Model selection

Day 2: Different approaches to model selection

Models as predictive tools

- How do we measure predictive ability and how do we estimate it?
- Loss function L(Y, f(X,$\hat{\theta}$))
  - L(Y, f(X,$\hat{\theta}$)) = (Y - f(X,$\hat{\theta}$))^2
  - L(Y, f(X,$\hat{\theta}$)) = |Y - f(X,$\hat{\theta}$)|
- Loss error for binary data
  - Likelihood: Σ Y x log($\widehat{Prob}$(Y=1)) + (1-Y) log($\widehat{Prob}$(Y=0))
  - Misclassification: $\Sigma\, Y_{\hat{Y}=0} + \ (1-Y)_{\hat{Y}=1}$

Models as Predictive Tools

- Training error Err_T = E[L(Y, f(X,$\hat{\theta}$))|T]
- Expected Err = E[$L(Y,$ f(X,$\hat{\theta}$))] = E[Err_T]

- Training error = $\frac{1}{n}\sum_{i=1}^{n} L(Y_i,$ f(X_i,$\hat{\theta}$))
  - Biased (optimistic) estimate of Expected error

Models as Predictive Tools

- Numerical vs Theoretical estimators of Expected Error
- Numerical: cross-validation and bootstrap
- Theoretical: C_p of Mallows
- AIC

Models as Predictive Tools

- Cross-validation
  - Divide the dataset in two, one to fit the model and one to assess the prediction error
  - Divide the data in k parts, use k-1 to fit the model and the last one to assess the prediction error AND repeat the process with a large number of splits
  - Jackknife or Leave-One-Out: remove one observation and predict it using all other observations
  - 2-fold = bias, jacknife = variance

Cross-validation

- Important to use cross-validation from the start
- Hastie et al. (2009): select «good» predictors from the whole data set (e.g. using marginal correlations), use cross validation to tune them (select a subset, estimate parameters). This is WRONG
- Select a subset of the data (e.g. k-fold cross validation). Select good predictors, …

Models as predictive tools

- Bootstrap
  - Generate data subsets using resampling
  - Fit the model on the subset and evaluate it on all observations
  - Obvious problem: the subset and the training set have observations in common (on average 0.632)
  - Various approaches to correct for the resulting optimism

Models as predictive tools

- C_p of Mallows
- Estimate the scaled prediction error

$$\frac{\sum_{i=1}^{n}(\hat{Y}_i - E\,Y_i)^2}{\sigma^2}$$

- $C_p = \frac{SSE}{\widehat{\sigma^2}} - (n - 2p)$
- $\widehat{\sigma^2}$ calculated from the most complex model
- C_p ~ p represents a good model

AIC

- Estimate the (relative) KL distance between different models and the «true» g
- Best predictive model when the distance between the predictive density and the true density is KL
- Efficient model selection criterion: select the best model in terms of expected prediction error (eg squared loss)

AIC vs BIC

- AIC is efficient, BIC is not
- BIC is consistent
  - Strong consistency: IF the true model belongs to the class of models used, BIC will select this model when n gets large
  - Weak consistency: when n gets large, select the model closest (KL) to the true model
- One cannot be efficient and consistent

(Yang; Claeskens and Hjort)

Models as tools to estimate parameters

- What is a good estimate?
- MSE = Bias^2 + Variance
- AIC does not optimize model selection to find the best model(s) for estimating parameters or a specific prediction (i.e. for one X)
- FIC (Claeskens and Hjort) does – difficult to calculate…