Model Selection: choosing the best classifier for your task

EE511 Statistical Learning
May 6, 2008

No Free Lunch Theorem

- Is there a learning algorithm that is inherently better than others?
  - Better = Lower generalization error
- “No Free Lunch Theorem” says: NO!
  - No algorithm is universally good for all problems.
  - Averaged over all problems, all algorithms have equal error
  - This is true independent of P(x) and data size

No Free Lunch: “Law of conservation”

- If a learning algorithm performs above-average in some problems, it will perform below-average in others (Figure from Duda, Hart, Stork, “Pattern Classification”)

Model Selection

- For a particular task/dataset, an optimal algorithm may exist
- The goal of model selection:
  - Using your training data, find the model that is expected to generalize best
  - Model = different learning algorithms, different parameter settings of the same algorithm
- Here, we’ll use “classifier” and “model” interchangeably

Related concepts:
- Model assessment: Given a model, estimate its generalization error, error bars, etc.
- Model averaging: Combine models as opposed to selecting the best model

Today’s Agenda

1. Re-visit Bias-Variance Tradeoff
2. General concepts in Model Selection
3. Extra-sample methods
   - Cross-Validation
   - Leave-one-out error
4. Intra-sample methods
   - Bayesian model selection & BIC
   - AIC
   - MDL
5. Vapnik-Chervonekis Theory

Bias-Variance Tradeoff for Regression

- Mean Squared Error = bias^2 + variance
  \[ E[(f(x) - y)^2] = E[(f(x) - \mu)^2 + E[(f(x) - E[f(x)])^2]] \]
  - Low bias: on average, f(x) is close to truth
  - Low variance: f(x) does not change much as training data varies

- Model selection: find the best balance
  - To reduce bias, increase model complexity (generally)
  - To reduce variance, decrease model complexity (generally)
Today’s Agenda
1. Re-visit Bias-Variance Tradeoff
2. General concepts in Model Selection
   - Extra-sample methods
     - Cross-Validation
     - Leave-one-out error
   - Intra-sample methods
     - Bayesian model selection & BIC
     - AIC
     - MDL
3. Extra-sample methods
4. Vapnik-Chervonenkis Theory

General concepts in Model Selection
- Basic ingredients:
  - Goodness-of-fit
  - Model complexity

- Training a classifier = optimizing "goodness-of-fit"
  - Objective criteria is defined
  - Learning algorithm implements an optimization procedure
to pick classifier that best fits the data
  - This "selects" the best classifier among all classifiers of the
same class, where "best" = attains best objective value

- Issue: Classifiers from different classes cannot
be compared using the objective

Why can’t we compare across different classes of classifiers?
- Task: binary classification
- Training objective: min error on training set
- Two different classes of classifiers: linear vs. very wiggly

- Which achieves better objective? Which will you use?

Occam’s Razor
- William of Ockham (14th century logician):
  - "entia non sunt multiplicanda praeter necessitatem"
  - "entities should not be multiplied beyond necessity"
- In the case of statistical learning:
  - Prefer models that fit the data but have minimal complexity

Bias-Variance Tradeoff for Classification
- For 0-1 loss with $y \in \{0,1\}$, $P(y=1|x) = f(x)$:
\[
\Phi(t) = \int_{-\infty}^{\infty} e^{-t/2} \text{Pr}(u|t) du
\]
- Different from regression case. Here:
  - Nonlinear & multiplicative interaction between bias/variance
  - If bias is negative, low variance reduces $Pr(error)$
  - If bias is positive, high variance reduces $Pr(error)$
  - Variance dominates bias
  - One reason why classifiers care so much about complexity

Figure from Duda, Hart, Stork, Pattern Classification
Optimism of the Training Error Rate

- Goal: Low test (generalization) error
  \[ TestErr = E\left[ L(Y, \hat{f}(X)) \right] \]
- Typically: training error rate < test error
  - Same data is used to fit the model & assess its error
  \[ TrainErr = \frac{1}{N} \sum_{i=1}^{N} L\left( y_i, \hat{f}(x_i) \right) \]
- Test error is a kind of extra-sample error
  - Features in the test set are not observed in the training set
  - Cross-validation methods estimate this test error using a held-out set

In-sample error

1. Assuming a data generation process: \( y = f(x) + \text{noise} \):
   \[ \hat{Err}_{in} = \frac{1}{N} \sum_{i=1}^{N} E \left[ L(\hat{f}(x_i), y_i) \right] \]
   \( Y_{new} = \text{new response values at each of training points } x_i, \ i=1, 2, ..., N \)
2. Define optimism
   \[ \text{op} = \hat{Err}_{in} - E\left( \hat{Err}_{in} \right) \]
   For squared error, 0-1 loss functions, it can be shown:
   i.e. if we try too hard to fit \( y \), then optimism is high
3. Model selection formula:
   \[ \hat{Err}_{in} = E\left( \hat{Err}_{in} \right) + \frac{1}{N} \sum_{i=1}^{N} \text{Cov}(\hat{f}(x_i), y_i) \]

What are properties of a good model selection method?

- Accurately estimates generalization error
  - Has low bias and variance itself
- Consistency:
  - Selects the true model (assuming it’s in the class of classifiers considered) as number of training samples increases to infinity
- Computable
- No tunable parameters
- Widely applicable

Today’s Agenda

1. Re-visit Bias-Variance Tradeoff
2. General concepts in Model Selection
3. Extra-sample methods
   - Cross-Validation
   - Leave-one-out error
4. Intra-sample methods
   - Bayesian model selection & BIC
   - AIC
   - MDL
5. Vapnik-Chervonekis Theory

Extra-sample methods

- Directly estimate test error by using a held-out (validation) set
  - Validation, cross-validation, Leave-one-out error
  - Bootstrap
- Assumptions:
  - Validation set is similar to test set
  - Validation error is sufficiently accurate
- These methods are often used in practice
  - Usually achieves good estimates
  - Does not assume anything about the model (parametric, non-parametric) or the task (classification, regression, density estimation)

Hold-out set for model selection

- Use a development (validation) set to choose parameters/models

Training Data → Training Data Dev Data

What’s the proportion of dev set?
- Usually small (10-20%), since training data is needed to estimate many model parameters
K-Fold Cross-Validation

To calculate a model's performance, average the performance of each fold.
- More robust than single dev set.
- Usually 5-fold or 10-fold used.
- Leave-one-out error in the extreme case.
- Stratified CV: maintain label proportions.

How many folds in K-fold cross-validation?

<table>
<thead>
<tr>
<th>Less Variance</th>
<th>Less Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>LARGER K</td>
<td></td>
</tr>
</tbody>
</table>

If K=N (leave-one-out error), we get approximately unbiased estimate for test error.
- But high variance due to N very similar training sets.
- High computational burden (but some algorithms have clever tricks to do this).
For low K, bias may be a problem depending on size of training set.

Learning Curve

Large bias here (overly pessimistic about true error)

If we operate at this range or above, Cross-validation error has small bias.

Covariance matrix of cross-validation errors has a block structure.

Summary of Extra-sample methods

- K-fold Cross-validation is effective in practice.
  - K=5 or 10 is usually good, but do think about bias-variance questions.
    - It’s not a cure-all, or else it’d violate “No Free Lunch”.
- Bootstrap is an alternative.
  - (not discussed here)
  - Basic idea: resamples the training set with replacement to reduce variance.

Today’s Agenda

1. Re-visit Bias-Variance Tradeoff.
2. General concepts in Model Selection.
3. Extra-sample methods.
   - Cross-Validation.
   - Leave-one-out error.
4. Intra-sample methods.
   - Bayesian model selection & BIC.
   - AIC.
   - MDL.
5. Vapnik-Chervonekis Theory.
Bayesian Model Selection & BIC (1/5)

- We’ll derive the BIC using Bayesian principles
- Notation:
  - \( D \) = data, \( w = \) parameters, \( H_i = \) model \( i \)
- Two levels of being Bayesian:
  - Bayesian parameter estimation: \( \text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}} \)
  - Bayesian model selection:
    - Evidence \( P(D|H_i) \) is important
    - Select models \( H_1 \) vs \( H_2 \) based on posterior odds:
      \[
      \frac{P(H_1|D)}{P(H_2|D)} = \frac{P(H_1)}{P(H_2)} \cdot \frac{P(D|H_1)}{P(D|H_2)}
      \]
    - Usually ignored; doesn’t affect solution

Bayesian model selection (2/5)

- Bayesian parameter estimation
  \[
  P(w|D,H_i) = \frac{P(D|w,H_i) \cdot P(w|H_i)}{P(D|H_i)}
  \]
- Bayesian model selection
  \[
  P(H_i|D) = \frac{P(D|H_i) \cdot P(H_i)}{P(D)}
  \]
  - Evidence \( P(D|H_i) \) is important
  - Select models \( H_1 \) vs \( H_2 \) based on posterior odds:
    \[
    \frac{P(H_1|D)}{P(H_2|D)} = \frac{P(H_1)}{P(H_2)} \cdot \frac{P(D|H_1)}{P(D|H_2)}
    \]
    - Bayes Factor

Bayesian model selection (3/5)

You might imagine we need to encode the priors to penalize complex models… But it turns out it is captured in evidence \( P(D|H) \) automatically

- Complex \( H \) will have large support because it can fit many datasets
- That implies it is less peaked

Bayesian model selection (4/5): How to evaluate the evidence?

- To compute evidence, integrate over all parameters:
  \[
  P(D|H_i) = \int P(D|w,H_i) \cdot P(w|H_i) dw
  \]
- Laplacian Approximation: suppose \( P(w|D,H_i) \) is peaky at MAP solution
  \[
  P(D|H_i) \approx \left( 1 \sigma_{w|D|H_i}^2 \right)^{-\frac{N}{2}}
  \]
  - Width of peak
  - Best fit likelihood
  - Occam Factor
- Occam Factor = factor in which \( H_i \)’s hypothesis space collapses after seeing data.
  - Penalizes complex models (i.e. high \( \sigma_w \))
  - Penalizes models that fit too much to the data (i.e. low \( \sigma_{w|D} \))

BIC: Bayesian Information Criteria (5/5)

- BIC can be derived from a Gaussian approximation of the evidence term
- Basic Form:
  \[
  \text{BIC} = -2 \cdot (\log \text{lik} + (\log N) \cdot d)
  \]
  - Choose the model that has the lowest BIC
  - BIC tends to penalize complex models heavily
  - Note dependence on \( N \)

Akaike Information Criteria (AIC)

- AIC is another estimate of in-sample prediction error
  - Derived from information theoretic arguments
    \[
    \text{For } N \to \infty:\quad -2E\left[\log \text{Pr}_\theta(Y)\right] = -\frac{1}{N} E[\log \text{lik}] + 2 \cdot \frac{d}{N}
    \]
    - \( \text{Pr}_\theta(Y) \)… family density for \( Y \) (containing the true density)
    - \( \hat{\theta} \)… ML estimate of \( \theta \)
    - \( \log \text{lik} = \sum_{i=1}^{N} \log \text{Pr}_\hat{\theta}(y_i) \)
    - Maximized log-likelihood due to ML estimate of theta
AIC or BIC?
- Recall we choose model with small BIC/AIC
  \[
  BIC = -2 \cdot \log(\text{lik}) + (\log N) \cdot d
  \]
  \[
  AIC = -2 \cdot \log(\text{lik}) + 2 \cdot d
  \]
- BIC is asymptotically consistent. AIC is not.
- AIC tends to choose more complex models as N increases
- BIC usually favors simpler models, due to strong penalty
- Asymptotically, AIC = leave-one-out; Asymptotically, BIC = cross validation with a particular K. To see this, note:
  \[
  \log(P(D|H)) = \log(P(d_1|H)) + \log(P(d_2|d_1,H)) + \log(P(d_3|d_1,d_2,H)) + \ldots
  \]
  Cross validation examines the average of this term over reorderings of data.

MDL: Minimum Description Length
- MDL can be derived from Bayesian model selection and vice versa. (But originally derived independently)
  - Replaces probability of events by code length required to communicate the event
  - \[ L(x) = -\log(2(P(x))): \text{ length of } x \text{ in bits} \]
- MDL criteria chooses the model with the least bits.
  \[
  MDL = -\log(P(H)) - \log(P(D|H)) + \text{constant}
  \]
  Data block
- Other issues: precision of parameters

Summary of In-Sample Estimates
- AIC, BIC, MDL:
  - Many possible derivations
  - All contain 2 terms: goodness-of-fit & complexity
- AIC/BIC/MDL vs. Cross-Validation?
  - Personally I’d choose CV (more practical, applies to more kinds of models) unless there’s good theoretical reasons to motivate AIC/BIC/MDL, etc.

Today’s Agenda
1. Re-visit Bias-Variance Tradeoff
2. General concepts in Model Selection
3. Extra-sample methods
   - Cross-Validation
   - Leave-one-out error
4. Intra-sample methods
   - Bayesian model selection & BIC
   - AIC
   - MDL
5. Vapnik-Chervonekis Theory
   - Save for later

Overall Summary
- No Free Lunch
- Central issues:
  - How to compare models of different complexity
  - Occam’s Razor
- BIC, AIC, MDL:
  - different goodness-of-fit & complexity terms
  - Bayesian model selection as motivation for BIC
- Direct estimate of test error by hold-out
  - Cross validation, Leave-one-out
- Bias-variance for classifiers, Bias-variance for estimators of test error