Deep Learning for Natural Language Processing

Evaluation of text classifiers

Marco Kuhlmann
Department of Computer and Information Science

LINKÖPING UNIVERSITY

Evaluation of text classifiers

Eisenstein § 4.4

- We require a **test set** consisting of a number of documents, each of which has been tagged with its correct class.

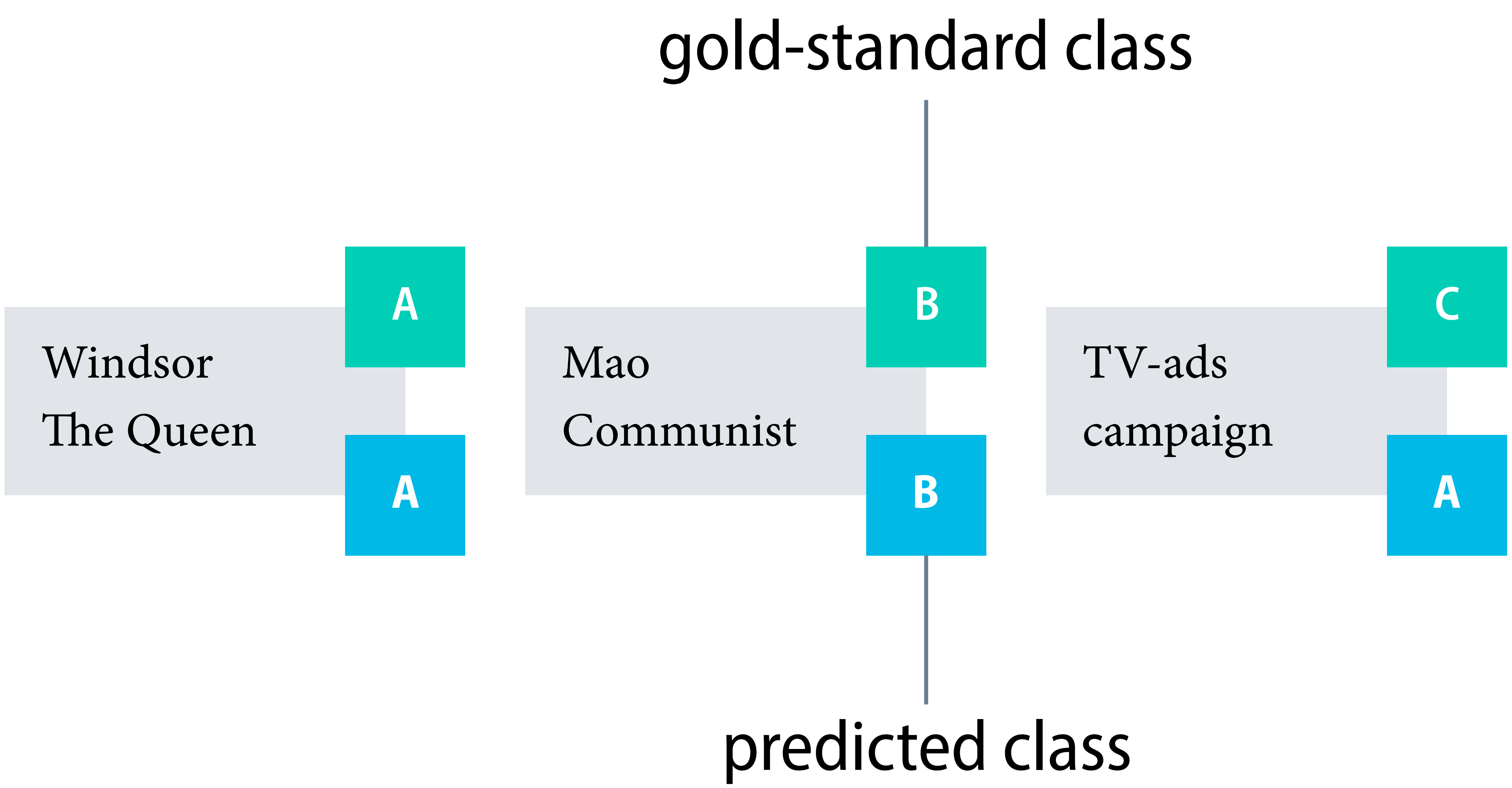
  typically part of a larger gold-standard data set

- To evaluate a classifier, we apply it to the test set and compare the predicted classes with the gold-standard classes.

- The result of this comparison allows us to estimate how well the classifier will perform on new, previously unseen documents.

  assume that all samples are drawn from the same distribution

Evaluation of text classifiers



Accuracy

The **accuracy** of a classifier is the proportion of documents for which the classifier predicts the gold-standard class:

$$\text{accuracy} = \frac{\text{number of correctly classified documents}}{\text{number of all documents}}$$

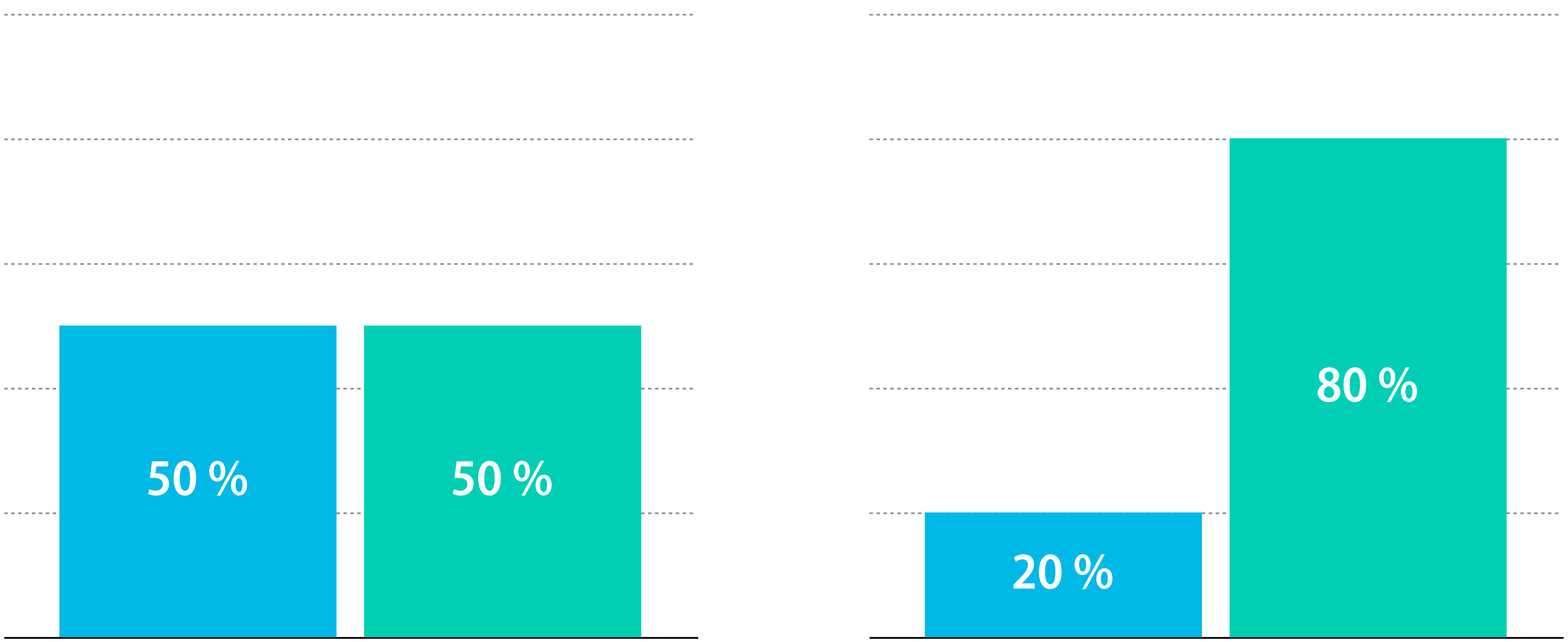
Accuracy

Document	Gold-standard class	Predicted class
Chinese Beijing Chinese	China	China
Chinese Chinese Shanghai	China	China
Chinese Macao	China	China
Tokyo Japan Chinese	Japan	China

accuracy for this example: 3/4 = 75%

Accuracy and imbalanced data sets

Is 80% accuracy good or bad?



The role of baselines

- Evaluation measures are no absolute measures of performance.

  Whether '80% accuracy' is good or not depends on the task at hand.

- Instead, we should ask for a classifier's performance relative to other classifiers, or other points of comparison.

  'The softmax classifier has a higher accuracy than the perceptron classifier.'

- When other classifiers are not available, a simple baseline is to always predict the **most frequent class** in the training data.

  alternative: random sampling from the class distribution in the training set

Confusion matrix

	classifier ‘positive’	classifier ‘negative’
gold standard ‘positive’	true positives	false negatives
gold standard ‘negative’	false positives	true negatives

Accuracy

	classifier 'positive'	classifier 'negative'
gold standard 'positive'	true positives	false negatives
gold standard 'negative'	false positives	true negatives

Precision and recall

Eisenstein § 4.4.1

- **Precision** and **recall** ‘zoom in’ on how good a system is at identifying documents of a specific class *k*.
- **Precision** is the proportion of correctly classified documents among all documents for which the system predicts class *k*.

  When the system predicts ‘positive’, how often is it correct?
- **Recall** is the proportion of correctly classified documents among all documents with gold-standard class *k*.

  If the movie review is ‘positive’, how often does the system predict it?

Precision with respect to the positive class

	classifier 'positive'	classifier 'negative'
gold standard 'positive'	true positives	false negatives
gold standard 'negative'	false positives	true negatives

Recall with respect to the positive class

	classifier 'positive'	classifier 'negative'
gold standard 'positive'	true positives	false negatives
gold standard 'negative'	false positives	true negatives

Precision and recall with respect to the positive class

$$\text{precision} = \frac{\#\text{ true positives}}{\#\text{ true positives} + \#\text{ false positives}}$$

$$\text{recall} = \frac{\#\text{ true positives}}{\#\text{ true positives} + \#\text{ false negatives}}$$

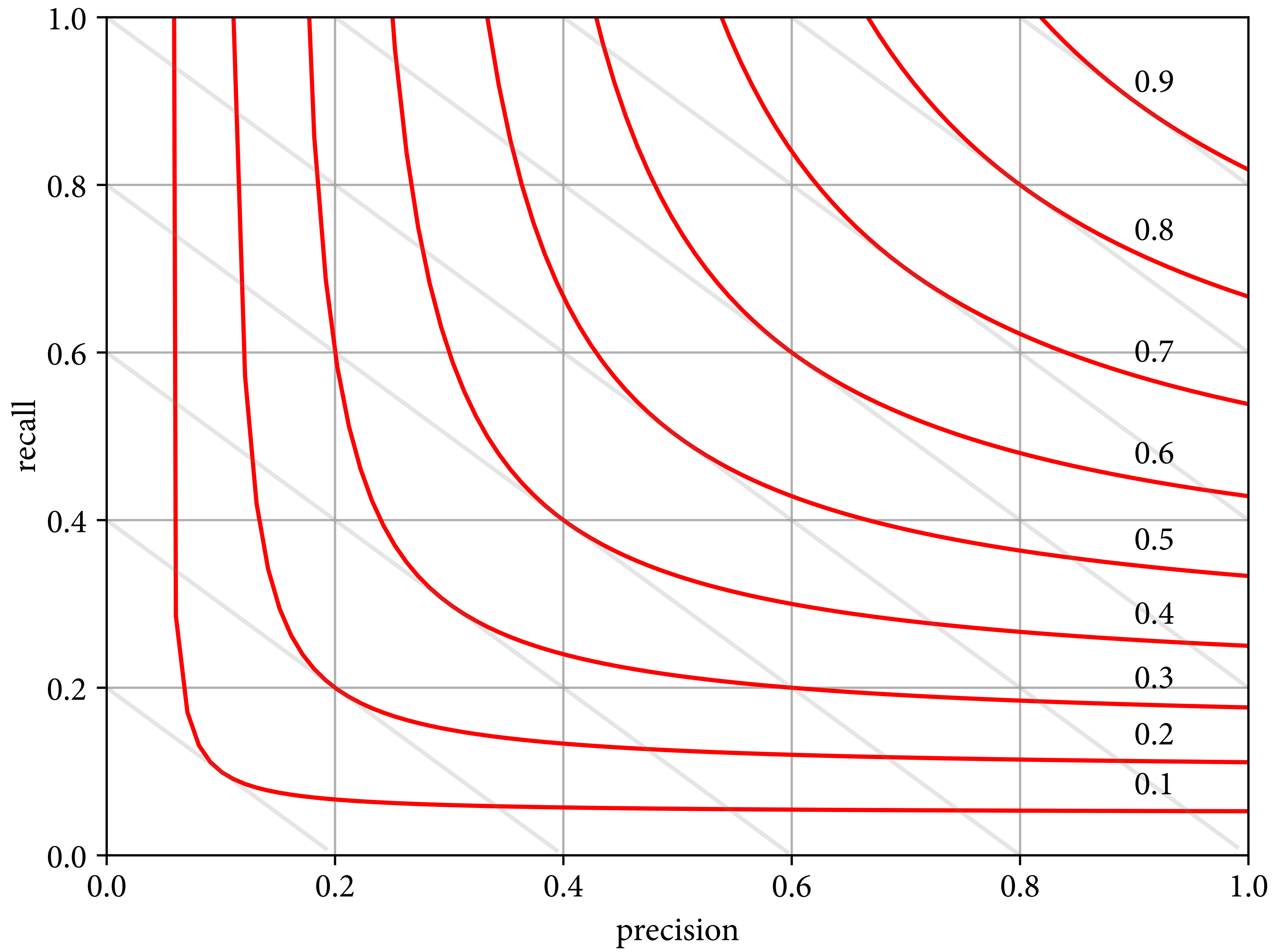
F1-measure

Eisenstein § 4.4.1

A good classifier should balance between precision and recall. The **F1-measure** is the harmonic mean of the two values:

$$\mathrm{F1} = \frac{2 \cdot \mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

F1-measure



0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
recall
precision

Accuracy with three classes

	A	B	C
A	58	6	1
B	5	11	2
C	0	7	43

Precision with respect to class B

	A	B	C
A	58	6	1
B	5	11	2
C	0	7	43

Recall with respect to class B

	A	B	C
A	58	6	1
B	5	11	2
C	0	7	43