Evaluating ML

Lyle Ungar

Learning objectives
Probability vs loss
Confusion matrix: TP/TN/FP/FN
Precision, Recall, Sensitivity, Specificity
ROC curves
Loss functions come from decision making

- We often optimize a loss function which is a surrogate for our true loss function
- Don’t confuse *probability* with loss
  - One can optimize a model for probability and then use the probability in a decision rule
  - Or just directly optimize the loss resulting from a decision rule
What does probability mean?

http://mathwithbaddrawings.com/2015/09/23/what does probability mean in your profession/
What does probability mean?

What does probability mean?

What does probability mean?

What does probability mean in your profession?

What does probability mean

Mission Impossible Agent

"5 for 5 so far"

[what would that even mean?]

probability of mission success

What does probability mean

Millennium Falcon Captain

NEVER TELL ME THE ODDS

0

1

probability of successfully navigating an asteroid field

Don’t confuse your probability estimate with your actual loss function.
For a linear regression predicting dollar amounts (e.g. income, housing prices)

- What is the loss function being optimized for
- What is the residual plot likely to look like?

Does this meet the assumptions of the linear regression model?

- If not, how could you fix it?
Precision, Recall, Sensitivity, Specificity, and ROC curves

Have you seen ROC curves?

A) Yes
B) No
## Ways to be right or wrong

<table>
<thead>
<tr>
<th>Claim\Is</th>
<th>True Yes</th>
<th>True No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classify Yes</td>
<td>True Positive</td>
<td>False Positive</td>
</tr>
<tr>
<td>Classify No</td>
<td>False Negative</td>
<td>True Negative</td>
</tr>
</tbody>
</table>

Accuracy = \( \frac{TP + TN}{TP + FP + FN + TN} \)
Measuring Performance

- **Accuracy (symmetric)**
  - % correctly classified

- **Asymmetric measures**
  - **Precision**
    - $P(\text{yes} \mid \text{predicted as yes})$
  - **Recall (or Sensitivity)**
    - $P(\text{predicted as yes} \mid \text{yes})$
  - **Specificity**
    - $P(\text{predicted as no} \mid \text{no})$
### Precision/Recall: Sensitivity/Specificity

<table>
<thead>
<tr>
<th>Claim is</th>
<th>True</th>
<th>True</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>True Positive</td>
<td>False Positive</td>
</tr>
<tr>
<td>No</td>
<td>False Negative</td>
<td>True Negative</td>
</tr>
</tbody>
</table>

- **Precision**
  - $P(\text{yes} \mid \text{predicted as yes}) = \frac{TP}{TP+FP}$

- **Recall (or Sensitivity)**
  - $P(\text{predicted as yes} \mid \text{yes}) = \frac{TP}{TP+FN}$

- **Specificity**
  - $P(\text{predicted as no} \mid \text{no}) = \frac{TN}{TN+FP}$
### Precision/Recall Example

<table>
<thead>
<tr>
<th>Claim\Is</th>
<th>True Good</th>
<th>True Not Good</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classify “Good”</td>
<td>70</td>
<td>50</td>
</tr>
<tr>
<td>Classify “Not good”</td>
<td>30</td>
<td>350</td>
</tr>
</tbody>
</table>

- **Precision**
  - \( P(\text{good } | \text{ predicted as good}) = \frac{70}{(70+50)} \)
- **Recall (or Sensitivity) = True Positive Rate (TPR)**
  - \( P(\text{predicted as good } | \text{ good}) = \frac{70}{(70+30)} \)
- **Specificity = 1 – (False Positive Rate)**
  - \( P(\text{predicted as bad } | \text{ bad}) = \frac{350}{(350+50)} \)
**F1 combines Precision and Recall**

<table>
<thead>
<tr>
<th>Claim\s</th>
<th>True Yes</th>
<th>True No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classify Yes</td>
<td>True Positive</td>
<td>False Positive</td>
</tr>
<tr>
<td>Classify No</td>
<td>False Negative</td>
<td>True Negative</td>
</tr>
</tbody>
</table>

- **Precision**
  - \( \frac{TP}{TP+FP} \)
- **Recall**
  - \( \frac{TP}{TP+FN} \)
- **F1**
  - \( \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \)
**ROC** (Receiver Operating Characteristic) Curve

- Sort all examples from highest probability (or score) of being ‘yes’, $p(y='yes'|x)$, to lowest
- Sweep the threshold for predicting an example to be labeled ‘yes’ from 1 down to 0
  - This varies *specificity* from 1 to 0.
- At each threshold compute the *sensitivity*
  - i.e., the fraction of the true positives you found
- Plot the curve

https://en.wikipedia.org/wiki/Receiver_operating_characteristic
ROC Chart Varies Threshold

True positive rate

$p(y|x) > \text{threshold}$ to be in class

AUC = Area Under Curve
ROC charts support comparison

$AUC = 0.5$ is random guessing
$AUC = 1.0$ is perfection

$AUC$ = Area Under the Curve
Where does google care about?

- Sensitivity
- 0.00
- 0.25
- 0.50
- 0.75
- 1.00

- Specificity
- 0
- 0.25
- 0.5
- 0.75
- 1

- AUC
- Area Under Curve

- \( p(y|x) > \text{threshold} \) to be in class

A) here
B) here
C) here
D) The AUC

A, B, C, or D
Which method is most likely to be better for generating an ROC curve?
A) Logistic regression
B) SVM
A confusion matrix shows the counts of the actual versus predicted class values.

Example (overall accuracy rate of 73.9%)

<table>
<thead>
<tr>
<th>Predicted Class</th>
<th>Actual Class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class A</td>
</tr>
<tr>
<td>Class A</td>
<td>20</td>
</tr>
<tr>
<td>Class B</td>
<td>6</td>
</tr>
<tr>
<td>Class C</td>
<td>4</td>
</tr>
</tbody>
</table>
For the confusion matrix

<table>
<thead>
<tr>
<th>Actual</th>
<th>Purchase</th>
<th>No Purchase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Purchase</td>
<td>10</td>
<td>60</td>
</tr>
<tr>
<td>No Purchase</td>
<td>20</td>
<td>200</td>
</tr>
</tbody>
</table>

- What is its precision?  
  - a) 10/20  
  - b) 10/(10+20)  
  - c) 10/60  
  - d) 10/(10+60)  
  - e) other

- What is its recall?  
  - a) 10/20  
  - b) 10/(10+20)  
  - c) 10/60  
  - d) 10/(10+60)  
  - e) other

- How do you
  - a) increase precision (but decrease recall)
  - b) increase both precision and recall
Optimizing for true utility

Could one directly learn a model to optimize
- An asymmetric loss function?
- AUC?
You should know

- Probability vs. loss
- Loss function vs. utility function
- Confusion matrix:
  - TP/TN/FP/FN
  - TPR/TNR/FPR/FNR
- Precision, Recall, Sensitivity, Specificity, F1
- ROC curves
  - AUC