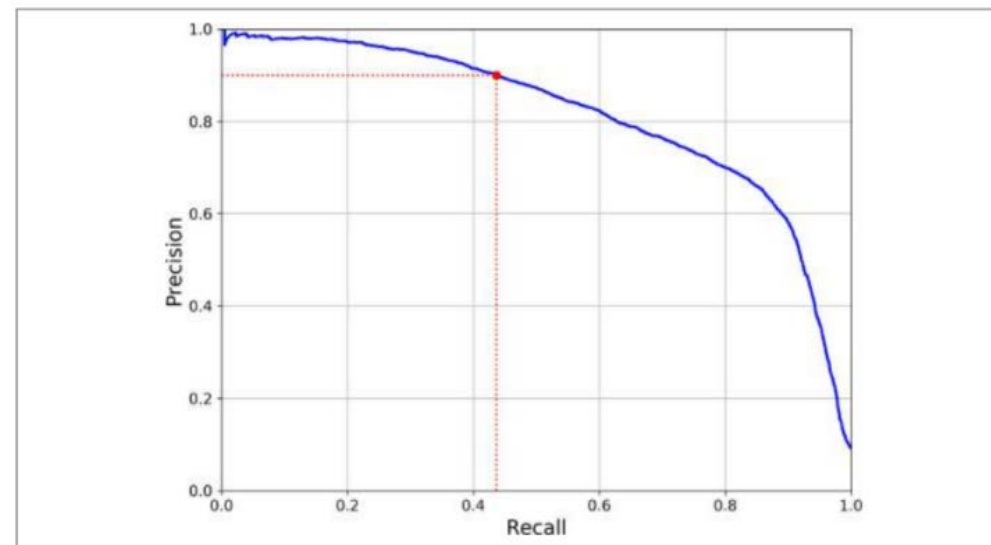
Precision	Recall	F1
0.75	1.00	0.857
0.80	0.67	0.729
1.00	0.50	0.667

Precision vs Recall

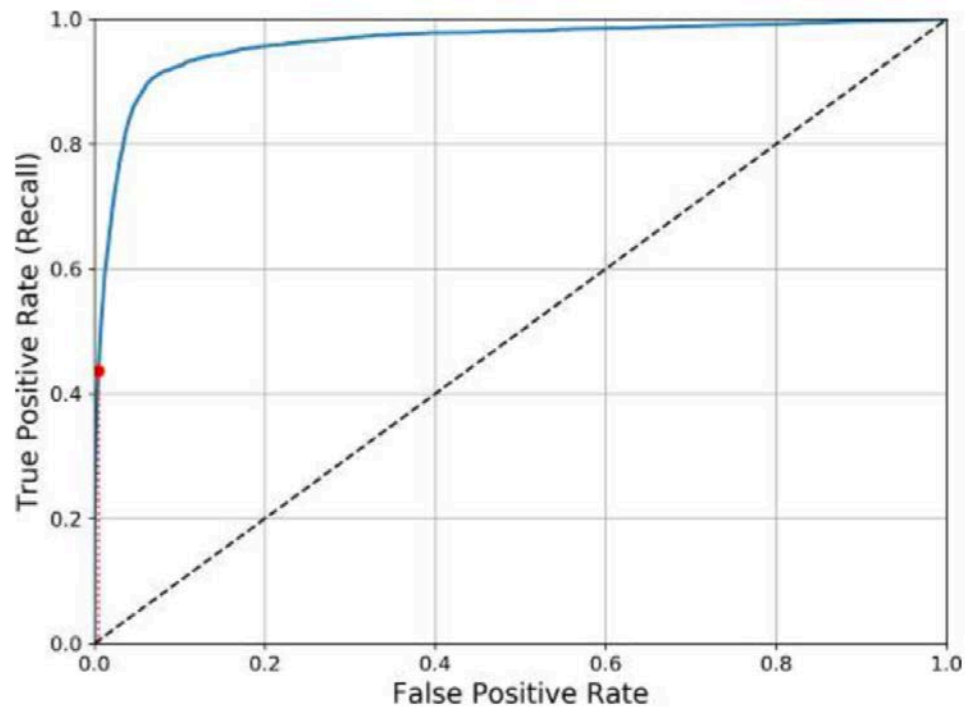
Precision-Recall vs Threshold



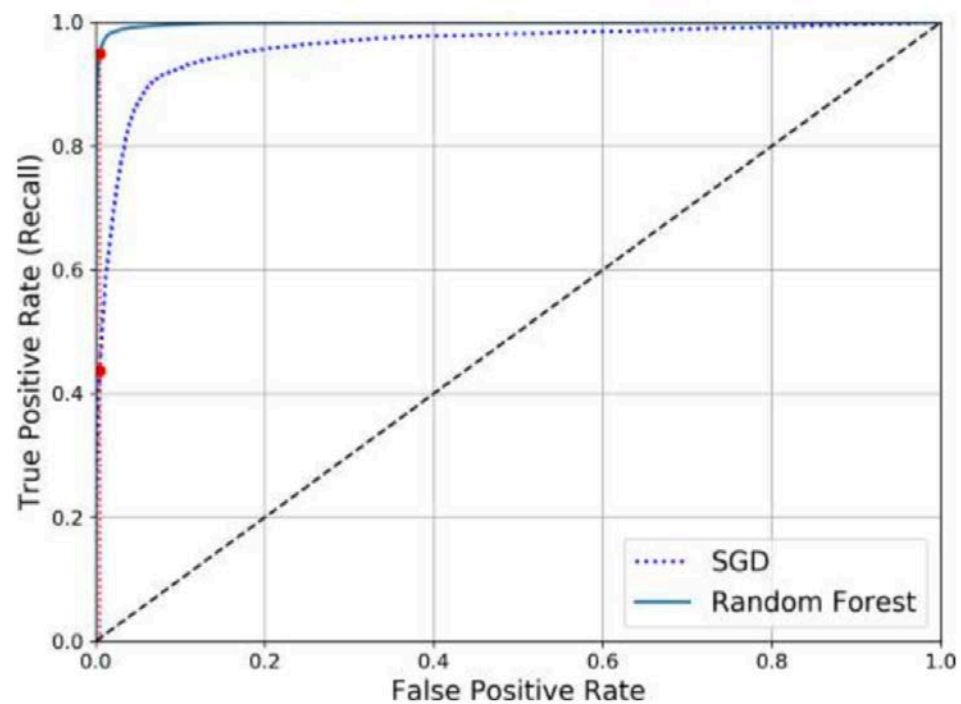
Precision vs Recall



ROC Curve: SGD Classifier

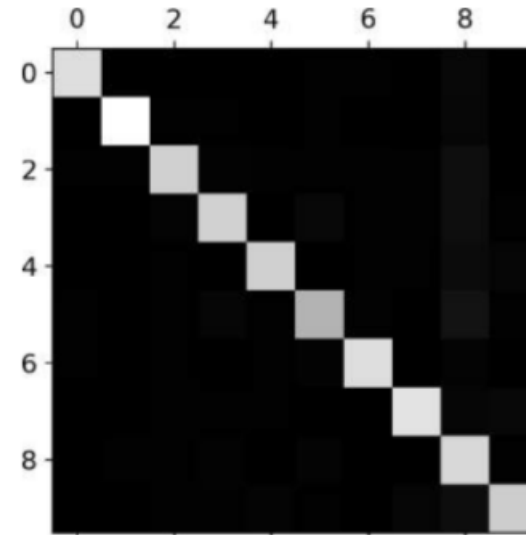


ROC Curve: SGD vs Random Forest

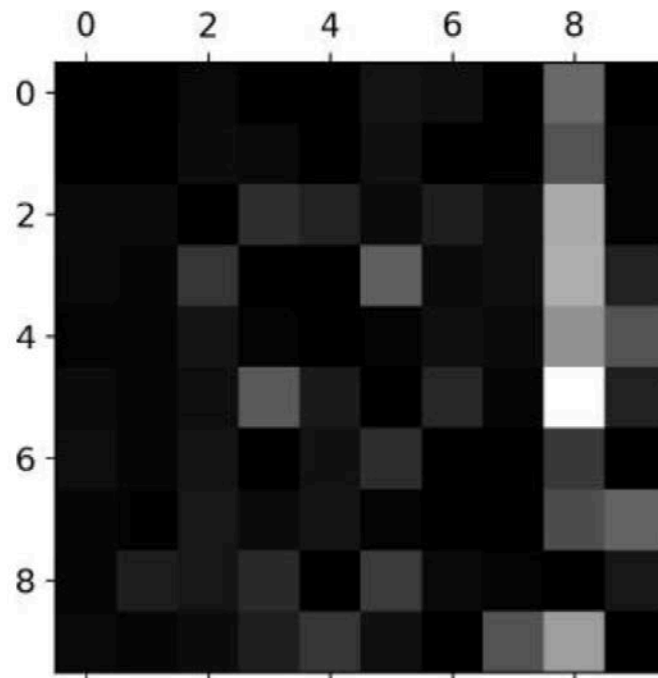


Error Analysis: Raw Counts

```
array([[5578,  0,  22,  7,  8,  45,  35,  5, 222,  1],  
       [  0, 6410,  35,  26,  4,  44,  4,  8, 198, 13],  
       [ 28,  27, 5232, 100,  74,  27,  68,  37, 354, 11],  
       [ 23,  18, 115, 5254,  2, 209,  26,  38, 373, 73],  
       [ 11,  14,  45,  12, 5219,  11,  33,  26, 299, 172],  
       [ 26,  16,  31, 173,  54, 4484,  76,  14, 482,  65],  
       [ 31,  17,  45,  2,  42,  98, 5556,  3, 123,  1],  
       [ 20,  10,  53,  27,  50,  13,  3, 5696, 173, 220],  
       [ 17,  64,  47,  91,  3, 125,  24,  11, 5421,  48],  
       [ 24,  18,  29,  67, 116,  39,  1, 174,  329, 5152]])
```



Error Analysis: Error Rates



Multi-s Target Variable Differences

	Num Class Categories	Num Targets/Y's
Multiclass	2 or more	1
Multilabel	2	2 or more
Multioutput	2 or more	2 or more