Chapter 3. Classification

MNIST

Shuffle the training set: 1. guarantee that all cross-validation folds will be similar; 2. some learning algorithms are sensitive to the order of the training instances, and they perform poorly if they get many similar instances in a row.

Training a Binary Classifier

Stochastic Gradient Descent_ (SGD) classifier: the true gradient of is approximated by a gradient at a single example.

Advantage: capable of handling very large datasets efficiently.

This is in part because SGD deals with training instances independently, one at a time (suitable for *online learning*)

Performance Measures

Measuring Accuracy Using Cross-Validation

K-fold cross-validation means splitting the training set into K-folds, then making predictions and evaluating them on each fold using a model trained on the remaining folds.

Confusion Matrix

Each row in a confusion matrix represents an *actual class*, while each column represents a *predicted class*.

true negatives (TN) false positives (FP) false negatives (FN) true positives (TP)

Precision, positive predictive value (PPV)

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

Recall, true positive rate (TPR):

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

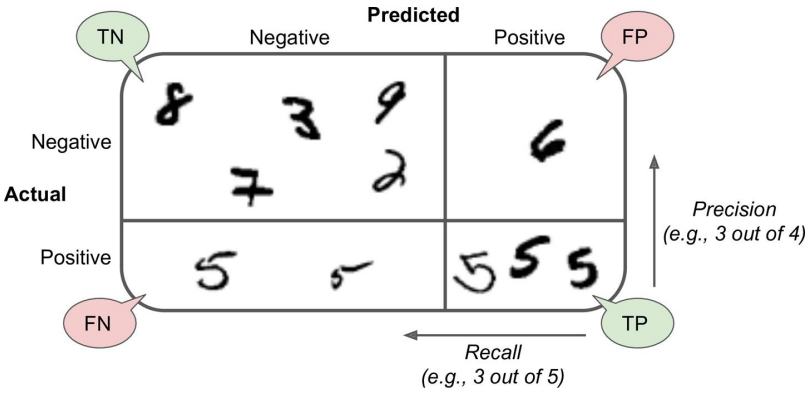

Figure 3-2. An illustrated confusion matrix

F-1 score is the harmonic mean of precision and recall:

$$F_1 = 2\frac{precision \cdot recall}{precision + recall}$$

F-measure:

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

The F_1 score favors classifiers that have similar precision and recall.Two other commonly used F measures are the F_2 measure, which weighs recall higher than precision (by placing more emphasis on false negatives), and the $F_{0.5}$ measure, which weighs recall lower than precision (by attenuating the influence of false negatives).

Precision/Recall Tradeoff

Classification, for each instance, it computes a score based on a *decision function*, and if that score is greater than a threshold, it assigns the instance to the positive class, or else it assigns it to the negative class.

Lowering the threshold increases recall and reduces precision.

Method 1 to decide threshold: precision and recall as functions of the threshold value.

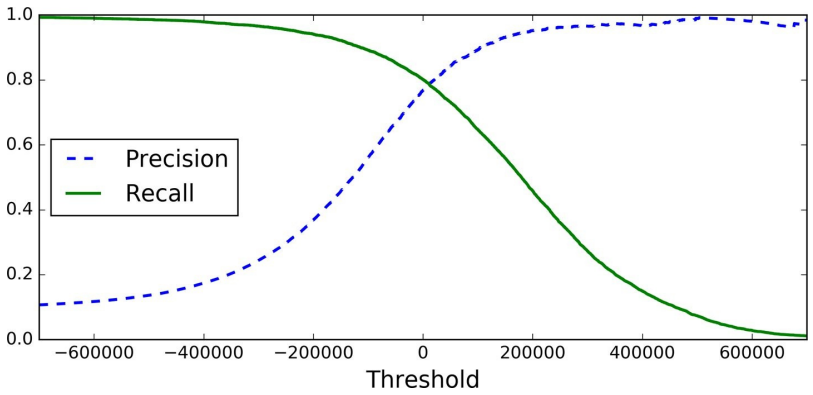

Figure 3-4. Precision and recall versus the decision threshold

Method 2: plot precision directly against recall.

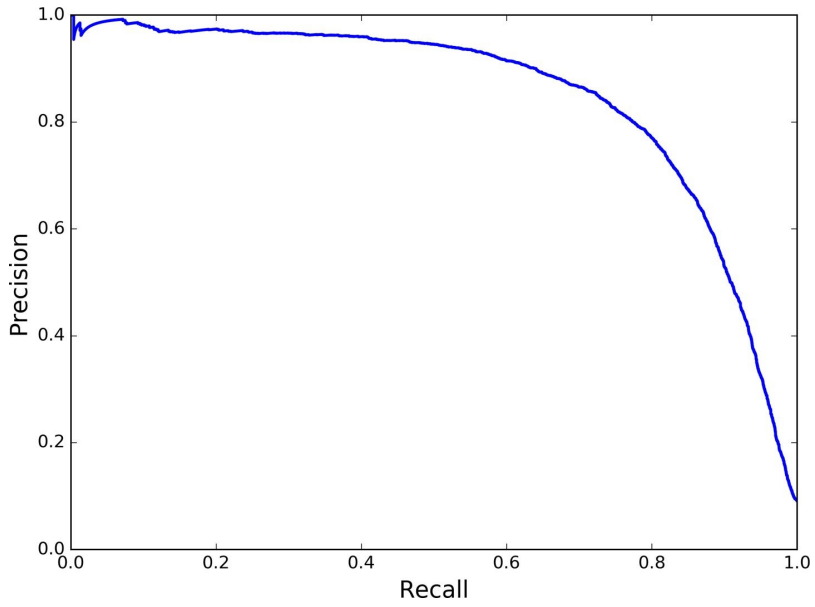

Figure 3-5. Precision versus recall

The ROC Curve

The *receiver operating characteristic* (ROC) curve is another common tool used with binary classifiers. It plots the *true positive rate* (TPR, sensitivity, recall) against the *false positive rate* (FPR), *sensitivity (recall)* versus *1 – specificity*.

$FPR = 1 - TNR$: the ratio of negative instances that are correctly classified as negative. TNR: the ratio of negative instances that are correctly classified as negative. The TNR is also called *specificity*.

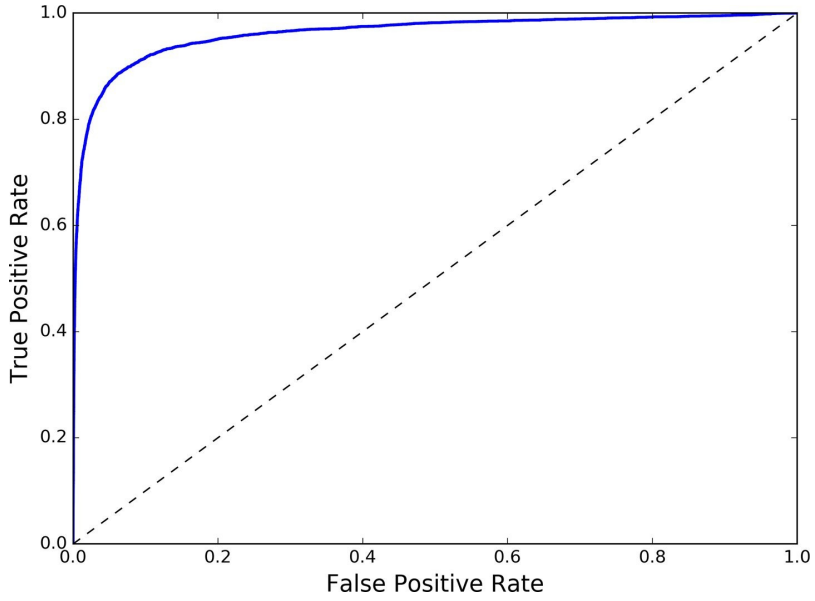

Figure 3-6. ROC curve

One way to compare classifiers is to measure the *area under the curve* (AUC).

TIP

> Prefer the PR curve whenever the positive class is rare or when you care more about the false positives than the false negatives, and the ROC curve otherwise.

Multiclass Classification

Multiclass classification algorithm: Random Forest, naive Bayes.

Binary classifier: Support Vector Machine (SVM), linear classifier.

Strategies that perform multiclass classification using multiple binary classifiers:

- one-versus-all (OvA) strategy (one-versus-the-rest).
- one-versus-one (OvO) strategy (advantage: each classifier only needs to be trained on the part of the training set for the two classes that it must distinguish.)

OvO is preferred for algorithms (SVM) scale poorly with the size of the training set. For most binary classification algorithms, however, OvA is preferred.

Error Analysis

First, plot confusion matrix. Divide each value in the confusion matrix by the number of images in the corresponding class, so you can compare error rates instead of absolute number.

Analyze individual errors.

Multilabel Classification

When measure F_1 score for each individual label, give each label a weight equal to its *support* (i.e., the number of instances with that target label).

Multioutput Classification

A generalization of multilabel classification where each label can be multiclass.