24. More Model Selection (AIC, BIC, MDL, Morozov...)

Sunday, March 10, 2024 10:35 AM

Technique #4a AIC Akaike Information Criterion (Akaike '73)

Model dependent, but something like

$$AIC = \frac{-2}{m} \cdot \log\text{-likelihood} + 2\frac{d}{m}$$

d ← effective number of parameters (we won't go into detail, but think of it qualitatively like VC dim)

$\frac{-2}{m} \cdot \log\text{-likelihood}$: like $\hat{L}_S$

$2\frac{d}{m}$: penalty related to complexity of model

See § 7.6 Hastie et al. for more

Ex: linear models, if $\hat{y} = P \cdot y$, then $d = \text{trace}(P)$

↳ OLS and Tikhonov cubic smoothing splines

Very "classical", especially in some fields

Ex In ARIMA models in time series, the AICC (= AIC Corrected) is a standard approach to find (p, d, g) parameters.

If noise is Gaussian, under some conditions,

$$AIC = C_p$$

Technique #4b BIC Bayesian Information Criterion aka Schwarz C.T. ('78)

$$BIC = -2 \cdot \log\text{-likelihood} + \log(m) \cdot d$$

d ← again, depends on details of model

So like AIC but changes a "2" into a "$\log(m)$" so it penalizes complex estimators more than AIC

It's asymptotically consistent (if the true model is in our class, it will be selected as $m \to \infty$) which isn't true for AIC

... but for $m < \infty$, benefit is less clear.

Some connections to MDL

Technique #5 Adjusted R^2 ← aka Coefficient of Determination

For linear models, $\hat{y} = X \cdot \omega$,

$$R^2 = 1 - \frac{\sum_{i=1}^{m} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{m} (y_i - \bar{y})^2}$$

$\sum (\hat{y}_i - y_i)^2$: SS_{resid}

$\sum (y_i - \bar{y})^2$: SS_{tot}

$\bar{y} = \frac{1}{m}\sum y_i$

$R^2 \approx 1$ is good

$\in [0, 1]$ (typically)

$$\bar{R}^2 = R^2_{adj} = 1 - \frac{SS_{resid}/df_{res}}{SS_{tot}/df_{tot}}$$

$df_{res} = m - d - 1$ if d = # variables in model (ie. like effective dimension)

$df_{tot} = m - 1$

24a. More Model Selection (AIC, BIC, MDL, Morozov...)

Sunday, March 10, 2024 10:54 AM

Technique #6 MDL (Minimum Description Length) §7.8 Hastie et al.

Idea: data is our message we wish to transmit, and we'll reward models that efficiently encode our message

Coding theory / information theory point of view

At a high level, similar to modern "information bottleneck" ideas

Ideal MDL: Kolmogorov Complexity

(ref: Grünwald's '05 tutorial)

The Kolmogorov complexity of a sequence is the length of the shortest computer program that can produce the sequence

which language? doesn't matter asymptotically since they're all effectively "equivalent norms" since can emulate each other

i.e. $(1,0,1,0,1,0,\dots)$ has low complexity, $O(1)$ independent of length

vs. random sequence of length n has $O(n)$ complexity.

Completely impractical though... equivalent to halting problem!

... in practice and in theory

Practical MDL

Coding theory intro

We want to send a message.

pre-written Hallmark cards

Not an arbitrary message: choose from among $\{z_1, \dots, z_m\}$

ex: $m = 26$, $z_1 =$ "a", $z_2 =$ "b", ...

--- and we'll send a sequence of such messages...

Convert each message to a bit string

Baseline: use $\lceil \log_2 m \rceil$ bits, eg. $m = 26$, use 5 bits

```
00000 = a
00001 = b
00010 = c
00011 = d  etc.
```

Can we do better? i.e. look at $\mathbb{E}$[message length]

"e" is more common than "z", so use 1 or 2 bits for "e",

5 or more for "z" i.e. Variable Length Code

With a variable length code, how do we tell when one message stops and another begins?

Two soln: ① use a "STOP" word in between (like a space) --- but you have to pay for it

② prefix-free codes

24b. More Model Selection (AIC, BIC, MDL, Morozov...)

Sunday, March 10, 2024 4:47 PM

(... coding theory) Instantaneous Prefix-free Code

ex: $m=4$

z_1	z_2	z_3	z_4
0	10	110	111

Idea: we send message z_1 more often than z_2

So 00101110 unambiguous

1 1 2 4 1

then $\mathbb{E}[\text{message length}] = \sum_{i=1}^{m} \mathbb{P}[z_i] \cdot (\#\text{ bits to encode } z_i)$

$\geq \sum_{i=1}^{m} \mathbb{P}[z_i] \cdot \log_2\left(\frac{1}{\mathbb{P}[z_i]}\right) =$ entropy of distribution over z

Shannon

i.e. if uniform over $[m]$, entropy is $\log_2(m)$

and our naive approach was optimal.

Due to discrete nature, we can't always achieve Shannon's lower bound, but there is an optimal code (Huffman coding '52)

Back to MDL ...

transmitting variable z takes $-\log(\mathbb{P}[z])$ bits of information

Data $z=(x,y)$, model param. θ, model M.

If Alice knows x but not y, to send her y, we could send over our classifier h. If h isn't perfect, should also send $y-\hat{y}$ $(\hat{y}=h(x))$.

Our criterion is

No actual coding is involved

$$\text{"message length"} = -\log\left(\mathbb{P}[\overbrace{y}^{\hat{y}} \mid \theta, M, x]\right) - \log\left(\mathbb{P}[\theta \mid M]\right)$$

avg. code length to transmit $y-\hat{y}$

avg. code length to transmit model param. (so you can't cheat)

Ex: model 1, codebook is $\{a,b,c,\ldots,z\}$ to transmit "hello" I encode each letter

model 2, codebook is {"hello", "goodbye"} easy to transmit "hello" but our model now takes up more space

Related to MML = Min. Message Length predates MDL by 10 yrs

simpler, Bayesian: criterion is $L(h) + L(\text{data} \mid h)$, L = length in bits

Rissanen '78, '96. Inspired by AIC See Grünwald '05 for details

24c. More Model Selection (AIC, BIC, MDL, Morozov...)

Sunday, March 10, 2024 5:07 PM

Technique #7 Morozov Discrepancy Principle (Morozov '66) Ref: §7.3 Vogel

Suppose $y = f(x) + \eta$ (additive noise model)
(f: unknown)

If we find a good hypothesis h, then hope $h(x) \approx f(x)$, so $y - h(x) \approx \eta$

$y - h(x)$: residual, which we observe = r

We don't observe noise η directly

of course, but sometimes we know it's distribution

So we want $\|\text{resid}\|^2 = \mathbb{E}\|\eta\|^2$ ($= m\cdot\sigma^2$ if iid noise of variance σ^2)

$\|\text{resid}\|^2$: our criterion

Ex: least squares

① Morozov
$$\min_{w\in\mathbb{R}^d} \tfrac{1}{2}\|w\|_2^2 \quad \text{s.t.} \quad \|Xw - y\|_2 \le \sqrt{m}\cdot\sigma$$

② Tikhonov
$$\min_{w\in\mathbb{R}^d} \frac{1}{m}\|Xw - y\|_2^2 + \alpha\|w\|_2^2$$

often easier to estimate than α or τ

③ [No Russian name]
$$\min_{w\in\mathbb{R}^d} \frac{1}{m}\|Xw - y\|_2^2 \quad \text{s.t.} \quad \|w\|_2 \le \tau$$

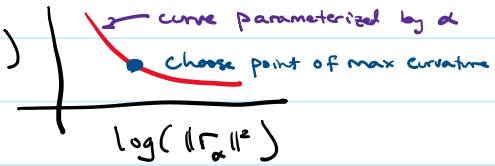
All 3 versions are convex optim. problems, easy to solve, and via convex analysis and Lagrange multipliers we see they're all essentially the same (ie. $\forall\ \alpha, \exists\ \sigma$ s.t. $w_\alpha = w_\sigma$ etc)

w/ a few minor technical caveats ...

Technique #8 "L-Curve Method" aka you can publish papers on weird stuff cf. Vogel

eg. w_α is Tikhonov soln, $r_\alpha = Xw_\alpha - y$ is residual.

Plot $\log(\|w_\alpha\|^2)$ vs $\log(\|r_\alpha\|^2)$



Not statistically consistent

A hack, but really so is everything else (since most assumptions not true in real world)

In clustering (to choose # of clusters) this is known as the elbow method