Advanced Modeling

Dr. Wray Buntine
Heuristicrats Research, Inc.

wray@Heuristicrat.COM
http://www.Heuristicrat.COM/wray/

1678 Shattuck Avenue, Suite 310 • Berkeley, CA, 94709-1631
Tel: +1 (510) 845-5810 • Fax: +1 (510) 845-4405

Section in the tutorial at *Maximum Entropy and Bayesian Methods,*
Sante Fe, New Mexico, June 31st, 1995.

Outline

- Overfitting
  - due to poor statistical criteria
  - due to inappropriate computational method
- Subjectivity versus Objectivity
  - “objectivity” needs to be carefully defined
  - its all a matter of your decision context
  - in some contexts, subjectivity is unavoidable
- Occam’s Razor

  "Entities are not to be multiplied except of necessity" (from Latin)
  - Bayesian methods provide a coherent implementation strategy
  - Bayesian factors provide a means of comparing models of different complexity

Overfitting: curve-fitting

- have 10-th degree polynomial with Gaussian(0,1) error
- take a fixed sample of 20 points
- fit the Maximum Likelihood (ML) model for different degrees of the polynomial
- notice transition from “underfit” to “overfit”
  - 17-th degree fit almost goes exactly through each point
  - 9-th degree is the closest in this case

(see Gelman et al. 95 for detail on Bayesian linear regression)

Overfitting: plots

See polyfit.ps

Overfitting: posterior samples

- assume "uniform prior" and sample from the posterior distribution for different degrees of the polynomial
- notice variation: overfitting disappears, but underfitting remains

Lesson: overfitting is due to poor statistical criteria or poor approximation

See samplepost.ps

Overfitting: justifying sampling

- sampling is one way to estimate the posterior expected value

$$E_{w|data}(\, y_w(x)\,) \;=\; \int_w y_w(x)\;\; p(w|data)\; dw$$

$$\approx\; \sum_{w\,\in\,\text{posterior-sample}} y_w(x)\;/\;\#\text{posterior-sample}$$

- in general, averaging over "multiple models" gives better estimates, and better quantifies uncertainty
- a large part of Bayesian computation is about approximating an integral such as this

Outline

- Overfitting
  - due to poor statistical criteria
  - due to inappropriate computational method
- Subjectivity versus Objectivity
  - "objectivity" needs to be carefully defined
  - its all a matter of your decision context
  - in some contexts, subjectivity is unavoidable
- Occam's Razor

  "Entities are not to be multiplied except of necessity" (from Latin)
  - Bayesian methods provide a coherent implementation strategy
  - Bayesian factors provide a means of comparing models of different complexity

Objectivity: some notions

- classical Fisherian view of objectivity
  - "Make an inference by *only* considering the data"
  - a noble view but requires inordinate amounts of data
    - from Bayesian perspective, this is required to "swamp" the prior

    e.g., see the work on uniform convergence and worst case bounds for learning (Vapnik, 82, Devroye, 91, Haussler, 92)
- intersubjectivity
  - Kant's notion of a group of scientists with different subjective opinions attempting to reach consensus
  - modeled with a range of priors, see Bernardo and Smith (94)

Objectivity: some notions, cont.

- Bayesian reasoning with an "objective" prior
  - so called non-informative or objective priors are controversial, but
  - *invariance* arguments provide reasonable priors for a range of problems (see Bernardo and Smith, 94; Jaynes 96), e.g.,
    - » scale invariant prior on magnitudes such as a standard deviation
    - » rotation-invariant priors on straight lines
- "public" decisions based on data
  - what is the decision context and who is making the decision?
  - is group concensus to be achieved?

Objectivity vs. decision context

Let X = a grounded proposition, with value in {0,1},

e.g. X = smoking increases propensity for lung cancer by 40%

- my best guess about X at the moment is that its true

  i.e., $p(X \mid data) > 0.5$ using my own subjective prior
- I believe X is probably true and I expect my belief wont change in the future

  i.e., $E_{future\text{-}data}\left(E\left((p(X|data) - X)^2 \mid future\text{-}data, data \right) \right)$ is small

  NB. this evaluates to p(X|data) being near 1

  NB. in practice, also need to consider for a range of priors since you may change your prior as well

Objectivity vs. decision context, cont.

- based on the data, most "reasonable" people should believe that X is probably true

  i.e., under a variety of different "reasonable" priors,

  $p(X|data) > 0.5$
- the "rational man" using an "objective prior" based on invariance argument Y would believe that X is probably true

  i.e., $p(X \mid data) > 0.5$ using the specific "objective" prior

Outline

- Overfitting
  - due to poor statistical criteria
  - due to inappropriate computational method
- Subjectivity versus Objectivity
  - "objectivity" needs to be carefully defined
  - its all a matter of your decision context
  - in some contexts, subjectivity is unavoidable
- Occam's Razor

  "Entities are not to be multiplied except of necessity" (from Latin)
  - Bayesian methods provide a coherent implementation strategy
  - Bayesian factors provide a means of comparing models of different complexity

Occam's Razor: valid justifications

It wont get us into too much trouble in the future as we get more data:

i.e. If we guess something too simple, as we get more data, we'll soon discover our mistakes. The Barron and Cover (91) stochastic complexity argument of convergence in the limit.

It wont get us in too much trouble in the future with our users:

i.e. We wont be blamed for vacillation.

i.e. If we guess something simple, we will have sufficient data at least to choose the best of the simple things, so we wont change our mind much later, except for adding more complexity. The Blumer *at al.* (87) argument of simpler spaces are easier to search statistically.

Its psychologically pleasing:

i.e. Most things we remember are simple too. (We've restructured our memory to make them that way.)

We've set the problem up that way:

i.e. Our choice of variables is carefully made using ones that made related things simpler.

Bayes factors for comparing models

- suppose we believe either a n or m-degree polynomial fits the data; call these models M_1 and M_2 respectively
- a prediction on a new case is then given by:

  E(y(x)|data) =

  $p(M_1|data)\ E(y(x)|data,M_1) + p(M_2|data)\ E(y(x)|data,M_2)$
- some useful quantities to consider here are:

  $p(M_1|data)/p(M_2|data)$ = *posterior odds ratio* for M_1 versus M_2

  $p(M_1)/p(M_2)$ = *prior odds ratio* for M_1 versus M_2

  $\log p(M_1|data)/p(M_2|data)$ = *(posterior) weight of evidence* (after Good)

  $p(data|M_1)/p(data|M_2)$ = *Bayes factor* for M_1 versus M_2

  $p(data|M_1)$ = *evidence* for M_1 = $\int p(data,w\,|M_1)\ dw$
- the following equations hold:

  posterior odds ratio = prior odds ratio * Bayes factor

  Bayes factor for M_1 versus M_2 = evidence for M_1 / evidence for M_2

Bayes factors for comparing models

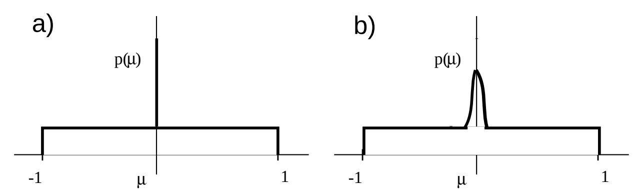
- the evidence can easily become order 2^{-100}, so computation is usually done as posterior weight of evidence
  - divide by the evidence for some “null” model and then calculate in log space
- most model comparison over heterogeneous models (e.g., n-degree polynomial for different values of n) uses:
  - computation or approximation of the Bayes factors
  - special prior to glom the heterogeneous models into a single model family over a well defined parameter space
- use of Bayes factors over heterogeneous models *assumes* Occam’s razor
  - 3-degree polynomials are a set of measure zero in 6-degree polynomials so the prior odds should be 0 (in general) !

Smooth versus discontinuous priors

prior on μ such that 0 is highly likely

a) composes 2 priors of dimension 0 (delta function at 0) & 1 respectively

b) makes a smooth peak at 0 so prior is of dimension 1

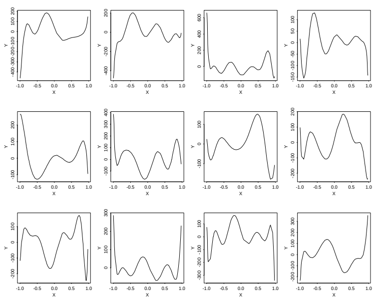


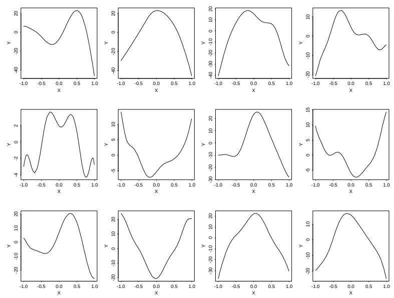
- this situation occurs frequently in complex models when Occam’s razor is required:
  - curve fitting with different dimensional poynomials
  - clustering where the number of hidden classes is unknown
- prior b) may be easy to write search algorithms for
- prior a) requires model comparison and use of Bayes factors to compare the two models of different dimension
- which situation is more realistic, and which is merely an approximation for convenience?

Occam's razor: example priors

- plots on left are sample curves drawn from a uniform prior on 10-degree Legendre polynomials (implicitly used by ML)
- plots on right are sample curves drawn from a prior on 10-degree Legendre polynomials with "almost" scale invariant prior on "average curvature" (see Buntine & Weigend, 91 for prior)

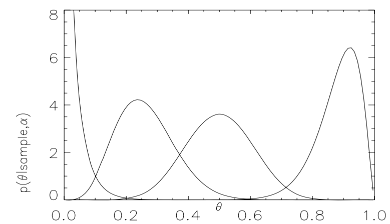
i.e.,





Large sample behavior of the joint

- a sample of posteriors (for different data) from a Bernoulli sampling problem is shown below
- notice they're all peaked, and to a good approximation the peaks are Gaussian (see Kass and Raftery, 93 or any text)



$p(\text{data},\theta|M) = p(\theta|M) \exp\left(- N \ (1/N)\sum_{i=1}^{N} \log 1/p(x_i|\theta,M) \right)$ data = $\{x_1, ..., x_N\}$

$\int f(\theta) \exp(- N g(\theta))\, d\theta$

$\approx \quad f(\theta_0) \int \exp(- N < \text{2nd-order taylor expansion of } g(\theta) \text{ at } \theta_0 >)\, d\theta$

$= \quad \exp(- N g(\theta_0)) \ f(\theta_0) \quad (2\pi / N)^{k/2} \ / \ \det^{1/2} \text{jacobian}_g(\theta_0)$

Laplace approximation where θ_0 is the unique local minima of $g(\theta)$ which must be interior, and $\dim(\theta)=k$

A coarse approximation for evidence

using the large sample behavior, we get p(data|M)

$\approx \quad p(\text{data},w=w_0|M)\ (2\pi)^{k/2} \prod_i (\lambda_i / \sqrt{N})$

where:

w_0 = unique local maxima of the posterior for w

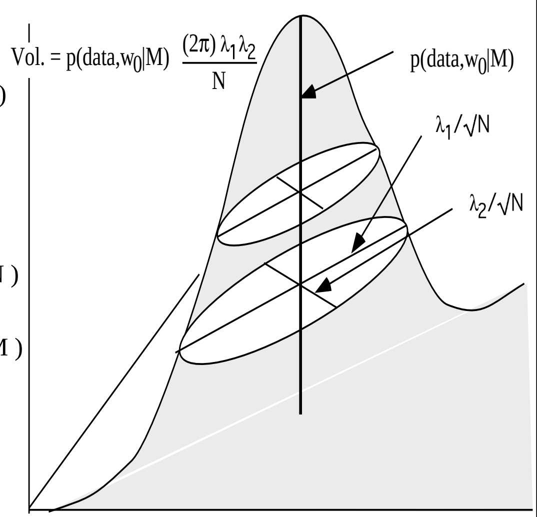
FI(w) = observed Fisher information

= Jacobian($\sum_i \log 1/p(x_i|w,M) / N$)

= Jacobian(average negative log. likelihood for data |w,M)

k = dim(w), N = dim(data)

$\lambda_1, \lambda_2, ...$ = eigenvalues of $FI^{1/2}(w_0)$



Understanding the Bayes factor

$$p(\text{data}|M_1)/p(\text{data}|M_2) \ = \ \frac{p(\text{data},w=w_{1,0}|M_1)}{p(\text{data},w=w_{2,0}|M_2)} \ \frac{(2\pi)^{k1/2} \prod_{i=1}^{k1} (\lambda_{1,i} / \sqrt{N})}{(2\pi)^{k2/2} \prod_{i=1}^{k2} (\lambda_{2,i} / \sqrt{N})}$$

comparative fit to the data

comparative precision of fit

- this approximation is coarse but informative
- all else being equal, the higher dimensional model gets killed by the 2nd term due to $(1/\sqrt{N})^{(k2-k1)}$
- we might expect the higher dimensional model to have a much better fit, but this needs to overcome the 2nd term first
- minimum description length (MDL, and many other names) gains its credence due to a related effect, see Barron and Cover, 91