Selecting Among Stochastic Models

Cosma Shalizi

Statistics Department, Carnegie Mellon University

Santa Fe Institute

5 December 2008

Why Isn't Model Selection Easy?

Naive model selection: Fit all the different model classes you'd consider, see which matches the data best, then use that

"How can $R^2 = 0.9$ be wrong?"

This is generally a *very bad idea*: leads to over-fitting; model will perform poorly

(Error in-sample) = (Error out-of-sample) + (capitalizing on luck)

If a stochastic model is correct, it *shouldn't* fit the data perfectly

Model Selection

Given: candidate model classes $\Theta_1, \Theta_2, \ldots$
data $z_1, z_2, \ldots z_n$
Unknown: the best model class Θ_{k^*} *Return:* a guess $\widehat{k}$ as to k^*
A good method is one which reliably selects the best model class:

$$\lim_{n \to \infty} \Pr\left(\widehat{k} \neq k^*\right) = 0$$

("consistency")
"Best model class" can mean

1. The one which will predict best in the future
2. The one which contains the (best approximation to) truth, the model which generated the data

These are distinct concepts!

The true model can be in a Θ_k with so many free parameters that estimation is hopeless, and we get better predictions *from limited data* with a systematically wrong but more tractable model

M-estimation

Additive loss function:

$$\begin{aligned} L(\theta; z_1, z_2, \ldots z_n) &= \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; z_i) \\ &\equiv L(\theta; z_1^n) \equiv L_n(\theta) \end{aligned}$$

e.g., normalized negative log likelihood

$$L_n(\theta) = -\frac{1}{n}\sum_{t=1}^{n} \log p(z_t | z_1^{t-1}; \theta)$$

or mean squared error:

$$L_n(\theta) = \frac{1}{n}\sum_{t=1}^{n} y_t - f(x_t; \theta)$$

or mis-classification rate, etc.

Minimum error estimators ("M-estimators"):

$$\widehat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}}\, L_n(\theta)$$

Can constrain θ to be in some $\Theta_k \subset \Theta$

$$\widehat{\theta_k} = \underset{\theta \in \Theta_k}{\operatorname{argmin}}\, L_n(\theta)$$

Sampling Fluctuations in the Loss

$$L_n(\theta) = \mathbf{E}\left[L(\theta)\right] + \eta_n(\theta)$$

Best model ("pseudo-truth") is

$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmin}}\, \mathbf{E}\left[L(\theta)\right]$$

We *select* a model to minimize the loss, so

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmin}}\, \mathbf{E}\left[L(\theta)\right] + \eta_n(\theta)$$

Law of large numbers/ergodic theorem: for any *one* θ, with probability one

$$L_n(\theta) \to \mathbf{E}\left[L(\theta)\right]$$

or

$$\eta(\theta) \to 0 \ a.s.$$

Consistency will be ensured by a uniform LLN/ergodicity:

$$\sup_{\theta} |\eta_n(\theta)| \to 0$$

Regularization

Simply minimizing $L_n(\theta)$ is unstable, small changes in data lead to big changes in solution

Typical approach from optimization: add a regularizing term, penalize the more unstable types of solutions

i.e., minimize

$$L_n(\theta) + \lambda g(\theta)$$

Need to pick λ very carefully for consistency

—Lasso/L1:

$$L_n(\theta) + \lambda \sum_{i=1}^{p} |\theta_i\|$$

encourages sparse solutions; not equivariant; model-selection consistent for a lot of problems

—Ridge/L2:

$$L_n(\theta) + \lambda \sum_{i=1}^{p} \theta_i^2$$

encourages small solutions; not equivariant

—Curvature:

$$L_n(\theta) + \lambda \int \nabla^2 f(x;\theta) dx$$

encourages flat solutions; very useful with splines

Information Criteria

Use negative normalized log-likelihood as the loss function
$\widehat{\theta}_k \equiv$ best-fitting model in Θ_k
Akaike's information criterion:

$$AIC(\Theta_k) = L_n(\widehat{\theta}_k) + \frac{\dim(\Theta_k)}{n}$$

RULE: $\widehat{k} = \operatorname{argmin} AIC(\Theta_k)$
Schwarz's Bayesian information criterion:

$$BIC(\Theta_k) = L_n(\widehat{\theta}_k) + \frac{\dim(\Theta_k)}{2}\frac{\log n}{n}$$

RULE: $\widehat{k} = \operatorname{argmin} BIC(\Theta_k)$
In general: add a penalty term $r(\Theta_k, n)$ which is $o_p(1)$
$r(\Theta_k, n)$ is generally an estimate of the over-fitting $\eta_n(\widehat{\theta}_k)$
could e.g. involve the Hessian of the likelihood at $\widehat{\theta}_k$

slightly non-standard notation; usually written in terms of log-likelihood $\mathcal{L}$

$$AIC(\Theta_k) = \mathcal{L}(\widehat{\theta_k}) - \dim(\Theta_k)$$

$$BIC(\Theta_k) = \mathcal{L}(\widehat{\theta_k}) - \frac{\dim(\Theta_k)}{2} \log n$$

Origin Myths: AIC

Sharpness of estimates depends on curvature of loss function:

$$H(\theta)_{ij} = \frac{\partial^2 L}{\partial \theta_i \partial \theta_j}$$

Pretend H is diagonal for the moment

$$\widehat{\theta}_i - \theta_i^* \approx n^{-1/2} H(\theta^*)_{ii}^{-1} U_i$$

with $U_i \sim \mathcal{N}(0,1)$, *if* the model is correct

Don't need the diagonal assumption but it saves a lot of linear algebra

Taylor expand to 2nd order:

$$\begin{aligned} L(\widehat{\theta}) &\approx L(\theta^*) + \frac{1}{2} \sum_{i=1}^{\dim(\Theta_k)} (\widehat{\theta}_i - \theta_i^*) H_{ii} (\widehat{\theta}_i - \theta_i^*) \\ \mathbf{E}\left[L(\widehat{\theta})\right] &\approx \mathbf{E}\left[L(\theta^*)\right] + \frac{1}{n} \dim(\Theta_k) \end{aligned}$$

so AIC is an unbiased estimate of the generalization error

the $\frac{1}{2}$ really does go away legitimately

Origin Myths: BIC

Introduce prior distribution $\rho(\theta)$ over Θ
Marginal/integrated likelihood:

$$\mathcal{L}(\Theta_k) = p(z_1^n | \theta \in \Theta_k) = \log \int_{\Theta_k} p(z_1^n; \theta) \rho(\theta | \theta \in \Theta_k) d\theta$$

RULE: $\widehat{k} = \text{argmax}\, \mathcal{L}(\Theta_k)$
Justification 1: $\rho(\theta \in \Theta_k | z_1^n) \propto e^{\mathcal{L}(\Theta_k)} \rho(\theta \in \Theta_k)$, so this is the Bayesian solution

Real Bayesians don't select models

Justification 2: With $\rho(\theta | \theta \in \Theta_k)$ diffuse, as $\dim(\Theta_k)$ grows, more of the prior mass goes on large parameter vectors $\|\theta\| \gg 0$, which are bad, so the average gets pulled down from the high-performing θ by their crazy relatives
As $n \to \infty$ the prior gets swamped (usually)

Most of the volume of a high-dimensional hypersphere is ϵ-close to the surface

$\therefore$ diffuse high-dimensional priors are weird

Calculating $\mathcal{L}(\Theta_k)$ is generally intractable
Approximate by Laplace's method:

$$\begin{aligned}
\log \mathcal{L}(\Theta_k) &\approx \log p(z_1^n; \widehat{\theta}) \\
&\quad + \frac{\dim(\Theta_k)}{2} \log 2\pi \\
&\quad - \frac{\dim(\Theta_k)}{2} \log n \\
&\quad - \frac{1}{2} \log |H(\widehat{\theta})| \\
&\quad + \log \rho(\widehat{\theta_k} | \theta \in \Theta_k)
\end{aligned}$$

Discard $O_p(1)$ terms (constant, Hessian, prior) $\Rightarrow$ BIC

The Truth About Information Criteria

— If the data-generating model is finite-dimensional, and the data are IID/regression/Markov/etc., BIC is consistent
— AIC in general is *not* consistent and will tend to over-fit even as $n \to \infty$
— AIC can give better generalization error than BIC when the truth is infinite-dimensional
— Nothing magical about the AIC and BIC penalties
— Even for estimating risk, number of parameters is not really what's wanted, unless model is well-behaved *and* well-specified

Can "robustify" by undoing some of the lies told in the derivations, leading to more complicated penalty terms

Cross-Validation

Generalization performance = expected error on new data from the same source

Fake this by pretending that some of your data is really new

Basic algorithm sketch:

- For $j = 1 : m$
  - Randomly divide z_1^n into z_{train_j} and z_{test_j}
  - For each Θ_k, estimate $\widehat{\theta_{k,j}}$ using only z_{train_j}
  - Calculate $L(\widehat{\theta_{k,j}}; z_{\text{test}_j})$
- $CV(\Theta_k) = m^{-1} \sum_j L(\widehat{\theta_{k,j}}; z_{\text{test}_j})$

RULE: $\widehat{k} = \text{argmin}\, CV(\Theta_k)$

CV works because

$$CV(\Theta_k) \approx \mathbf{E}\left[L(\theta_k^*)\right]$$

Extremely reliable, robust, and practical; 10-fold CV is basically "industry standard"
Standard way of picking λ in penalization methods
Relies on dividing data into *independent* training/testing sets
CV is not so easy even for sequential data (but possible)
CV for networks/relational data would be a Very Good Thing

Penalizing Capacity to Over-Fit

If over-fitting is the problem, *penalize over-fitting*

$$\eta_n(\Theta_k) \equiv \sup_{\theta \in \Theta_k} |\eta_n(\theta)|$$

Size depends on

- Pointwise rate at which $|\eta_n(\theta)| \to 0$
  Large deviations theory (usually) says:
  $\Pr\left(|\eta_n(\theta)| > \epsilon\right) \approx e^{-Cn\epsilon^2}$
- Size (in some sense) of Θ_k

The number of *effectively* distinct θ in Θ_k grows with n

With $n = 5$ there are at most 32 distinguishable classifiers

Similarly, but scale-dependently, for regression, etc.

How rapidly does the number of distinguishable models grow with the amount of data?

- Exponentially-small error probabilities $\times$ exponentially-large number of models $\Rightarrow$ trouble
- Exponentially-small error probabilities $\times$ polynomial number of models $\Rightarrow$ consistency

Growth rate in number of distinguishable models can be quantified in various ways

covering numbers, bracketing numbers, Vapnik-Chervonenkis dimension, Pollard pseudo-dimension, fat-shattering dimension, Rademacher complexity, ...

These are the complexities of the *model classes*, not of the system being modeled

Generally, complexity $\neq$ number of parameters

Different size measures lead to different bounds on $\eta_n(\Theta_k)$

Many bounds are distribution-free (though worst case)

Bias-variance trade-off

High capacity model classes have lower approximation bias:

$$\underset{\theta \in \Theta_k}{\operatorname{argmin}} \, \mathbf{E}\left[L(\theta)\right] - \mathbf{E}\left[L(\theta^*)\right]$$

shrinks as the capacity grows

High capacity model classes have higher (worst-case) estimation variance:

$$\eta_n(\Theta_k)$$

grows as the capacity grows

Capacity control: picking a model class to optimize this bias/variance trade-off

Cross-validation $\approx$ capacity control without math (or guarantees)

Structural Risk Minimization

Let $B(\Theta_k, n) =$ your favorite learning-theory bound on $\eta_n(\Theta_k)$
RULE: $\widehat{k} = \text{argmin}\, L(\widehat{\theta_k}) + B(\Theta_k, n)$
— Consistent, because model classes are penalized *directly* by how badly they could be over-fitting
— Tends to work extremely well when it can be applied; major difficulty is getting suitable bounds

Method of Sieves

$\Theta_k \subset \Theta_{k+1}$
With n samples, estimate in $\Theta_{k(n)}$
Let $k(n) \to n$ as $n \to \infty$, but slowly
Handles the bias/variance trade-off as well
Examples: smoothing methods for density estimation and regression

Model Selection by Hypothesis Testing

Model selection basically *is* a hypothesis test
Why not do it directly?
Consider two *fixed* parameter values θ and ψ; assume IID
Pointwise log likelihood ratio:

$$\Lambda_i = \log p(z_i; \theta) - \log p(z_i; \psi)$$

By law of large numbers,

$$\begin{aligned}
\overline{\Lambda}_n &\equiv \frac{1}{n}\sum_{i=1}^{n}\Lambda_i \\
&\rightarrow \mathbf{E}\left[\Lambda_i\right] \\
&= \mathbf{E}\left[\log p(Z; \theta)\right] - \mathbf{E}\left[\log p(Z; \psi)\right] \\
&= \mathbf{E}\left[L(\psi)\right] - \mathbf{E}\left[L(\theta)\right]
\end{aligned}$$

so the sign of $\overline{\Lambda}_n$ indicates which model fits better

Suppose $\mathbf{E}[L(\psi)] - \mathbf{E}[L(\theta)] \approx 0$ — fluctuations become significant and sign becomes unreliable
Use central limit theorem:

$$V_n \equiv \sqrt{n}\frac{\overline{\Lambda}_n}{\sigma(\Lambda)} \rightsquigarrow \mathcal{N}(\mathbf{E}[L(\psi)] - \mathbf{E}[L(\theta)], 1)$$

If the two models are equally good, $V_n \rightsquigarrow \mathcal{N}(0, 1)$

Assumes two fixed models, but that's never interesting

$$\widehat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}}\, L_n(\theta)$$
$$\widehat{\psi} = \underset{\psi \in \Psi}{\operatorname{argmin}}\, L_n(\psi)$$

Vuong (1989) showed that in reasonable models, V_n is *still* normally distributed, even with estimated distributions

RULE: Select Θ if $V_n > V_c$, select Ψ if $V_n < -V_c$, and refuse to pick otherwise

—If one model *is* better, this will eventually pick it out with probability 1; probability of false selections controllable by null distribution of V

—Generalizes to regression, time series, can include $o_p(1)$ correction terms (like AIC or BIC or Hessian), etc.

Specification/Adequacy Testing

Information theory: the correct distribution, and only the correct distribution, lets us compress the data to white noise

$\therefore$ non-white-noise compression $\Rightarrow$ the wrong distribution

— Compression in more familiar statistical terms: residuals are IID from the specified distribution

— Ordinary linear regression: Can test for independence, Gaussianity, homoskedasticity, etc.

Recent paper by Aris Spanos: compare Ptolemy to Kepler on orbit of Mars

AIC prefers epicycles because residuals are smaller

but Ptolemaic residuals are *very* structured, so it fails specification tests

More general (but less specifically-powerful) specification tests look at the difference between

$$\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j}$$

and

$$(\nabla_\theta \ell)^T (\nabla_\theta \ell)$$

which should converge if the model is correct
— Specification/adequacy testing is not *directly* comparative
can give a collection of model classes which are all compatible with the data
$\approx$ a confidence set of model classes

Structure Learning

With graphical models, each graph implies a different set of conditional independence relations among variables

and any other conditional independencies require measure-0 conspiracies among parameters (faithfulness)

$\therefore$ if the graph is right, those independencies will hold

$\therefore$ can *eliminate* graphs, reliably, if their independence relations *don't* hold

$\therefore$ construct a confidence set of possible graphs

asymptotically the confidence set narrows to the true graph

Encompassing

Rather than compare distinct model classes, find a single meta-specification of which they are all special cases

usually by continuous interpolation

Estimate in the meta-specification

If you really need to make a choice, look at which special case the grand estimate is closest too

Model Averaging

Why chose *a* model in the first place?
Get a weight $w(k)$ for each Θ_k
Then predict/estimate using $\widehat{\theta}_k$ with weight $w(k)$
For sensible loss functions

$$\begin{aligned}(\text{Generalization error of averaged predictions}) \\ = \quad (\text{Average prediction errors}) - (\text{Diversity of predictions})\end{aligned}$$

(smoothing effect)

Where do the weights come from?
AIC weights (my notation):

$$w(k) \propto e^{-nAIC(\Theta_k)}$$

BIC weights:

$$w(k) \propto e^{-nBIC(\Theta_k)}$$

General exponentially weighted "mixture of experts"

$$w(k) \propto e^{-nL_n(\widehat{\Theta_k})}$$

Exponentially weighted predictors have a strong property:

$$\begin{aligned} &(\text{Accumulated prediction error of the mixture}) \\ \leq\ &(\text{Error of the best predictor}) + C \log n \end{aligned}$$

no matter what the data are

This is very strong but it's retrospective (small regret)

Does assume that the number of experts is finite (C grows with this number)

Bagging

— Generate a bootstrap sample from the original data

i.e. resample with replacement *n* times

— Fit a model to the bootstrap sample

— Repeat many times

— Average the predictions

— Many variants

"boosting" weights later bootstrap samples by how hard it's been to predict sample points

Hierarchical Ensembles

Stacking: High-level model looks at data and picks a low-level model to use for actual prediction

Hierarchal mixture: High-level model looks at data and picks weights over low-level models

Bayesian Model Averaging

Posterior distribution over parameters:

$$\rho(\theta|z_1^n) = \frac{\rho(\theta)p(z_1^n;\theta)}{\int_\Theta \rho(\theta')p(z_1^n;\theta')d\theta'}$$

Bayesian model averaging $\equiv$ use posterior weights to predict

do not go through the posterior weights for individual model classes

Often hoever similar because posterior tends to concentrate around the best-fitting parameter value in each class

NEVER actually select a model, always keep a distribution

Predictive distribution:

$$p(z_{n+1}|z_1^n;\rho) = \int_\Theta p(z_{n+1}|z_1^n;\theta)\rho(\theta|z_1^n)d\theta$$

— This often works because of the bias/variance trade-off: using ρ introduces a bias towards certain parts of model space but $p(z_{n+1}|z_1^n;\rho)$ is less vulnerable to fluctuations than $p(z_{n+1}|z_1^n;\widehat{\theta})$
plus the smoothing effect

Why Doesn't Model Averaging Lead to Over-Fitting?

Why isn't capacity of a big ensemble of experts so large that it always overfits?
Two reasons:

- Benefits of diversity/smoothing
- Models in the ensemble are *not* independent
  all trained on the same/very similar data
  number of effectively-distinct ensembles $\ll$ product of individual capacities

Consistency of Bayesian model averaging is hard, need to handle correlations among posterior densities

also impose constraints like a sieve

unless you cheat and assume that the true model is in your prior

Bayesian updating turns out to be equivalent to replicator dynamics with its own large deviations principle

$$\log \rho(\theta \in A | z_1^n) \approx -n \left(\inf_{\theta \in A} \mathbf{E}\left[L(\theta)\right] - \inf_{\theta \in \Theta} \mathbf{E}\left[L(\theta)\right] \right) + O_p(n^{1/2})$$

Conclusion

Reliability is fundamental; specific criteria are not

Appropriate model selection depends on the purpose of the model

Realistic representation or predictive instrument?

Control of capacity and background assumptions

Consistency of specific techniques for relational data are (largely) open questions

Someone needs to figure our network cross-validation/bootstrapping