ML in Practice:

CMSC 422
Slides adapted from Prof. CARPUAT and Prof. Roth
Expressivity

\[ f(x) = \text{sgn} \{ x \cdot w - \theta \} = \text{sgn}\{\sum_{i=1}^{n} w_i x_i - \theta \} \]

- Many functions are Linear
  - Conjunctions:
    - \( y = x_1 \land x_3 \land x_5 \)
    - \( y = \text{sgn}\{1 \cdot x_1 + 1 \cdot x_3 + 1 \cdot x_5 - 3\}; \quad w = (1, 0, 1, 0, 1) \quad \theta = 3 \)
  - At least \( m \) of \( n \):
    - \( y = \text{at least 2 of } \{ x_1, x_3, x_5 \} \)
    - \( y = \text{sgn}\{1 \cdot x_1 + 1 \cdot x_3 + 1 \cdot x_5 - 2\}; \quad w = (1, 0, 1, 0, 1) \quad \theta = 2 \)

- Many functions are not
  - Xor: \( y = x_1 \land x_2 \lor \neg x_1 \land \neg x_2 \)
  - Non trivial DNF: \( y = x_1 \land x_2 \lor x_3 \land x_4 \)

- But can be made linear
Functions Can be Made Linear

- Data are not linearly separable in one dimension
- Not separable if you insist on using a specific class of functions
Data are separable in $<x, x^2>$ space

Key issue: Representation:
- what features to use.
- Computationally, can be done implicitly (kernels)
- Not always ideal.
Exclusive-OR (XOR)

- $y = x_1 \land x_2 \lor \neg x_1 \land \neg x_2$

- In general: a parity function.

- $x_i \in \{0,1\}$

- $f(x_1, x_2, \ldots, x_n) = 1$ iff $\sum x_i$ is even

This function is not linearly separable.
Practical Issues

• “garbage in, garbage out”
  – Learning algorithms can’t compensate for useless training examples
    • e.g., if we only have irrelevant features
  – Feature design often has a bigger impact on performance than tweaking the learning algorithm
Practical Issues

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy on test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team A</td>
<td>80.00</td>
</tr>
<tr>
<td>Team B</td>
<td>79.90</td>
</tr>
<tr>
<td>Team C</td>
<td>79.00</td>
</tr>
<tr>
<td>Team D</td>
<td>78.00</td>
</tr>
</tbody>
</table>

Which classifier is the best?
– this result table alone cannot give us the answer
– solution: statistical hypothesis testing
Practical Issues

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy on test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team A</td>
<td>80.00</td>
</tr>
<tr>
<td>Team B</td>
<td>79.90</td>
</tr>
<tr>
<td>Team C</td>
<td>79.00</td>
</tr>
<tr>
<td>Team D</td>
<td>78.00</td>
</tr>
</tbody>
</table>

Is the difference in accuracy between A and B statistically significant?

What is the probability that the observed difference in performance was due to chance?
A confidence of 95%
• does NOT mean
“There is a 95% chance than classifier A is better than classifier B”
• It means
“If I run this experiment 100 times, I expect A to perform better than B 95 times.”
Practical Issues: Debugging

• You’ve implemented a learning algorithm, you try it on some train/dev/test data, but it doesn’t seem to learn.

• What’s going on?
  – Is the data too noisy?
  – Is the learning problem too hard?
  – Is your implementation buggy?
Practical Issues: Debugging

• You probably have a bug
  – if the learning algorithm cannot overfit the training data
  – if the predictions are incorrect on a toy 2D dataset hand-crafted to be learnable
N-fold cross validation

Instead of a single test-training split:

- Split data into N equal-sized parts

- Train and test N different classifiers

- Report average accuracy and standard deviation of the accuracy
Topics

• A few practical issues
  – CIML Chapter 5

• Dealing with imbalanced learning problems
  – Evaluation metrics (CIML 5.5)
  – Learning with imbalanced data (CIML 6.1)
Evaluation metrics: beyond accuracy/error

• Example 1
  – Given medical record,
  – Predict whether a patient has cancer or not

• Example 2
  – Given a document collection and a query
  – Find documents in collection that are relevant to query

• Accuracy is not a good metric when some errors matter more than others!
Imagine we are addressing a document retrieval task for a given query, where +1 means that the document is relevant and -1 means that the document is not relevant.

We can categorize predictions as:
- true/false positives
- true/false negatives

<table>
<thead>
<tr>
<th></th>
<th>Gold label = +1</th>
<th>Gold label = -1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction = +1</td>
<td>tp</td>
<td>fp</td>
</tr>
<tr>
<td>Prediction = -1</td>
<td>fn</td>
<td>tn</td>
</tr>
</tbody>
</table>
Precision and recall

- **Precision**: % of positive predictions that are correct
- **Recall**: % of positive gold labels that are found

<table>
<thead>
<tr>
<th></th>
<th>Gold label = +1</th>
<th>Gold label = -1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction = +1</td>
<td>tp</td>
<td>fp</td>
</tr>
<tr>
<td>Prediction = -1</td>
<td>fn</td>
<td>tn</td>
</tr>
</tbody>
</table>
A combined measure: F

- A combined measure that assesses the P/R tradeoff is F measure

\[ F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \]

- People usually use balanced F-1 measure
  - i.e., with \( \beta = 1 \) (that is, \( \alpha = 1/2 \)):
  - \( F = 2PR/(P+R) \)