Model Selection

• The bias-variance tradeoff provides conceptual framework for determining a good model
  – Bias-variance tradeoff not directly useful
• Need a practical method for optimizing bias-variance tradeoff
• Practical aim is to pick a model that minimizes a criterion:

\[ f_1(\text{fitting error from given data}) + f_2(\text{model complexity}) \]

where \( f_1 \) and \( f_2 \) are increasing functions
• All methods based on a tradeoff between fitting error (high variance) and model complexity (low bias)
• Criterion above may/may not be explicitly used in given method
Methods for Model Selection

• Among many popular methods are:
  – Akaike Information Criterion (AIC) (Akaike, 1974)
    • Popular in time series analysis
  – Bayesian selection (Akaike, 1977)
  – Bootstrap-based selection (Efron and Tibshirini, 1997)
  – Cross-validation (Stone, 1974)
  – Minimum description length (Risannen, 1978)
  – V-C dimension (Vapnik and Chervonenkis, 1971)
    • Popular in computer science

• Cross-validation appears to be most popular model fitting method
Cross-Validation

• Cross-validation is simple, general method for comparing candidate models
  – Other specialized methods may work better in specific problems

• Cross-validation uses the training set of data

• Method is based on iteratively partitioning the full set of training data into training and test subsets

• For each partition, estimate model from training subset and evaluate model on test subset
  – Number of training (or test) subsets = number of model fits required

• Select model that performs best over all test subsets
Choice of Training and Test Subsets

- Let $n$ denote total size of data set, $n_T$ denote size of test subset, $n_T < n$
- Common strategy is leave-one-out: $n_T = 1$
  - Implies $n$ test subsets during cross-validation process
- Often better to choose $n_T > 1$
  - Sometimes more efficient (sampling w/o replacement)
  - Sometimes more accurate model selection
- If $n_T > 1$, sampling may be with or without replacement
  - “With replacement” indicates that there are “$n$ choose $n_T$” test subsets, written $\binom{n}{n_T}$
  - With replacement may be prohibitive in practice: e.g., $n = 30$, $n_T = 6$ implies nearly 600K model fits!
- Sampling without replacement reduces number of test subsets to $n/n_T$ (disjoint test subsets)
Conceptual Example of Sampling Without Replacement: Cross-Validation with 3 Disjoint Test Subsets
Typical Steps for Cross-Validation

**Step 0 (initialization)** Determine size of test subsets and candidate model. Let $i$ be counter for test subset being used.

**Step 1 (estimation)** For the $i$th test subset, let the remaining data be the $i$th training subset. Estimate $\theta$ from this training subset.

**Step 2 (error calculation)** Based on estimate for $\theta$ from Step 1 ($i$th training subset), calculate MSE (or other measure) with data in $i$th test subset.

**Step 3 (new training and test subsets)** Update $i$ to $i + 1$ and return to step 1. Form mean of MSE when all test subsets have been evaluated.

**Step 4 (new model)** Repeat steps 1 to 3 for next model. Choose model with lowest mean MSE as best.
Numerical Illustration of Cross-Validation (Example 13.4 in ISSO)

• Consider true system corresponding to a sine function of the input with additive normally distributed noise

• Consider three candidate models
  – Linear (affine) model
  – 3rd-order polynomial
  – 10th-order polynomial

• Suppose 30 data points are available, divided into 5 disjoint test subsets (sampling w/o replacement)

• Based on RMS error (equiv. to MSE) over test subsets, 3rd-order polynomial is preferred

• See following plot
Numerical Illustration (cont’d): Relative Fits for 3 Models with Low-Noise Observations