FIN311人工智能及金融应用

Lecture 4: Classification

Instructor: Dr. Shijie (Kenny) Yang
Email: yangsj6@sustech.edu.cn

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Review and Preview

Introduction to Artificial Intelligence (Weeks 1-2)

1

- Fundamentals of Artificial Intelligence
- Python Basics and Some Technical Background
- End-to-End Machine Learning Project

Machine Learning (Weeks 3-6)

2

- Classification
- Training Models
- Support Vector Machines
- Decision Trees
- Ensemble Learning and Random Forests
- Dimensionality Reduction
- Unsupervised Learning Techniques

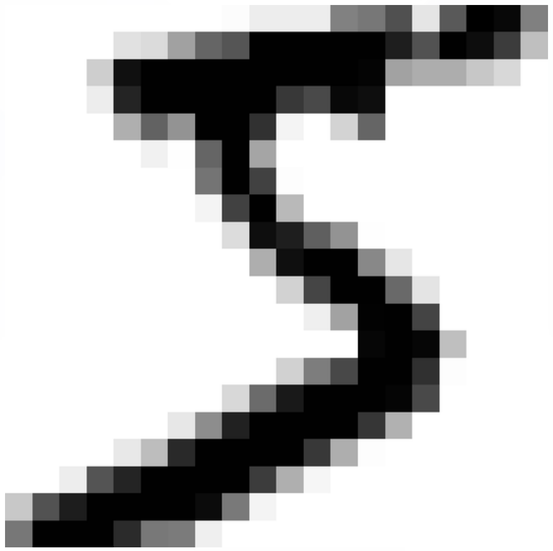
MNIST digits classification dataset

- The **MNIST** database (Modified National Institute of Standards and Technology database)
  - is a large database of handwritten digits that is commonly used for training various image processing systems
  - is also widely used for training and testing in the field of machine learning
- The set of images in the MNIST database was created in 1994
- The MNIST database contains
  - 60,000 training images
  - 10,000 testing images
- http://yann.lecun.com/exdb/mnist/

MNIST image

Digits from the MNIST dataset

- MNIST has 70,000 images
  - each image has 784 features
  - each image is 28 × 28 pixels
  - each feature simply represents one pixel's intensity, from 0 (white) to 255 (black).



Example of an MNIST image

- Let's grab an instance's feature vector, reshape it to a 28 × 28 array, and display it using Matplotlib's imshow() function

Training a Binary Classifier

- Let's simplify the problem and only try to identify one digit
- For example, the number 5
  - This "**5-detector**" will be an example of a **binary classifier**
  - capable of distinguishing between just two classes, 5 and non-5
- Select a simple classifier and train it
  - **Stochastic gradient descent** (SGD, or stochastic GD) **classifier**
  - This classifier is capable of handling very large datasets efficiently
  - SGD deals with training instances independently, one at a time, which also makes SGD well suited for online learning

```
from sklearn.linear_model import SGDClassifier
sgd_clf = SGDClassifier(random_state=42)
sgd_clf.fit(X_train, y_train_5)
```

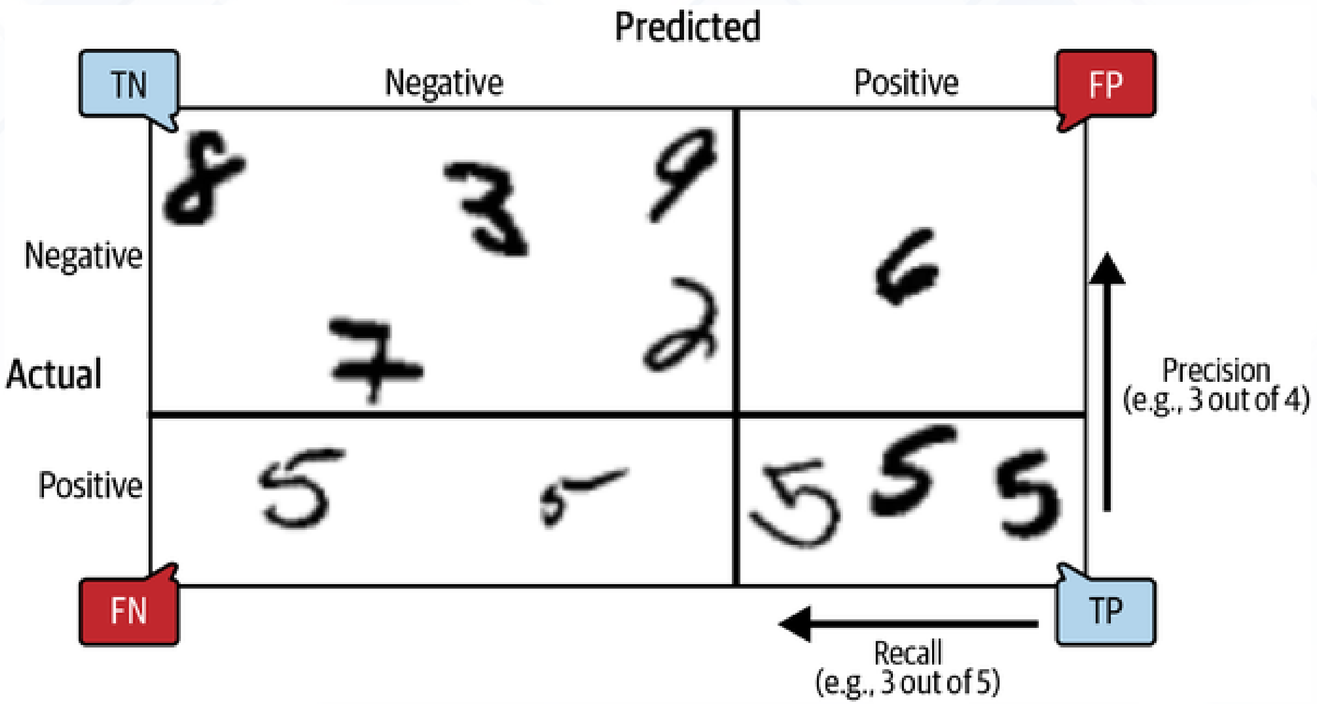
Performance Measures

- Evaluating a classifier is often significantly trickier than evaluating a regressor
- New concepts and acronyms
  - Measuring Accuracy Using Cross-Validation
  - Confusion Matrices
  - Precision and Recall, F1 Score
  - The Precision/Recall Trade-off
  - The ROC Curve

A Confusion Matrix

		Predicted	
		Negative **(N)** **-**	Positive **(P)** **+**
Actual	Negative **-**	True Negative **(TN)**	**False Positive (FP)** **Type I Error**
	Positive **+**	**False Negative (FN)** **Type II Error**	True Positive **(TP)**

Performance Measures

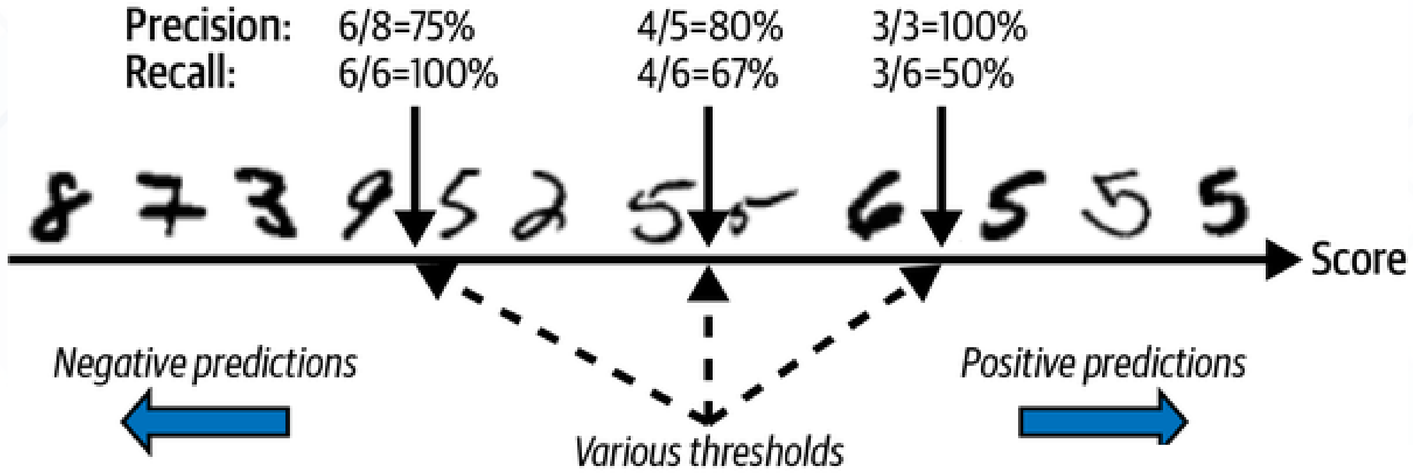


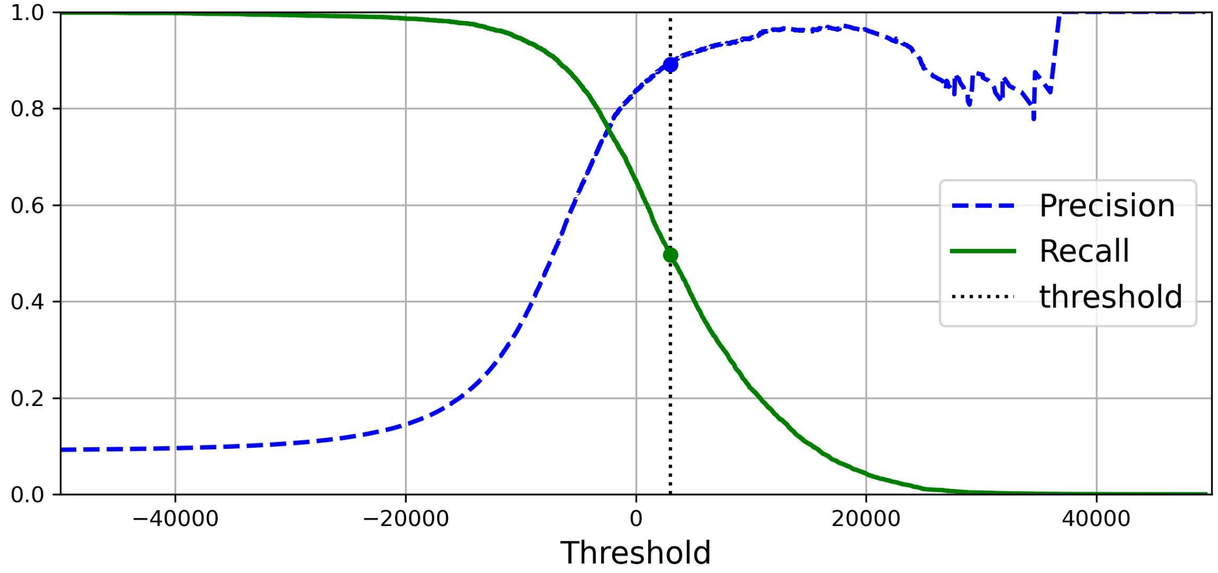
$$\text{precision} = \frac{TP}{TP + FP}$$

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

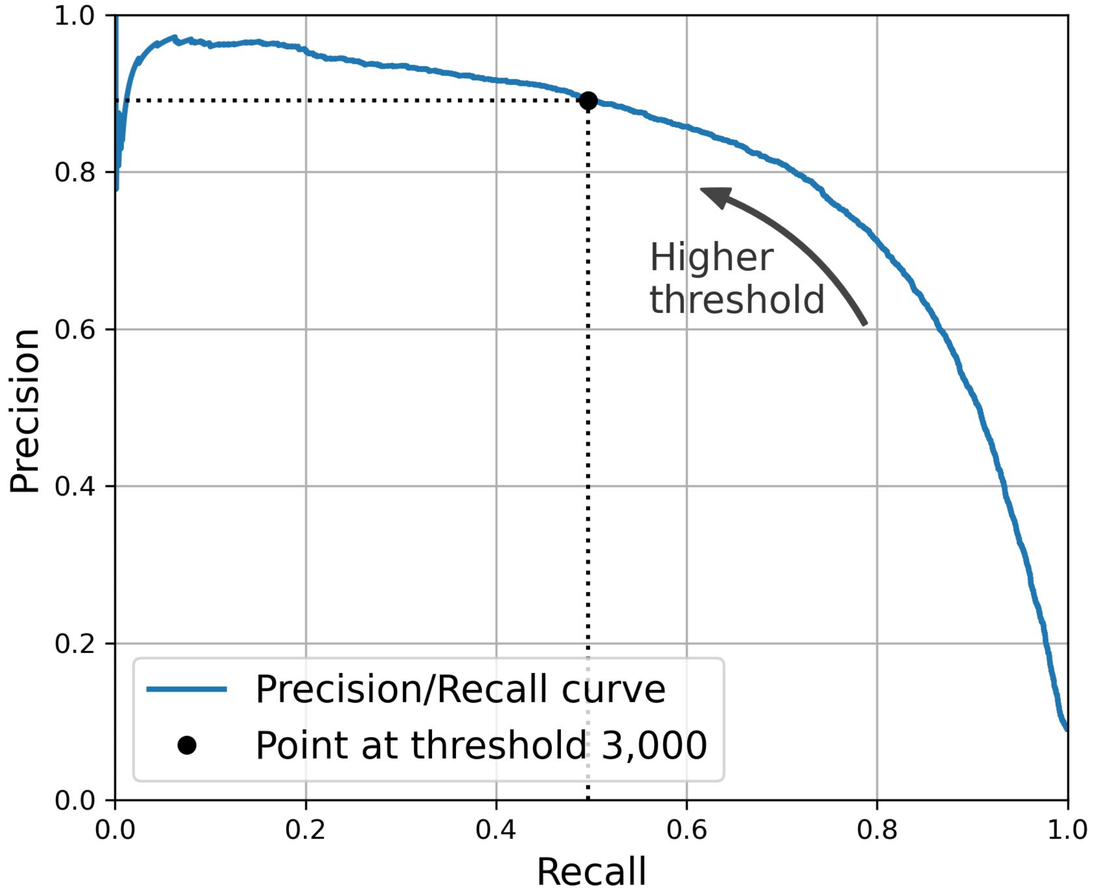
$$F_1 = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{TP}{TP + \frac{FN+FP}{2}}$$

The Precision/Recall Trade-off

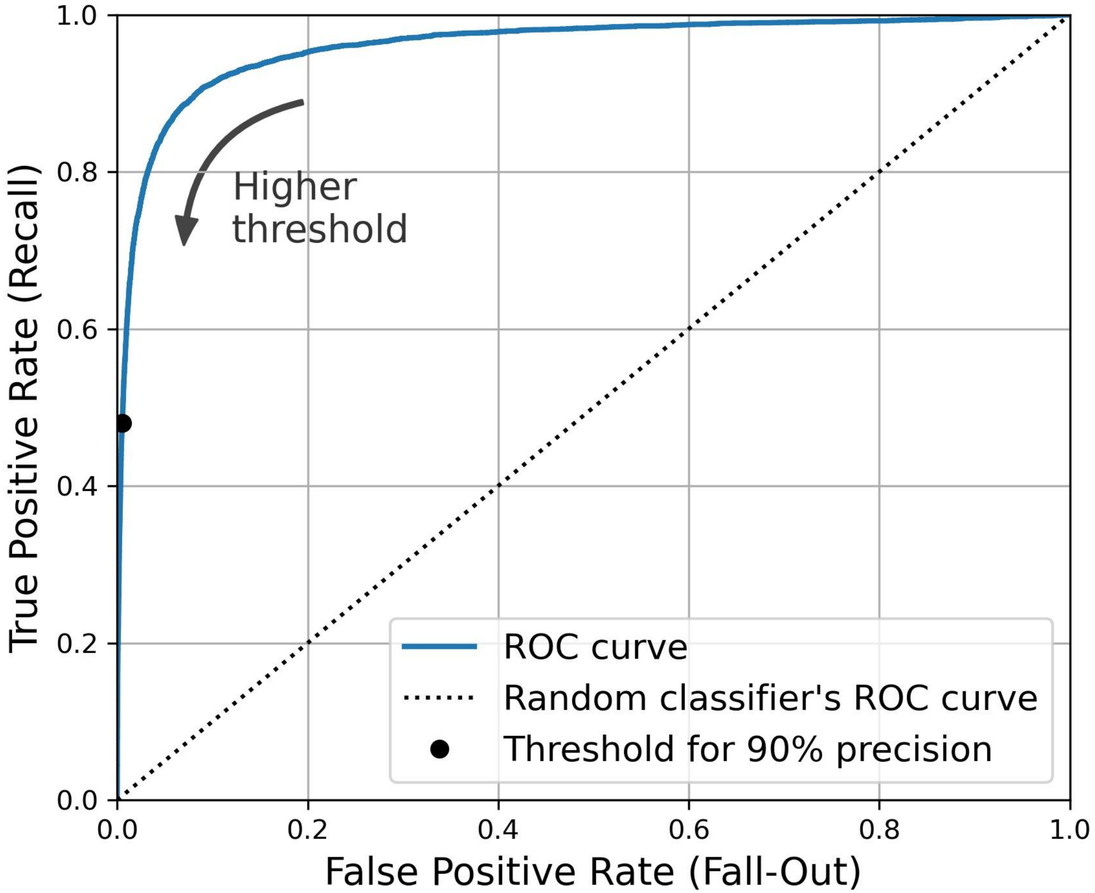




The Precision/Recall Trade-off



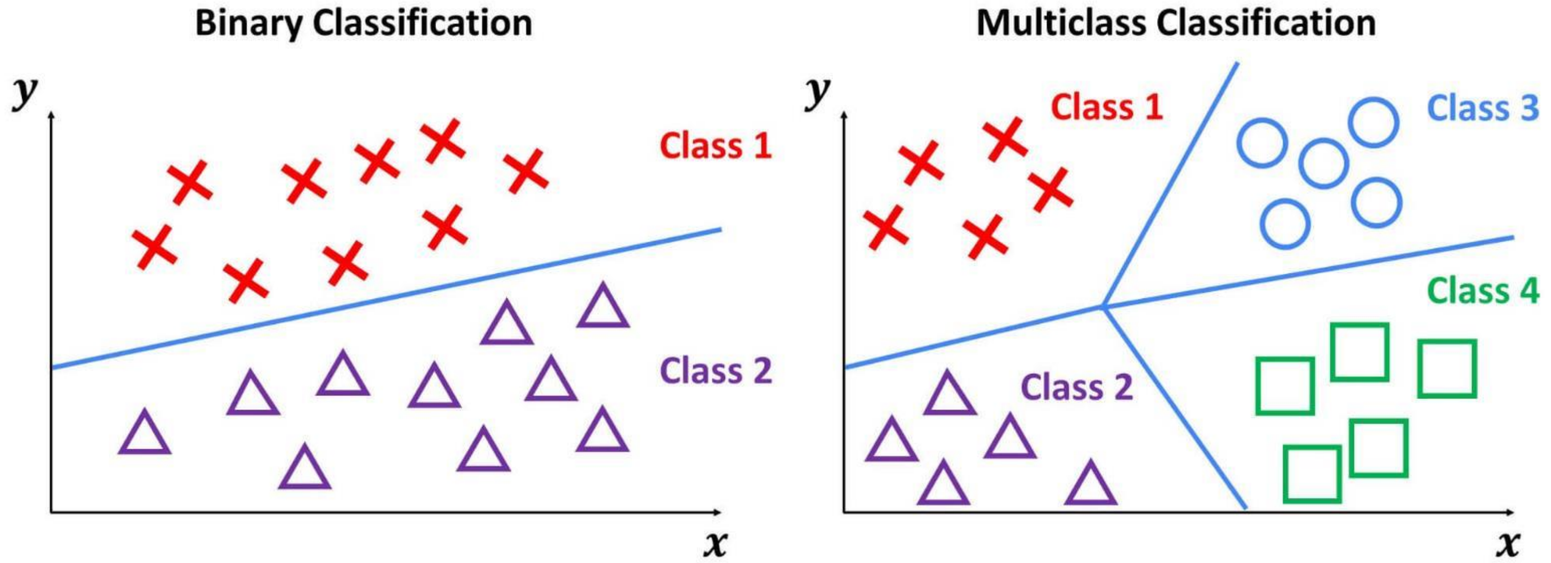
the precision/recall curve



the ROC curve

Multiclass Classification

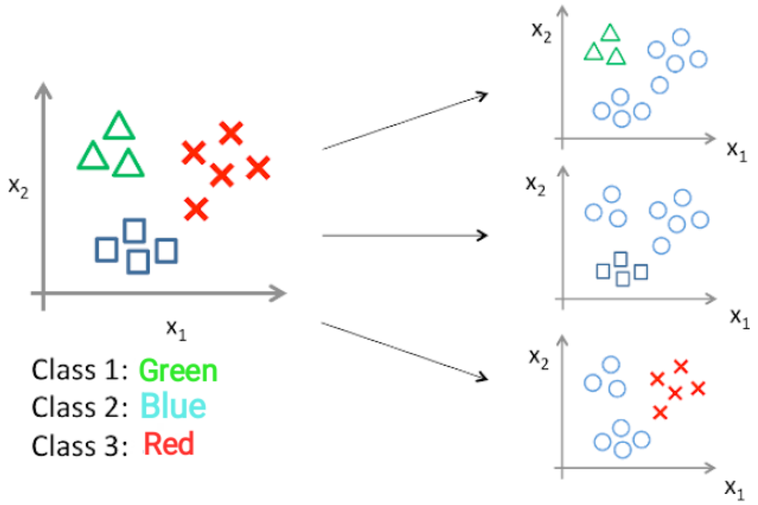
- **Binary classifiers** can distinguish between two classes
  - Examples: SGDClassifier and SVC
- **Multiclass classifiers** (also called multinomial classifiers) can distinguish between more than two classes
  - Examples: LogisticRegression, RandomForestClassifier, and GaussianNB



Multiclass Classification

- There are various strategies that you can use to perform multiclass classification with **multiple binary classifiers**

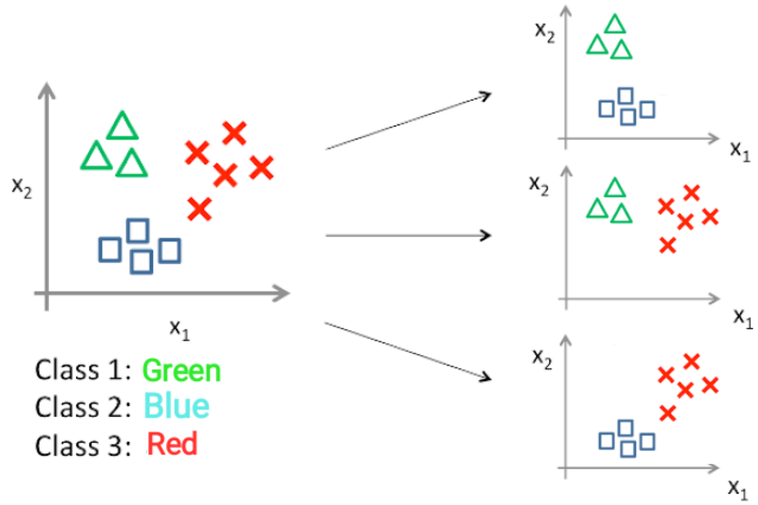
one-versus-the-rest (OvR) strategy
(or one-versus-all (OvA))



Classifier 1: [Green] vs [Red, Blue]
Classifier 2: [Blue] vs [Green, Red]
Classifier 3: [Red] vs [Blue, Green]

} N binary classifiers

one-versus-one (OvO) strategy

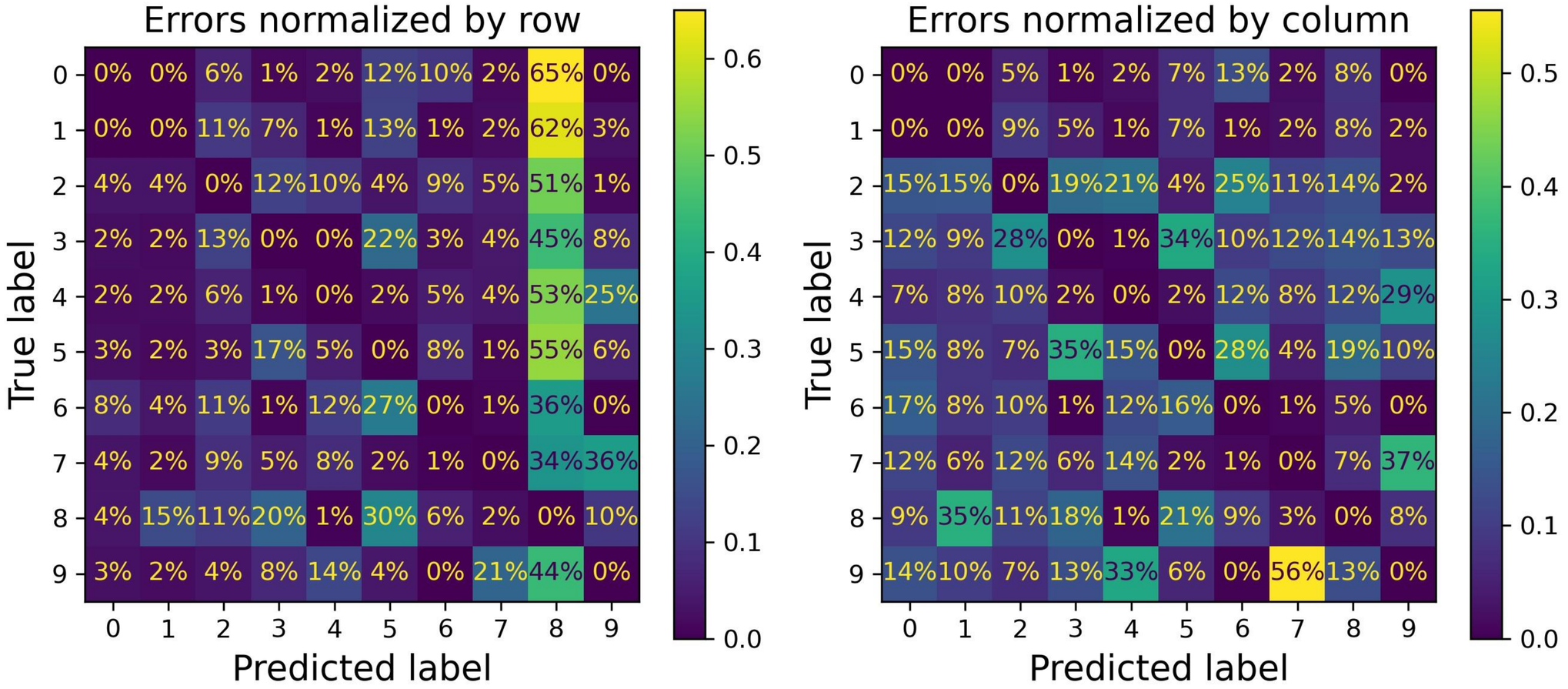


Classifier 1: Green vs. Blue
Classifier 2: Green vs. Red
Classifier 3: Blue vs. Red

} N* (N-1)/2 binary classifiers

Error Analysis

- Confusion matrix with errors only, normalized by row (left) and by column (right)



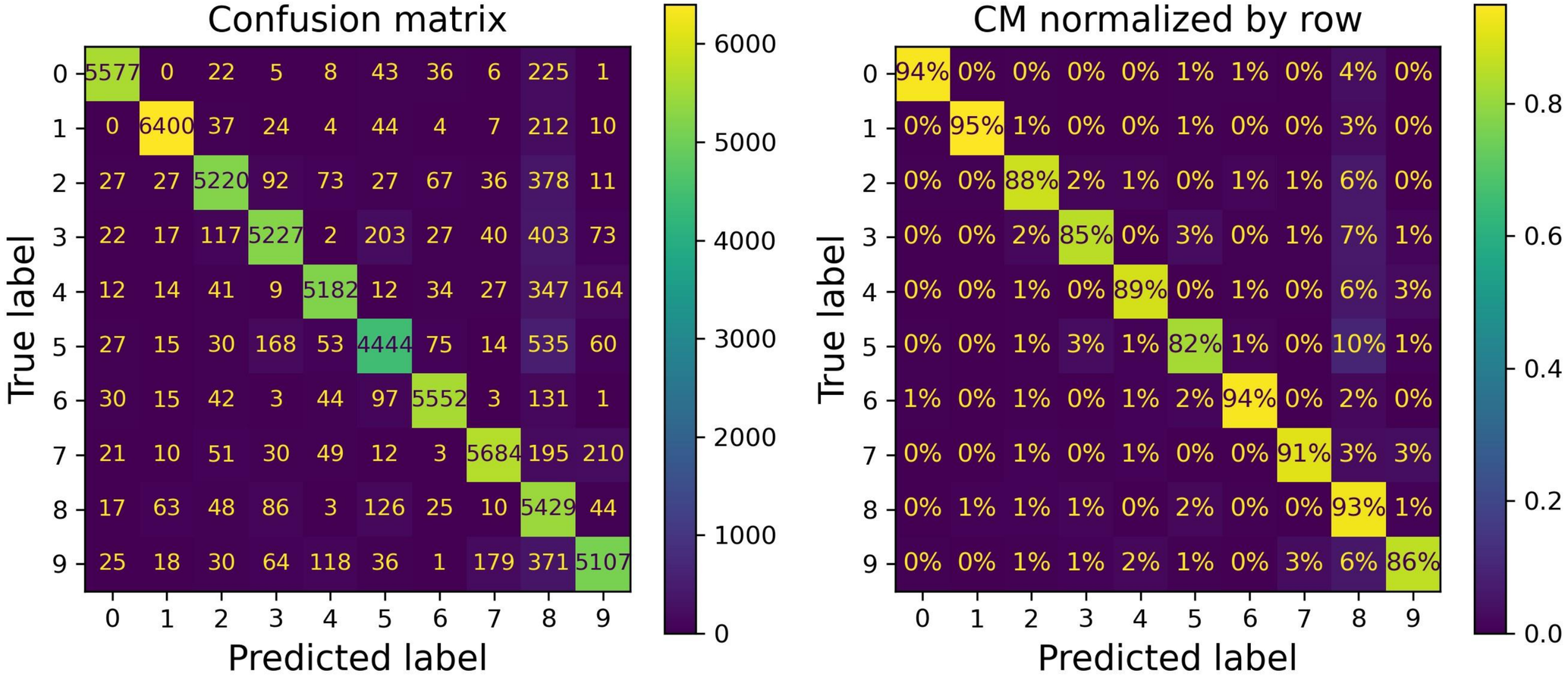
Multiclass Classification

- Scikit-Learn detects when you try to use a binary classification algorithm for a multiclass classification task, and it automatically runs OvR or OvO, depending on the algorithm
  - Some algorithms (such as support vector machine classifiers) scale poorly with the size of the training set. For these algorithms OvO is preferred because it is faster to train many classifiers on small training sets than to train few classifiers on large training sets.
  - For most binary classification algorithms, however, OvR is preferred.
- Below is an example running a support vector machine classifier using the sklearn.svm.SVC class. By default, it uses OvO

```
from sklearn.svm import SVC
svm_clf = SVC(random_state=42)
svm_clf.fit(X_train[:2000], y_train[:2000])
```

Error Analysis

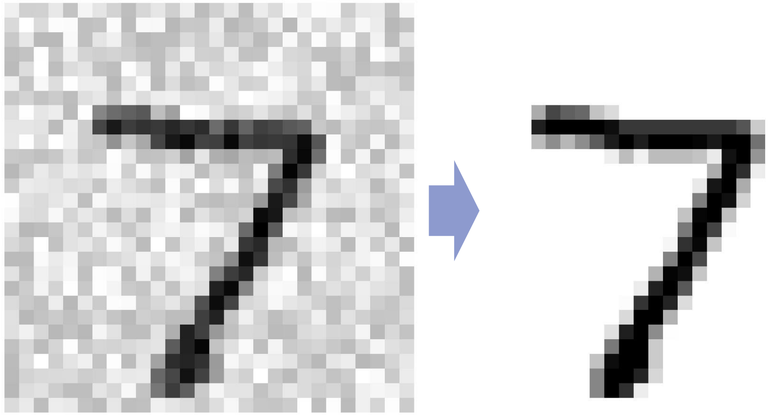
- Confusion matrix (left) and the same CM normalized by row (right)



Other topics

- Multilabel Classification
  - A classification system that outputs **multiple binary tags**
  - KNeighborsClassifier
  - ChainClassifier

- Multioutput Classification
  - A generalization of multilabel classification where **each label can be multiclass** (i.e., it can have more than two possible values)



Glossary

- Binary classifier (二分类器)
- Multiclass classifiers (多类分类器)
- Stochastic gradient descent classifier (随机梯度下降分类器)
- Accuracy (准确度)
- Confusion matrix (混淆矩阵)
- True negatives (真负例)
- False positives / Type I errors (假正例/第一类错误)
- False negatives / Type II errors (假负例/第二类错误)
- True positives (真正例)
- Precision (精确率/查准率)
- Recall (召回率/查全率)
- F1 Score (F1分数)
- Decision threshold (决策阈值)
- Precision/Recall Trade-off (精确率/召回率折衷)
- ROC Curve (受试者工作特征曲线曲线)
- Area under the curve (AUC) (ROC曲线下面积)
- One-versus-the-rest (OvR) strategy (一对多策略)
- One-versus-one (OvO) strategy (一对一策略)
- Support vector machine classifier (支持向量机分类器)
- Multilabel classification (多标签分类)
- Multioutput classification (多输出分类)