Multiclass classification evaluation

Multiple Classes – One vs. One

- With k classes confusion matrix becomes a k x k matrix
- No clear notion of positives and negatives.

		Ground Truth			
		Class A	**Class B**	**Class C**	**Class D**
Prediction	**Class A**	Correct	Wrong	Wrong	Wrong
	Class B	Wrong	Correct	Wrong	Wrong
	Class C	Wrong	Wrong	Correct	Wrong
	Class D	Wrong	Wrong	Wrong	Corrent

Multiple Classes – One vs. All

- Choose one of k classes as positive
- Combine all other classes into negative to obtain k different binary confusion matrices

		Ground Truth	
		Class A	**Other**
Pred.	**Class A**	True positive	False positive
	Other	False negative	True negative

- Combine the results for each class with the microaverage or macroaverage techniques

Macroaveraging vs. microaveraging

	label	prediction
	apple	orange
	orange	orange
	apple	apple
	banana	pineapple
	banana	banana
	pineapple	pineapple

microaveraging: average over examples

macroaveraging: calculate evaluation score (e.g. accuracy) for each label, then average over labels

?

Exempley by:
David Kauchak

Macroaveraging vs. microaveraging

	label	prediction
	apple	orange
	orange	orange
	apple	apple
	banana	pineapple
	banana	banana
	pineapple	pineapple

microaveraging: 4/6

macroaveraging:

apple = 1/2

orange = 1/1

banana = 1/2

pineapple = 1/1

total = (1/2 + 1 + 1/2 + 1)/4

= 3/4

Exempley by:
David Kauchak

average : *{'binary', 'micro', 'macro', 'samples','weighted'}, default=None*

If `None`, the scores for each class are returned. Otherwise, this determines the type of averaging performed on the data:

`'binary'`:
Only report results for the class specified by `pos_label`. This is applicable only if targets (`y_{true,pred}`) are binary.

`'micro'`:
Calculate metrics globally by counting the total true positives, false negatives and false positives.

`'macro'`:
Calculate metrics for each label, and find their unweighted mean. This does not take label imbalance into account.

`'weighted'`:
Calculate metrics for each label, and find their average weighted by support (the number of true instances for each label). This alters 'macro' to account for label imbalance; it can result in an F-score that is not between precision and recall.

`'samples'`:
Calculate metrics for each instance, and find their average (only meaningful for multilabel classification where this differs from `accuracy_score`).

Source: Scikit learn
sklearn.metrics.precision_recall_fscore_support

Multiple class-classification

Two ways to compute multi-class accuracy

- Consider the TN
  - microaverage = macroaverage

- Without the TN

	ant	bird	cat
ant	2	0	0
bird	0	0	1
cat	1	0	2