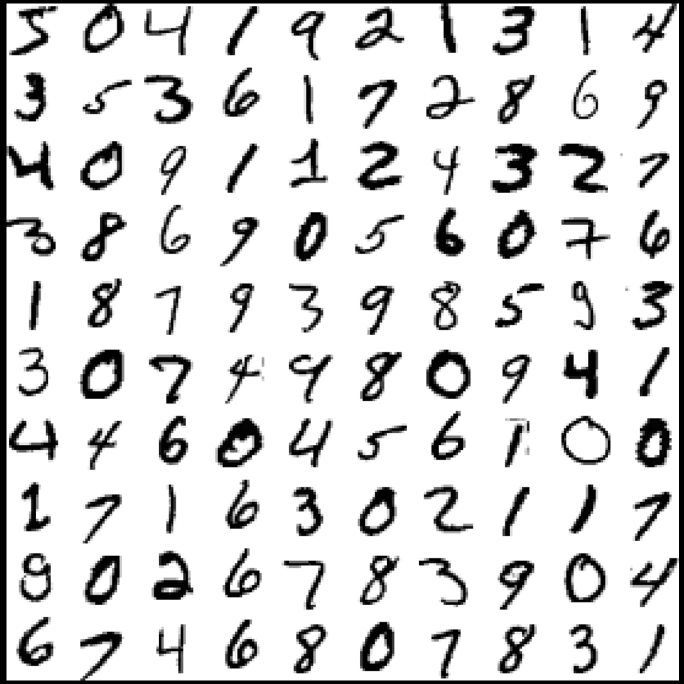
Classification

Chapter 3: pp 85 – 109

MNIST Dataset

- 70,000 28x28 pixel labeled images of handwritten digits of high school students
  - 784 features

- Train a binary classifier to distinguish two classes
  - “5” and “not 5”
  - Use SGD-Classifier



Performance Measures

- SGD-Classifier with three-fold cross validation
  - Accuracy = 93%
- Dumb classifier (classify every digit as “not-5” class)
  - Accuracy = 90%
  - Why?
- Accuracy is not sufficient as a performance measure when dealing with skewed datasets.

Confusion Matrix

- Each row represents an actual class
- Each column represents a predicted class

	Predicted Class “Not-5”	Predicted Class “5”
Actual Class “Not-5”	53,057 (true negatives)	1,522 (false positives)
Actual Class “5”	1,325 (false negatives)	4,096 (true positives)

Precision & Recall

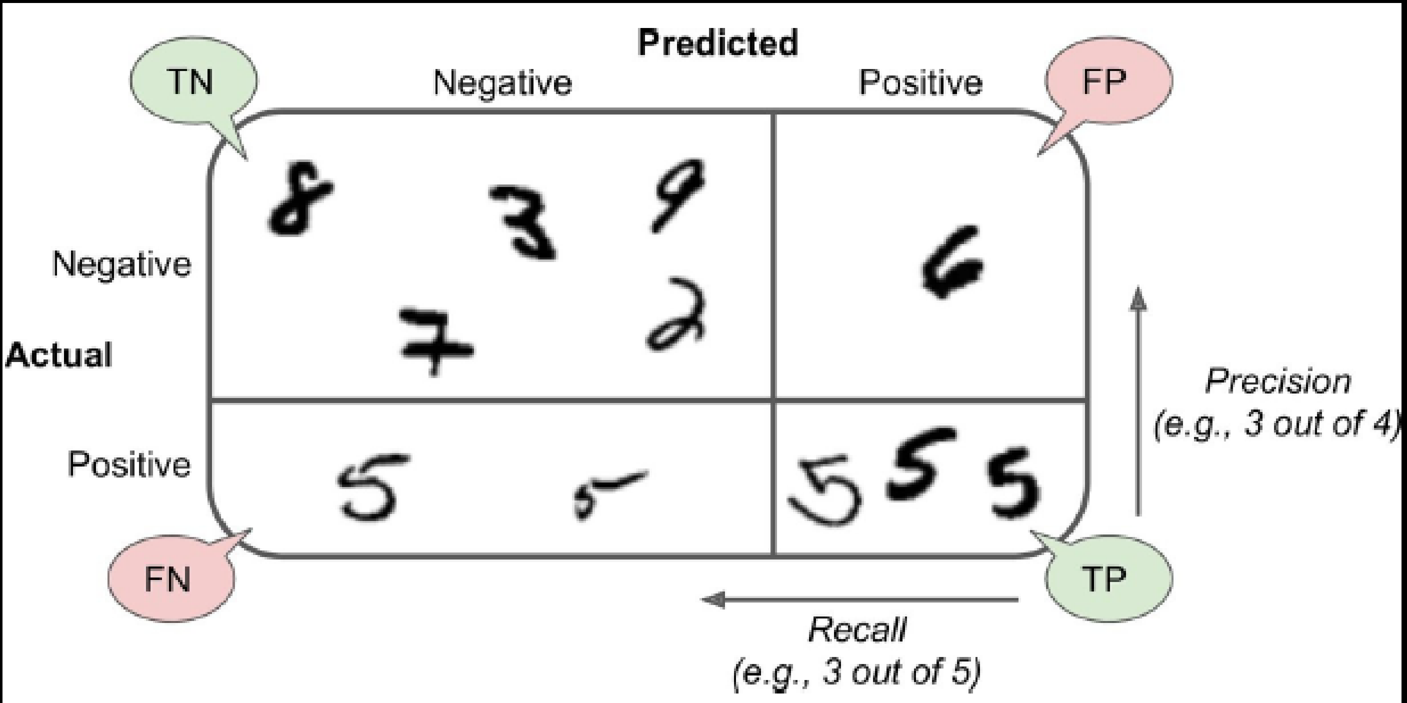


Figure 3-2. An illustrated confusion matrix

Classify digit "5"

Equation 3-1. Precision

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

Equation 3-2. Recall

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

F1 Score

$$F_1 = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{TP}{TP + \frac{FN + FP}{2}}$$

- F1 score is the harmonic mean of precision and recall
  - Arithmetic mean treats all values equally
  - Harmonic mean gives more weight to low values
- Good for balanced dataset
- High F1 score means recall and precision are high, but sometimes you care more about precision than recall (and vice versa)

Low Recall, High Precision

- Detecting videos that are safe for children
  - Reject many potentially good videos (low recall)
  - Keep only safe ones (high precision)

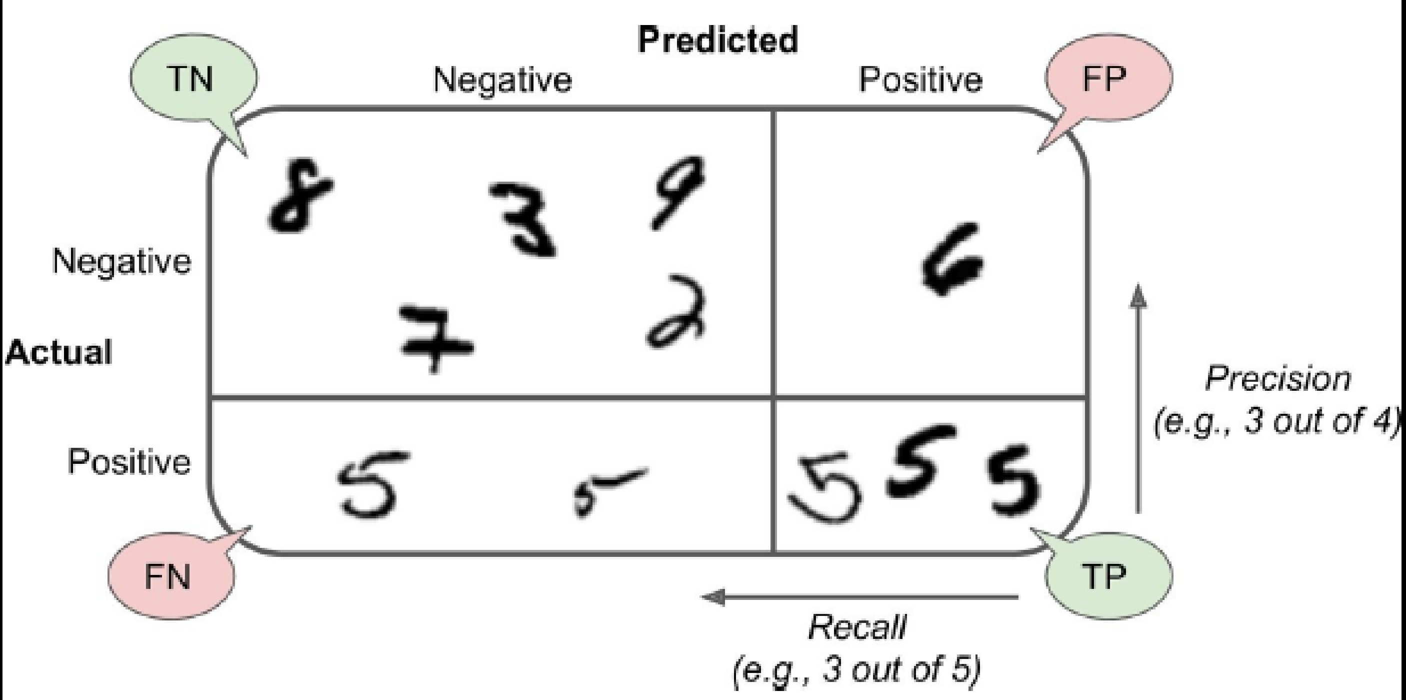


Figure 3-2. An illustrated confusion matrix

- Good video → positive class

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

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

High Recall, Low Precision

- Detecting shoplifter
  - Catch almost all shoplifter (high recall)
  - Deal with unhappy customers (low precision)

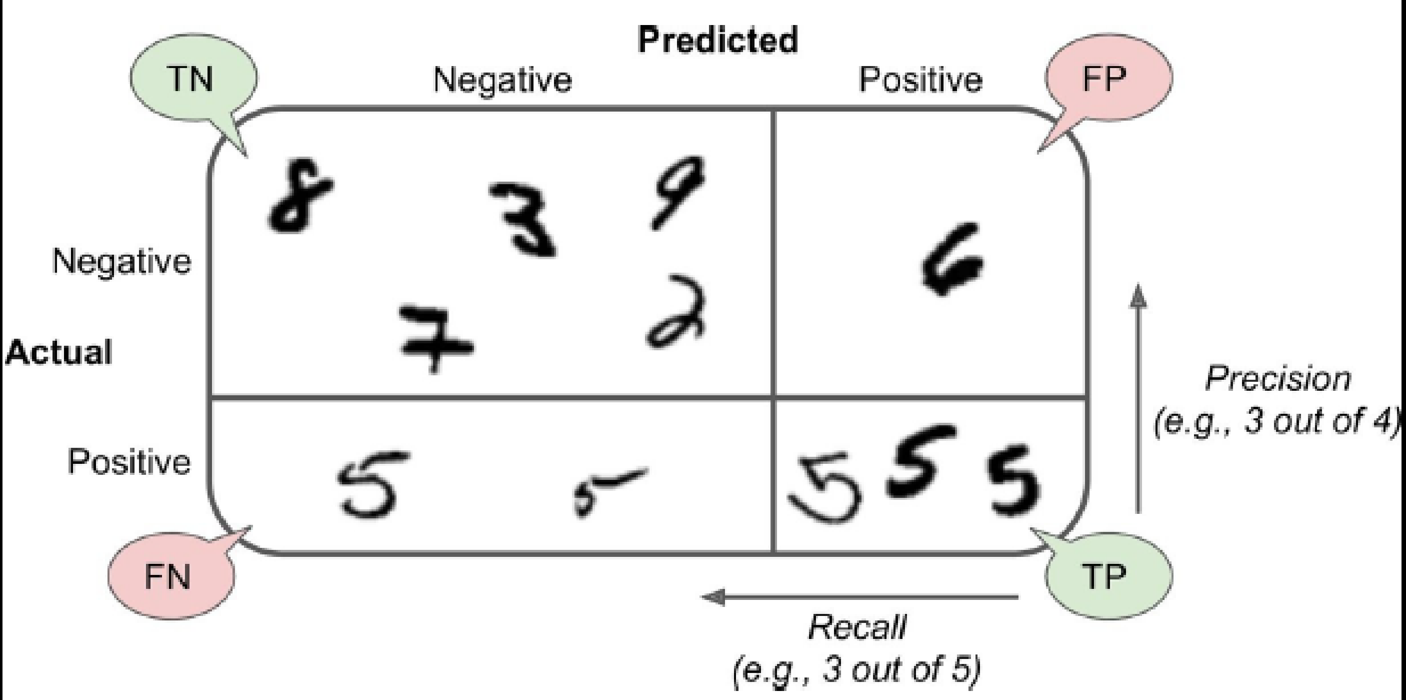


Figure 3-2. An illustrated confusion matrix

- Shoplifter → positive class

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

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

Precision/Recall Tradeoff

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

- Consider a decision function
  - If Score > Threshold then *Positive Class*
  - Else *Negative Class*

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

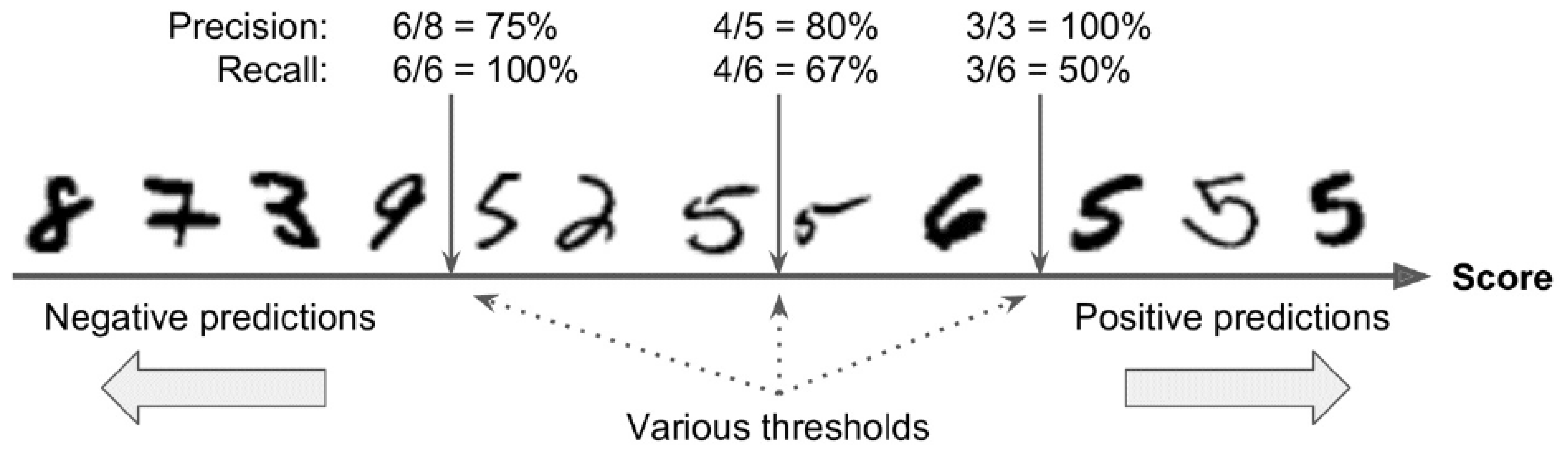


Figure 3-3. Decision threshold and precision/recall tradeoff

Precision/Recall Tradeoff

- High precision classifier is not very useful the recall is too low
- If someone says “we have 99% precision”, ask “what’s the recall?”

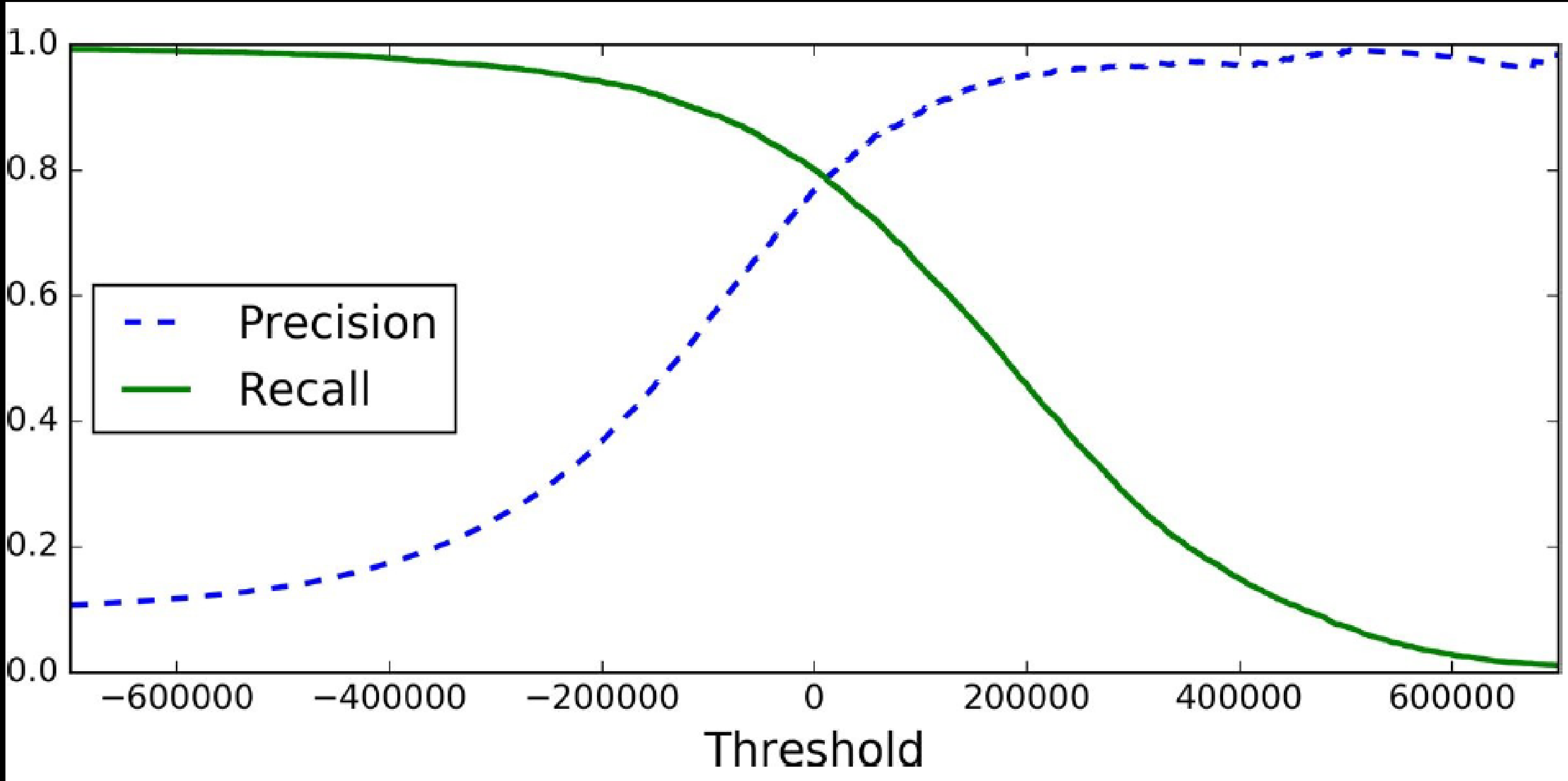


Figure 3-4. Precision and recall versus the decision threshold

How to choose the threshold?

- Find a threshold that produces results before the sharp drop of the precision vs recall function

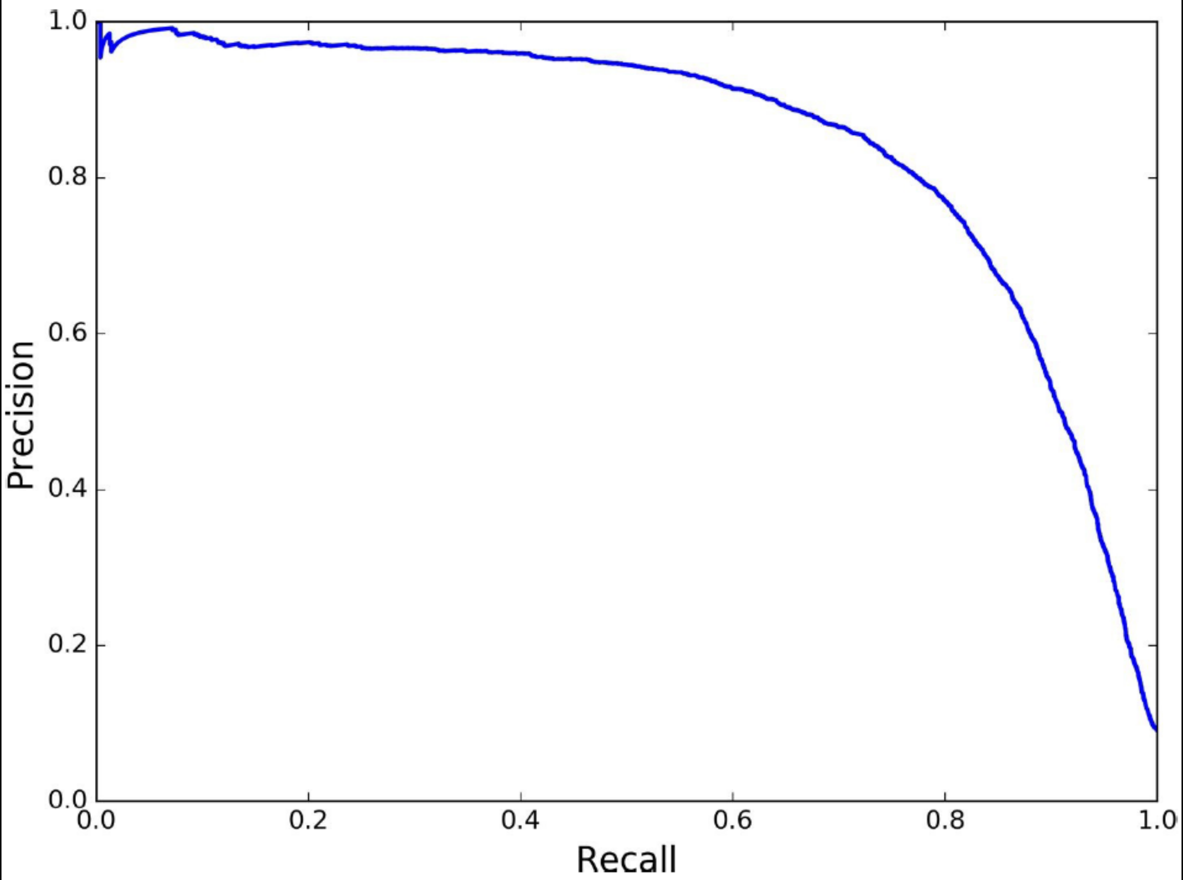


Figure 3-5. Precision versus recall

Receiver Operating Characteristic (ROC) Curve

- Plots True Positive Rate (*Recall*) vs False Positive Rate (*1-Specificity*)
  - Specificity = True Negative Rate
- Good classifier produces results closer to top left corner
  - Perfect classifier has AUC = 1
- Dotted line is random classifier
  - AUC = 0.5

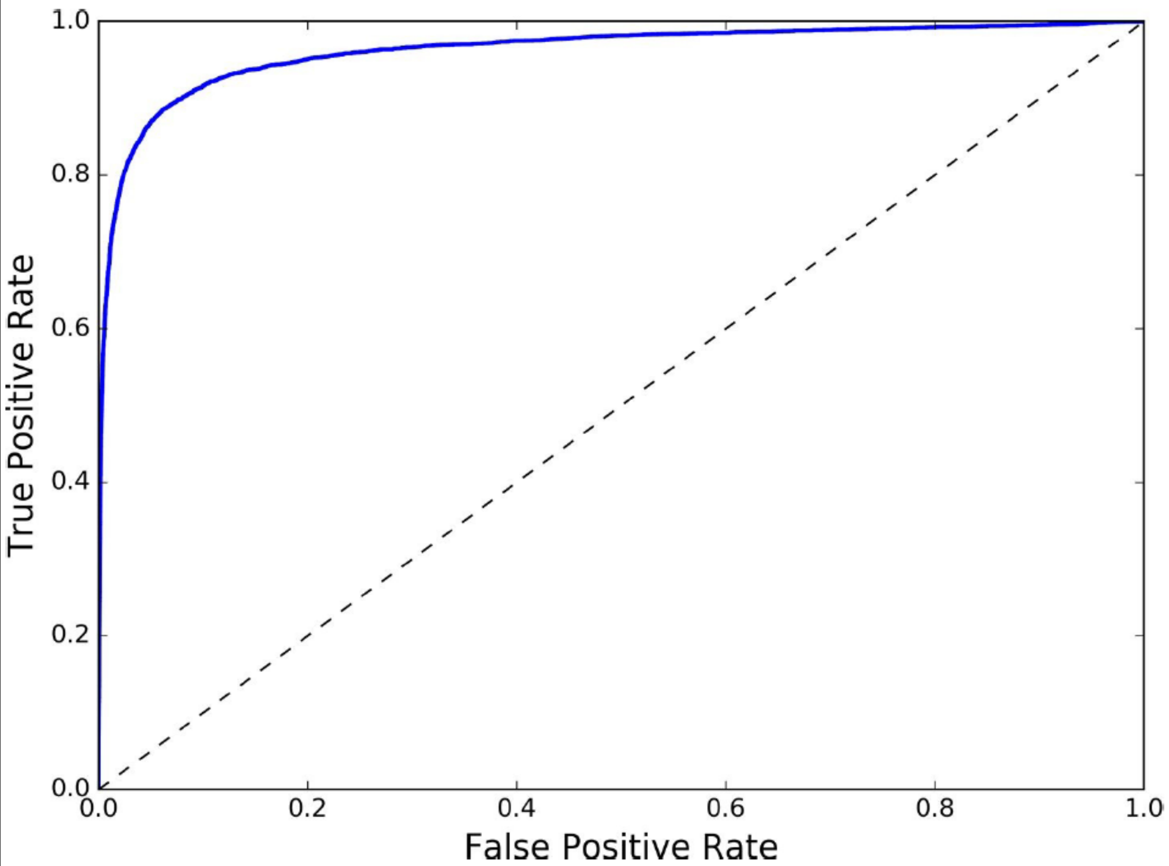


Figure 3-6. ROC curve

How to choose a model

- Train binary classifiers
- Evaluate using cross-validation
- Select precision/recall tradeoff
- Compare models using ROC curves & AUC scores

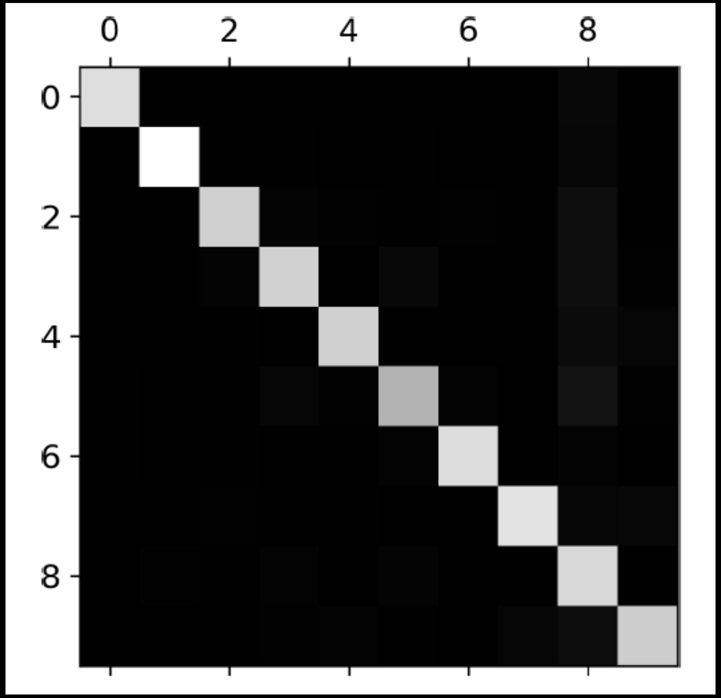
Multiclass Classification

- Binary classifiers (SVM) vs multiclass classifiers (Random Forest, Naïve Bayes)
- Multiple binary classifiers can be used for multiclass classification
- One-vs-all / one-vs-rest / one-against-rest
  - Train N classifiers, each to distinguish its own label from the remaining classes
    - N = number of classes
  - Apply new data to all N classifiers. The one that produces the largest output is chosen
- All-vs-all / one-vs-one / all-pairs
  - Train N x (N-1)/2 classifiers to distinguish each pairs of labels
  - Apply new data to all N classifiers. The final class label is determined by majority voting

Error Analysis
for MNIST Dataset with SGD Classifier

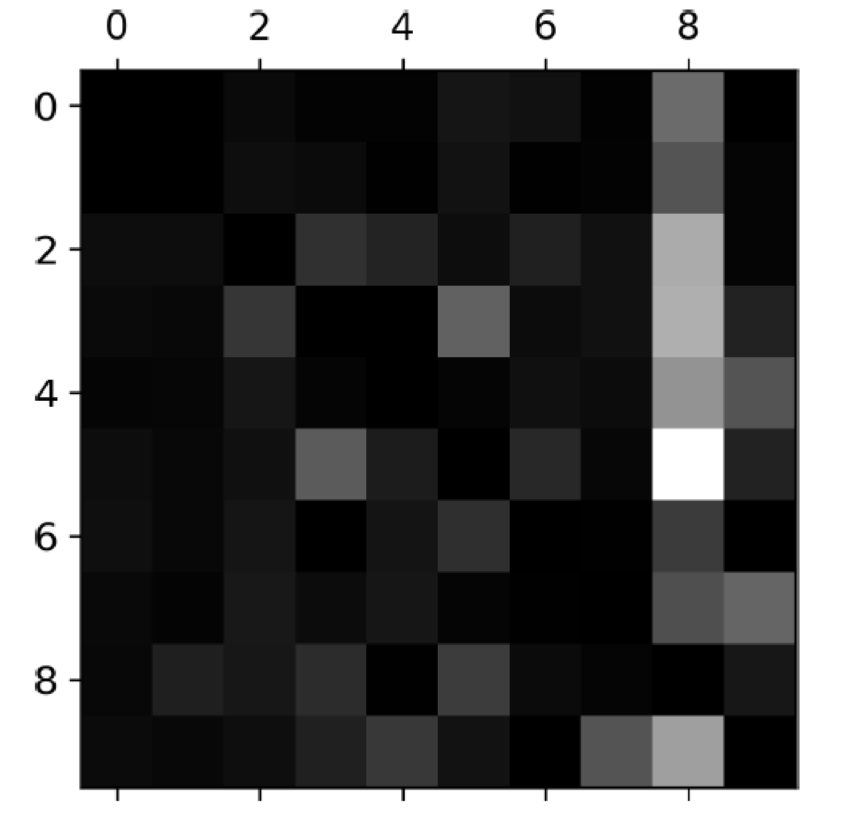
- Confusion matrix

```
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]])
```



“5” looks darker, smaller accuracy

- Fill the diagonal with zeros to keep only the errors.
- Column for class 8 is bright, but the row for class 8 is not
  - Many images misclassified as 8s
  - Actual 8s classified correctly
- Focus on reducing false 8’s
  - Find more training data for digits that look like 8s (but are not) to distinguish them from real 8s



Multilabel Classification

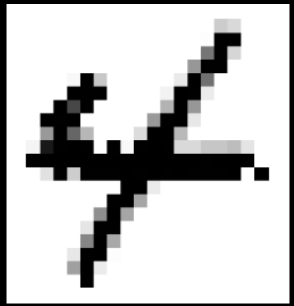
- Outputs multiple classes
- Classifier trained to recognize three faces (Alice, Bob, Charlie)
  - Output = [1 0 1] → Alice & Charlie on the picture but not Bob

Multioutput Classification

- aka Multioutput-Multiclass Classification
- Generalization of multilabel classification
  - Classifier's output is multilabel
  - Each label can be multiclass (can have more than 2 values)
- Classifier is trained to output a clean digit image from a noisy digit image
  - One label per pixel (clean or noisy) → multilabel
  - Each pixel can have multiple values (0-255) → multioutput

Noisy input image

Clean target image



Classifier output