CSI 436/536
Introduction to Machine Learning

Binary classifications

Professor Siwei Lyu
Computer Science
University at Albany, State University of New York
Binary classification

- classifier: a function that maps data to categories
- also known as decision boundary
- output can also include a confidence score
- binary classification, e.g.,
  - assigning an email into "spam or "non-spam
  - Determine if two faces are of the same person
terminology

• prediction function (model) $f(x;w)$ maps input to $\mathbb{R}$
• A classifier will then use a threshold in decision function

• Typically the training of the binary classifier is with a loss function, and regularization
• The evaluation of binary classification uses a different set of metrics
evaluate binary classifiers

- A binary classifier with a threshold (positive on right)
  - actual = positive, classification = positive (TP)
    - 5 + examples > threshold
  - actual = negative, classification = negative (TN)
    - 5 - examples < threshold
  - actual = negative, classification = positive (FP)
    - 2 - examples > threshold
  - actual = positive, classification = negative (FN)
    - 2 + examples < threshold

negative ← - - - + - + - + + + + - + - + → positive
evaluate binary classifiers

- actual positive samples = TP + FN = 7
- classified positive samples = TP + FP = 7
- actual negative samples = TN + FP = 7
- classified negative samples = TN + FN = 7
- total samples = TP + TN + FP + FN = 14
evaluate binary classifiers

- accuracy (Acc) = \( \frac{TP + TN}{TP + TN + FP + FN} \)
  - 10/14 \( \approx 71.4\% \)
- error rate (Err) = \( \frac{FP + FN}{TP + TN + FP + FN} \)
  - 4/14 \( \approx 28.6\% \)
- Acc = 1 - Err, 0 \( \leq \) Acc, Err \( \leq \) 1
evaluate binary classifiers

- true positive rate (TPR) = TP/(TP+FN) = 5/7 ≈ 71.4%
- false positive rate (FPR) = FP/(TN+FP) = 2/7 ≈ 28.6%
- precision (PR) = TP/(TP+FN) = TPR = 5/7 ≈ 71.4%
- recall (RC) = TP/(TP+FP) = 5/7 ≈ 71.4%
- F1 score = 2*PR * RC/(PR+RC) ≈ 71.4%

negative  - - - + - + - + + + + - + - + positive
evaluate binary classifiers

- Confusion matrix

<table>
<thead>
<tr>
<th>TP</th>
<th>FN</th>
<th>Actual positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>FP</td>
<td>TN</td>
<td>Actual negatives</td>
</tr>
<tr>
<td>Classified positives</td>
<td>Classified negatives</td>
<td>All examples</td>
</tr>
</tbody>
</table>

- Diagonal is correct classification
- Anti-diagonal is incorrect classification
evaluate binary classifiers

- Confusion matrix

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>FN</th>
<th>Actual positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classified positives</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classified negatives</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All examples</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- TPR & FPR are ratios along the row direction
evaluate binary classifiers

- Confusion matrix

<table>
<thead>
<tr>
<th></th>
<th>Actual positives</th>
<th>Actual negatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>Classified positives</td>
<td>Classified negatives</td>
</tr>
<tr>
<td>FN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All examples</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- RC is on the column direction
Why single rates are not good

- consider a data set with 10 positive and 90 negative examples, and two binary classifiers
  - Classifier 1

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>82</td>
<td>90</td>
</tr>
<tr>
<td>18</td>
<td>82</td>
<td>100</td>
</tr>
</tbody>
</table>

- Classifier 2

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>0</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>2</td>
<td>98</td>
<td>100</td>
</tr>
</tbody>
</table>
The ROC curve

- Receiver Operator Characteristics (ROC) curve
  - tracing the curve of (FPR, TPR) with varying classification threshold
  - it is derived from the cumulative data distribution
ROC curve

- The two axes are integration over the class probabilities
  - connecting two diagonal points (0,0) and (1,1)
  - non-decreasing: as the two distributions are nonnegative
- In practice, ROC may not be smooth if there are not enough number of data
how to read ROC: summary

• ROC curve is trickier to understand
  • ideal ROC is line segment connecting (0,0) to (0,1) to (1,1)
  • worst ROC (random classifier) is the diagonal line connecting (0,0) and (1,1)
  • a good ROC is a convex curve connecting (0,0) and (1,1)
  • a ROC symmetric to the diagonal line can be obtained by flipping the class labels
  • support overlapping causes inflective ROC curve
  • multi-modal distribution causes multi-turn ROC curve
Reading ROC curves
Choosing threshold

- using maximum classification separation rule we get the intersection of the two distributions
  - The point with the minimum overall classification error
  - The point on ROC with (1,1) derivative
Area under ROC curve (AUR)

- \( \text{AUC} \in [0.5, 1.0] \)
  - 0.5 is the random binary classifier
  - 1.0 is the perfect binary classifier

ROC Curve

AUC = 0.89
Computing AUC on finite data

- AUC on finite data set can be computed using the Mann-Whitney-Wilcoxon (MWW) statistics
  \[
  \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{1}(x_i > x_j)
  \]

- Intuition: MWW statistics is the fraction of pairs with wrong orders
  - \( (4 + 3 + 2 + 2 + 2 + 1 + 0)/(5*5) = 56\% \)

- Proof:
  \[
  \Pr(X > Y) = \int \int_{x > y} f(x) g(y) \, dx \, dy = \int \int \delta(x > y) f(x) g(y) \, dx \, dy
  \]
  \[
  = \int f(x) \, dx \int \delta(x > y) g(y) \, dy = \int G(x) \, f(x) \, dx
  \]
  \[
  = \int G(x) \, dF(x) = \int \text{ROC}(F) \, dF = \text{AUC}
  \]
Evaluation of multi-classifier

- confusion matrix
  - Diagonal: correct examples
  - Off-diagonal: errors
- Acc and Err can be computed as in binary classification
- Cannot compute TPR/FPR
- Can still compute PR/RC

```
<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10</td>
<td>2</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>12</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>0</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>
```