Metrics

CS230 Fall 2018
Section 10
Overview

● Shortcomings of Accuracy
● Recall
● Precision
● F1
Ground truth

Normal

Pneumonia
Model 1: Accuracy = 9/10

Model 2: Accuracy = 8/10

Model 3: Accuracy = 5/10

Ground truth:
Is Model1 the best model?
Recall
Recall

- Out of all the patients with pneumonia, how many did the model predict as having pneumonia?

- \( \frac{\text{(No. of patients that have pneumonia and are predicted as having pneumonia by the model)}}{\text{(No. of patients that have pneumonia)}} \)
Recall: (No. of patients that have pneumonia and are predicted as having pneumonia by the model) / (No. of patients that have pneumonia)

Model1
Recall = ?

Model2
Recall = ?

Model3
Recall = ?
Recall: (No. of patients that have pneumonia and are predicted as having pneumonia by the model) / (No. of patients that have pneumonia)
Are Model2 and Model3 equally good?
Precision
Precision

- Out of all the patients that are predicted to have pneumonia, how many actually have pneumonia?

- \( \frac{\text{No. of patients that have pneumonia and are predicted as having pneumonia by the model}}{\text{No. of patients that are predicted to have pneumonia}} \)
**Precision:** (No. of patients that have pneumonia and are predicted as having pneumonia by the model) / (No. of patients that are predicted to have pneumonia)
**Precision:** (No. of patients that have pneumonia and are predicted as having pneumonia by the model) / (No. of patients that are predicted to have pneumonia)

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Ground truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model1</td>
<td>0/0</td>
<td><img src="model1_ground_truth.png" alt="Ground truth" /></td>
</tr>
<tr>
<td>Model2</td>
<td>1/3 = 0.3</td>
<td><img src="model2_ground_truth.png" alt="Ground truth" /></td>
</tr>
<tr>
<td>Model3</td>
<td>1/6 = 0.16</td>
<td><img src="model3_ground_truth.png" alt="Ground truth" /></td>
</tr>
</tbody>
</table>
F1 score
Which one is better?

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>95%</td>
<td>90%</td>
</tr>
<tr>
<td>B</td>
<td>98%</td>
<td>85%</td>
</tr>
</tbody>
</table>

Image credits - Machine Learning Yearning, Andrew Ng, Pg 21
Hard to compare!

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>95%</td>
<td>90%</td>
</tr>
<tr>
<td>B</td>
<td>98%</td>
<td>85%</td>
</tr>
</tbody>
</table>
Other metrics

- PR AUC
- ROC AUC
- Sensitivity (used mostly in the medical field)
- Specificity (used mostly in the medical field)

Look at papers to find the metric used in the space that you’re working in!
F1 score

\[ F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \]
# F1 score

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>95%</td>
<td>90%</td>
<td>92.4%</td>
</tr>
<tr>
<td>B</td>
<td>98%</td>
<td>85%</td>
<td>91.0%</td>
</tr>
</tbody>
</table>

Image credits - Machine Learning Yearning, Andrew Ng, Pg 22
Back to our original example
Model1

F1 = ?

Ground truth

Model2

F1 = 0.5

Model3

F1 = 0.29