Evaluation

Princeton AI4ALL NLP Group

Recap

- Linear regression and standard deviation
- Python functions
- File organization

Evaluation Metrics

Material adapted from Stanford AI4ALL
2017

Plan for today

- How to evaluate the performance of a classifier?
- Implement an evaluation metric and evaluate your rule-based classifier

Evaluation

- How can we determine whether one classifier is better than another?
- How can we determine whether making changes to a classifier actually leads to improvements?
    - E.g. does writing an additional rule help, or make things worse?

Overview

1. You have labeled data (each input labeled with its true category)
2. Split the data into *training examples* and *test examples*
3. Develop your classifier using the training examples
4. Use the classifier to label all the test examples
5. Compare the classifier's labels with the true labels and measure performance

Example	Predicted label	True label
1	Benign	Malignant
2	Benign	Malignant
3	Malignant	Malignant
4	Benign	Benign
5	Malignant	Benign
6	Benign	Benign
7	Benign	Benign
8	Malignant	Malignant
9	Benign	Benign
10	Benign	Benign

How would you measure the performance of the classifier?

Accuracy

$$\text{Accuracy} = \frac{\text{Number of correctly classified examples}}{\text{Number of examples}}$$

Example	Predicted label	True label	Correct?
1	Benign	Malignant	No
2	Benign	Malignant	No
3	Malignant	Malignant	Yes
4	Benign	Benign	Yes
5	Malignant	Benign	No
6	Benign	Benign	Yes
7	Benign	Benign	Yes
8	Malignant	Malignant	Yes
9	Benign	Benign	Yes
10	Benign	Benign	Yes

Accuracy?

Example	Predicted label	True label	Correct?
1	Benign	Malignant	No
2	Benign	Malignant	No
3	Malignant	Malignant	Yes
4	Benign	Benign	Yes
5	Malignant	Benign	No
6	Benign	Benign	Yes
7	Benign	Benign	Yes
8	Malignant	Malignant	Yes
9	Benign	Benign	Yes
10	Benign	Benign	Yes

Accuracy = 70%

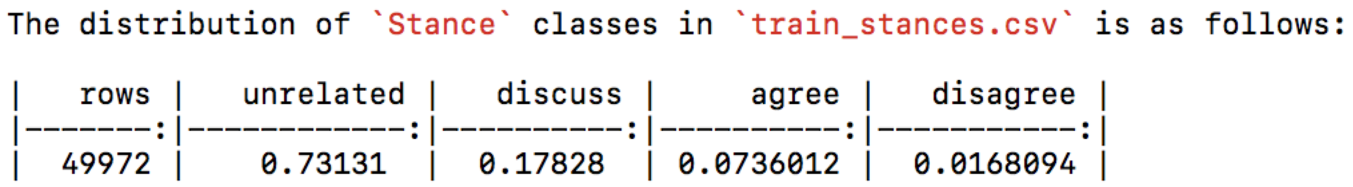
Example	Predicted label	True label	Correct?
1	Benign	Malignant	No
2	Benign	Malignant	No
3	Malignant	Malignant	Yes
4	Benign	Benign	Yes
5	Malignant	Benign	No
6	Benign	Benign	Yes
7	Benign	Benign	Yes
8	Malignant	Malignant	Yes
9	Benign	Benign	Yes
10	Benign	Benign	Yes

Just classify everything as benign!

Accuracy?

Accuracy

- Accuracy not a good when uneven distribution of labels in the data



```
The distribution of `Stance` classes in `train_stances.csv` is as follows:

|   rows |   unrelated |   discuss |     agree |   disagree |
|-------:|------------:|----------:|----------:|-----------:|
|  49972 |    0.73131  |   0.17828 | 0.0736012 |  0.0168094 |
```

		Predicted label	
		malignant	**benign**
True label	**malignant**	True positive	False negative
	benign	False positive	True negative

Example	Predicted label	True label	
1	Benign	Malignant	
2	Benign	Malignant	
3	Malignant	Malignant	
4	Benign	Benign	
5	Malignant	Benign	
6	Benign	Benign	
7	Benign	Benign	
8	Malignant	Malignant	
9	Benign	Benign	
10	Benign	Benign	

Example	Predicted label	True label	
1	Benign	Malignant	FN
2	Benign	Malignant	FN
3	Malignant	Malignant	TP
4	Benign	Benign	TN
5	Malignant	Benign	FP
6	Benign	Benign	TN
7	Benign	Benign	TN
8	Malignant	Malignant	TP
9	Benign	Benign	TN
10	Benign	Benign	TN

What about for spam classification?

True positives?

False positives?

True negatives?

False negatives?

What next?

Precision and Recall

Precision: "Of all those labeled positive, how many were correctly labeled?"

$$\text{Precision} = \frac{tp}{tp + fp}$$

Recall: "Of all the true positive examples, how many did the classifier detect?"

$$\text{Recall} = \frac{tp}{tp + fn}$$

Precision and Recall

Why do precision and recall matter?

- High recall:
- High precision:

- For detecting malignant tumors, which is more important?
- For convicting someone of a crime, which is more important?
- For detecting spam email, which is more important?

Precision and Recall

Why do precision and recall matter?

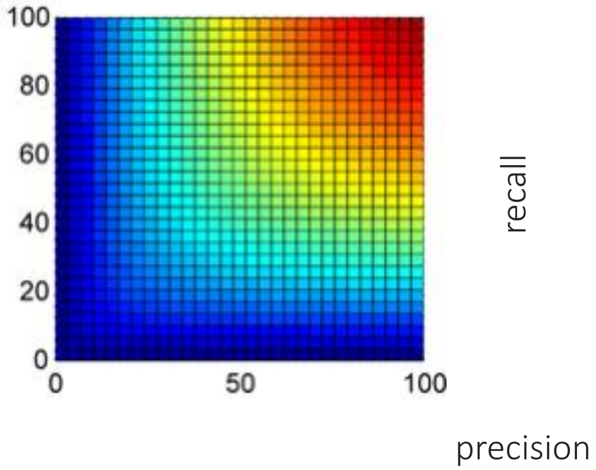
- High recall: classifier has few false negatives
  - e.g. few malignant tumors go undetected
- High precision: classifier has few false positives
  - e.g. few "false alarms" on benign tumors

- For detecting malignant tumors, which is more important?
  - High recall
- For convicting someone of a crime, which is more important?
  - High precision
- For detecting spam email, which is more important?
  - Both are important

How to combine precision and recall into a single balanced measure?

Harmonic mean

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$



How should we measure performance for the Fake News Challenge?

Go to AI4ALL_NLP_Student folder in Terminal

Type `git pull`

Type `unzip Day6_evaluation`

Go to the Day6_evaluation folder in Terminal

Type `source ~/miniconda3/bin/activate`

Type `jupyter notebook`