Stages of (Batch) Machine Learning

**Given:** labeled training data \( X, Y = \{ \langle x_i, y_i \rangle \}_{i=1}^{n} \)

- Assumes each \( x_i \sim \mathcal{D}(X) \) with \( y_i = f_{\text{target}}(x_i) \)

**Train the model:**

\[
\text{model} \leftarrow \text{classifier.train}(X, Y)
\]

**Apply the model to new data:**

- Given: new unlabeled instance \( x \sim \mathcal{D}(X) \)

\[
y_{\text{prediction}} \leftarrow \text{model.predict}(x)
\]
Classification Metrics

\[
\text{accuracy} = \frac{\# \text{ correct predictions}}{\# \text{ test instances}}
\]

\[
\text{error} = 1 - \text{accuracy} = \frac{\# \text{ incorrect predictions}}{\# \text{ test instances}}
\]
Confusion Matrix

• Given a dataset of $P$ positive instances and $N$ negative instances:

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>Predicted Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>No</td>
<td>FP</td>
</tr>
</tbody>
</table>

accuracy = \frac{TP + TN}{P + N}

• Imagine using classifier to identify positive cases (i.e., for information retrieval)

precision = \frac{TP}{TP + FP}

recall = \frac{TP}{TP + FN}

Probability that a randomly selected result is relevant

Probability that a randomly selected relevant document is retrieved
Training Data and Test Data

• Training data: data used to build the model
• Test data: new data, not used in the training process

• Training performance is often a poor indicator of generalization performance
  – Generalization is what we really care about in ML
  – Easy to overfit the training data
  – Performance on test data is a good indicator of generalization performance
  – i.e., test accuracy is more important than training accuracy
Simple Decision Boundary

TWO-CLASS DATA IN A TWO-DIMENSIONAL FEATURE SPACE

Decision Region 1

Decision Region 2

Decision Boundary

Feature 1

Feature 2

Slide by Padhraic Smyth, UCIrvine
More Complex Decision Boundary

TWO-CLASS DATA IN A TWO-DIMENSIONAL FEATURE SPACE

Decision Region 1

Decision Region 2

Decision Boundary

Feature 1

Feature 2

Slide by Padhraic Smyth, UCIrvine
Example: The Overfitting Phenomenon
A Complex Model

$Y = \text{high-order polynomial in } X$
The True (simpler) Model

$Y = a \times X + b + \text{noise}$
How Overfitting Affects Prediction

Predictive Error

Error on Training Data

Model Complexity

Slide by Padhraic Smyth, UCIrvine
How Overfitting Affects Prediction

Predictive Error

Error on Test Data

Error on Training Data

Model Complexity

Slide by Padhraic Smyth, UCIrvine
How Overfitting Affects Prediction

Predictive Error

Underfitting

Error on Training Data

Error on Test Data

Overfitting

Ideal Range for Model Complexity

Model Complexity

Slide by Padhraic Smyth, UCIrvine
Comparing Classifiers

Say we have two classifiers, $C1$ and $C2$, and want to choose the best one to use for future predictions.

Can we use training accuracy to choose between them?

- No!
  - e.g., $C1 =$ pruned decision tree, $C2 =$ 1-NN
    
    
    \[
    \text{training\_accuracy}(1\text{-NN}) = 100\%, \text{ but may not be best}
    \]

Instead, choose based on test accuracy...
Idea:
Train each model on the “training data”...
...and then test each model’s accuracy on the test data.
$k$-Fold Cross-Validation

• Why just choose one particular “split” of the data?
  – In principle, we should do this multiple times since performance may be different for each split

• $k$-Fold Cross-Validation (e.g., $k=10$)
  – randomly partition full data set of $n$ instances into $k$ disjoint subsets (each roughly of size $n/k$)
  – Choose each fold in turn as the test set; train model on the other folds and evaluate
  – Compute statistics over $k$ test performances, or choose best of the $k$ models
  – Can also do “leave-one-out CV” where $k=n$
Example 3-Fold CV

Full Data Set

1st Partition

Test Data

Training Data

2nd Partition

Training Data

Test Data

Training Data

kth Partition

Training Data

Test Data

Test Performance

Test Performance

Summary statistics over k test performances
Optimizing Model Parameters

Can also use CV to choose value of model parameter $P$
- Search over space of parameter values $p \in \text{values}(P)$
  - Evaluate model with $P = p$ on validation set
- Choose value $p'$ with highest validation performance
- Learn model on full training set with $P = p'$

Choose value of $p$ of the model with the best validation performance
More on Cross-Validation

• Cross-validation generates an approximate estimate of how well the classifier will do on “unseen” data
  – As \( k \to n \), the model becomes more accurate (more training data)
  – ...but, CV becomes more computationally expensive
  – Choosing \( k < n \) is a compromise

• Averaging over different partitions is more robust than just a single train/validate partition of the data

• It is an even better idea to do CV repeatedly!
Multiple Trials of $k$-Fold CV

1.) Loop for $t$ trials:

a.) Randomize Data Set

```
Randomize Data Set
```

b.) Perform $k$-fold CV

```
Perform $k$-fold CV
```

2.) Compute statistics over $t \times k$ test performances

```
Compute statistics over $t \times k$ test performances
```
Comparing Multiple Classifiers

1.) Loop for \( t \) trials:

a.) Randomize Data Set

b.) Perform \( k \)-fold CV

2.) Compute statistics over \( t \times k \) test performances

- Allows us to do paired summary statistics (e.g., paired t-test)
Learning Curve

• Shows performance versus the # training examples
  – Compute over a single training/testing split
  – Then, average across multiple trials of CV
Building Learning Curves

1.) Loop for $t$ trials:

   a.) Randomize Data Set

   b.) Perform $k$-fold CV

2.) Compute statistics over $t \times k$ learning curves