Avoiding Overfitting

- We have a choice of different techniques:
  - Decision trees
  - Nearest neighbors
  - Bayes classifier
  - Neural networks
- For each we have different levels of complexity:
  - Depth of trees
  - Number of neighbors in K-NN
  - Number of layers and hidden units
  - ....
- How to choose the right one?
- Overfitting: A complex enough model (e.g., large enough trees, ...) will always be able to fit the training data well
Example

• Construct a predictor of $y$ from $x$ given this training data

Which model is best for predicting $y$ from $x$??
Which model is best for predicting $y$ from $x$? We want the model that generates the best predictions on future data. Not necessarily the one with the lowest error on training data.

Using a Test Set

1. Use a portion (e.g., 30%) of the data as test data
2. Fit a model to the remaining training data
3. Evaluate the error on the test data
Using a Test Set:
+ Simple
- Wastes a large % of the data
- May get lucky with one particular subset of the data
“Leave One Out” Cross-Validation

- For \( k = 1 \) to \( R \)
  - Train on all the data leaving out \((x_k, y_k)\)
  - Evaluate error on \((x_k, y_k)\)
- Report the average error after trying all the data points

\[ \text{Error} = 2.12 \]
Error = 0.962

Error = 3.33
“Leave One Out” Cross-Validation

“For $k=1$ to $R$:
- Train on all the data leaving out $(x_k, y_k)$
- Evaluate error on $(x_k, y_k)$

Report the average error after trying all the data points

K-Fold Cross-Validation

• Randomly divide the data set into $K$ subsets
• For each subset $S$:
  – Train on the data not in $S$
  – Test on the data in $S$
• Return the average error over the $K$ subsets

Example: $K = 3$, each color corresponds to a subset
Classification Problems

• The exact same approaches apply for cross-validation except that the error is the number of data points that are misclassified.
Example: CV for KNN

- For each kNN, evaluate the error using K-fold Cross-Validation
- Choose the one with the minimum cross-validation error

Cross-Validation Summary

<table>
<thead>
<tr>
<th></th>
<th>-</th>
<th>+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Set</td>
<td>Wastes a lot of data</td>
<td>Simple/ Efficient</td>
</tr>
<tr>
<td></td>
<td>Poor predictor of future performance</td>
<td></td>
</tr>
<tr>
<td>Leave One Out</td>
<td>Inefficient</td>
<td>Does not waste data</td>
</tr>
<tr>
<td>K-Fold</td>
<td>Wastes 1/K of the data</td>
<td>Wastes only 1/K of the data!</td>
</tr>
<tr>
<td></td>
<td>K times slower than Test Set</td>
<td>Only K times slower than Test Set!</td>
</tr>
</tbody>
</table>