STAT 206
3 Mar 21

DD Am discussion section

Q: How do we know when a statistical model is 'good'?

useful

A: A model is 'good' if it makes accurate & well-calibrated predictions of data ① using not used in fitting. it

out-of-sample predictive validation

out-of-bag (OOB)

models, fitting means estimating its parameters & obtaining well-calibrated uncertainty assessments about (a) those parameters & (b) predictions of new data

with parametric

To instantiate ⊛ we need a measure of goodness of fit (GoF) of a model

single example

$$\left\{ \begin{array}{l} (\mu, \sigma) \sim \text{LI}(\mu, \sigma) \\ I_i \mid \mu, \sigma \overset{\text{IID}}{\sim} N(\mu, \sigma^2) \end{array} \right\} \quad (i = 1, \ldots, n) \qquad ⊛⊛$$

for even moderate n we showed (when looking at Bayesian analyses of Gaussian sampling models) that the predictive distribution for a new data value Y_{new} is

$$\underline{y} = (y_1, \ldots, y_n) \qquad \bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$(Y_{new} \mid \underline{y}\ [SIGMA][PM:LI]\ B) \sim N(\bar{y}_n, s_n^2)$$

$$s_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

with a few batch of data

$$\underline{y}_{new} = (y_{n+1}, \ldots, y_{n+\ell})$$

a natural goodness-of fit measure is

$$\frac{1}{\ell}\sum_{i=1}^{\ell}(y_{n+i} - \bar{y}_n)^2 \triangleq GoF$$

this is the squared Euclidean distance between the vector $\underline{y}_{new}$ and the scalar $\bar{y}_n$, scaled by sample size to create a mean discrepancy

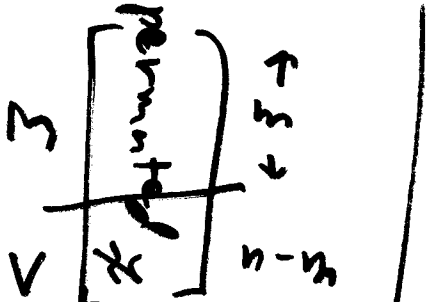
but we can't compute this until $\underline{y}_{new}$ arrives

solutions: ① cross-validation | ② penalized complexity

Cross validation (CV) ① "1-way CV": use all of $\mathcal{Y}$ (cheating) to assess predictive quality by calculating $\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2$: this is easy but may well understate our actual uncertainty about $\mathcal{Y}$ now, because we used the dataset twice: once to set $\bar{y}_n$, and once to see how good $\bar{y}_n$ is (this can be remedied with a complexity penalty [see ② below])

③ $\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \mathcal{Y}_n$, observed data vector

"2-way CV" (A) partition randomly $\mathcal{Y}$ into modeling (M) and validation (V) subsets of size m & $(n-m)$ (respectively) for a choice of $1 \le m \le n-1$; (B) use the mean of the (M) subset values to predict the (V) subset values & compute the GoF summary on (V)



(C) repeat many times (A),(B) & summarize (4) & average the results (R code)

Q: How does this generalize to other models?

A: The "2 way CV" structure is already completely general for settings in which the dataset $\mathcal{D}$ is a vector $\underline{y}$ of real numbers

But the natural goodness of fit measure in other models is less clear. It turns out that looking at the log likelihood function is illuminating. In model ⊛⊛ (p. 1 of Hw) this is $\ell\ell(\mu, \sigma \mid \underline{y}\ [SM:N]\ B) = -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2$ recall that the MLEs in this model are $\hat{\mu}_{MLE} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$, so the maximum log lik.

value is $ll(\hat{\mu}_{MLE}\ \hat{\sigma}_{MLE} \mid x[SM:NV]\ B) =$

$$-\frac{n}{2}\log\left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\right] - \frac{n}{2}\log 2\pi - \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\frac{2}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

$$= -\frac{n}{2}\left\{\log\left[\left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\right]\right] - \log 2\pi - 1\right\}$$

But this is the natural GoF measure for the Gaussian model (!) This suggests a general rule: use the maximum log likelihood value to define goodness of fit

in practice people use

$$-2\cdot ll(\underset{\sim}{\hat{\theta}} \mid D[SM]\ B)$$ as the general GoF measure, with small values ↔ good fit

$-2\,ll(\hat{\underset{\sim}{\theta}} \mid D\,[SM]\,B) \triangleq -2ll_{max}$ is called the frequentist deviance of the model based on IID realizations from the sampling model [SM]

The multiplication by (−2) is because of a frequentist story under which it turns out that $(-2ll_{max})$ can be calibrated by comparing it to values of the χ^2 distribution (a special case of the $\Gamma(a,b)$ family)

However, the deviance is based on the "1 way CV" idea of using the entire dataset to compute ll_{max}; to avoid over-fitting

it turns out that we need to penalize the model leading to $\mathcal{L}_{max}$ according to its complexity (otherwise we could set a superb value of $\mathcal{L}_{max}$ just by fitting the data perfectly

scatterplot

model (polynomial of order $n=7$, n = # data points) with superb value of $\mathcal{L}_{max}$ but that validates well out of sample

model (linear) that validates poorly out of sample (high complexity)

has a small $\mathcal{L}_{max}$ but low complexity

Important note

Q: Since the likelihood function is only defined up to a positive constant multiple, can't we just make $\mathcal{L}_{max}$

as large as we like just by choosing that constant?

A: Yes; that's temporarily embarrassing. The fix is as follows: when computing the likelihood function for model comparison & assessment of GoF, (a) write down the joint sampling distribution with its correct normalizing constant; (b) define the likelihood function for GoF evaluation as equal to the function in (a), including the normalizing constant from (a) in this definition (I did this in the example above but didn't mention it until now).

3 math impulses following development of a good idea: ① try to break it, to find out its scope of validity; ② try to generalize it; ③ use it to unify ideas that were previously regarded as separate

key statistical unifying idea for parametric sampling models: the Exponential Family (EF)

(not to be confused with the Exponential sampling model, which is (confusingly) itself a member of the EF)

In 1935-1936, 3 sets of researchers

(~~Cambridge~~) E.J.G. Pitman & John Wishart

(École Normale Supérieure) G. Darmois

(Columbia) B.O. Koopman

independently & more or less simultaneously published the same excellent idea, on which the EF is based

(Lecture Notes part 6 pp. 149 →)

General Parametric model for a data vector

$\underset{\sim}{y} = (y_1, \ldots, y_n)$

$\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k)$

sampling model

$p(y_i \mid \underset{\sim}{\theta}\ [SM]\ B)$

↑ marginal PMF/PDF for a single observation y_i

joint sampling model

$(Y_i \mid \underset{\sim}{\theta}\ [SM]\ B) \overset{IID}{\underset{\sim}{}}$

$p(y_i \mid \underset{\sim}{\theta}\ [SM]\ B)$

$(i = 1, \ldots, n)$

$p(\underset{\sim}{y} \mid \underset{\sim}{\theta}\ [SM]\ B) = \prod_{i=1}^{n} p(y_i \mid \underset{\sim}{\theta}\ [SM]\ B)$

← this is a member of the EF

iff it can be written →

$f(\underset{\sim}{y})\, g(\underset{\sim}{\theta}) \cdot \exp\left[\sum_{j=1}^{k} \phi_j(\underset{\sim}{\theta}) \sum_{i=1}^{n} h_j(y_i)\right]$

2 immediate advantages

① a set of sufficient statistics is then (minimal) obvious:

$$\left\{ \sum_{i=1}^{n} h_1(y_i), \ldots, \sum_{i=1}^{n} h_k(y_i) \right\}$$

---

② this sampling model must have a conjugate prior, which is as follows:

$$p(\underset{\sim}{\theta} \mid [cp]\ \beta) = c\, g(\theta)^{\tau_0} \exp\left[\sum_{j=1}^{k} \phi_j(\theta) \tau_j \right]$$

(conjugate prior)

for some vector $\underset{\sim}{\tau} = (\tau_0, \tau_1, \ldots \tau_k)$

---

example | $(Y_i \mid \theta\ \beta) \overset{IID}{\sim} \text{Bernoulli}(\theta)$ (each) $(i = 1, \ldots, n)$

$k=1$ | $(0<\theta<1)$

$\underset{\sim}{\theta} = (\theta)$ | $p(y_i \mid \theta\ \beta) = \theta^{y_i}(1-\theta)^{1-y_i}$

$$p(\underline{x} \mid \theta B) = \prod_{i=1}^{n} p(x_i \mid \theta B) = \theta^s (1-\theta)^{n-s}$$

for $s = \sum_{i=1}^{n} x_i$ — to be in the EF club this has to be expressible as $f(\underline{x}) g(\theta) \exp\left[\phi_1(\theta) \sum_{i=1}^{n} h_1(x_i)\right]$ $(k=1)$

They really don't look the same, but here's a final EF trick: for all $x > 0$, $x = \exp[\log(x)]$.

$$\theta^s (1-\theta)^{n-s} = \exp\left\{\log\left[\theta^s (1-\theta)^{n-s}\right]\right\}$$

$$= \exp\left[s \log\theta + (n-s)\log(1-\theta)\right]$$

$$= \exp\left[s \log\theta + n\log(1-\theta) - s\log(1-\theta)\right]$$

$$\theta^s(1-\theta)^{n-s} = \exp\left[n\log(1-\theta) + s\log\left(\frac{\theta}{1-\theta}\right)\right]$$

$$= \exp\left[n\log(1-\theta)\right] \cdot \exp\left[(s)\log\left(\frac{\theta}{1-\theta}\right)\right]$$

$$= \underbrace{1}_{f(y)} \times \underbrace{(1-\theta)^n}_{g(\theta)} \exp\left[\underbrace{\sum_{i=1}^n h_1(y_i)}_{\text{with } h_1(y_i) = y_i} \cdot \underbrace{\log\left(\frac{\theta}{1-\theta}\right)}_{\phi_1(\theta)}\right]$$

logistic regression

EF ✓✓

$\phi_1(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$ is called the natural parameterization of the Bernoulli(θ) sampling distribution

minimal

The sufficient statistic is $s = \sum_{i=1}^n h_1(y_i)$

and the conjugate prior, which we know is

$$p(\theta \mid [CP]\,\mathcal{B}) = c\,\theta^{\alpha-1}(1-\theta)^{\beta-1},$$

can be written as $c\,g(\theta)^{\tau_0}\exp\left[\phi_1(\theta)\tau_1\right]$:

$$= c\,(1-\theta)^{n\tau_0}\exp\left[\tau_1 \log\left(\frac{\theta}{1-\theta}\right)\right]$$

Some EF sampling models are regular

& some are not

Def. An EF sampling model is regular if the support of the marginal sampling dist. of $\underset{\sim}{Y}$ does not depend on $\underset{\sim}{\theta}$

with parameter vector $\underset{\sim}{\theta}$

ex. ① $(Y_i \mid \mu, \sigma \; [SM:N] \; \underset{\sim}{\theta}) \overset{IID}{\sim} N(\mu, \sigma^2)$ $(i = 1, \dots, n)$

$\underset{\sim}{\theta} = (\mu, \sigma)$ or also (μ, σ^2) or also $(\mu, \frac{1}{\sigma^2})$

$\text{support}(Y_i) = \mathbb{R}$ does not depend on $\underset{\sim}{\theta}$

so this SM is a regular member of the EF

$V(\hat{\mu}_{MLE}) = O(\frac{1}{n})$

ex. ② $(Y_i \mid \theta \; [SM:U] \; \underset{\sim}{\theta}) \overset{IID}{\sim}$ Uniform$(0, \theta)$ $(i = 1, \dots, n)$ $(\theta > 0)$

$\text{support}(Y_i) = (0, \theta)$ this depends on θ

this is a non-regular member of EF

$V(\hat{\theta}_{MLE}) = O(\frac{1}{n^2})$ (!)

Edwards, Lindman(n), & Savage ~ 1960s

psychometrics

LI prior:

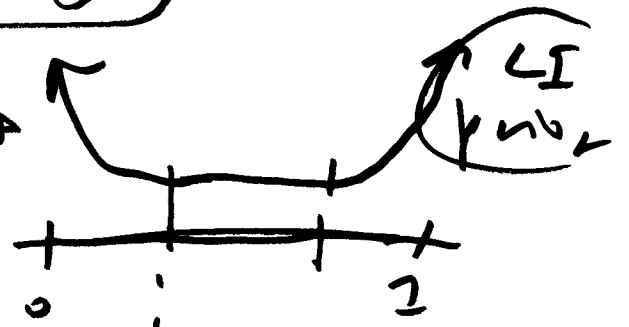
low info =

flat =

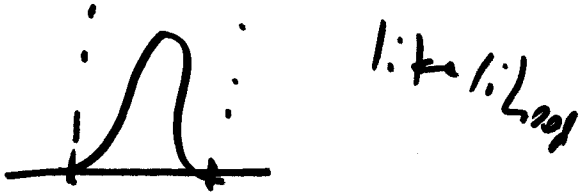
diffuse =
(large spread)

= noninformative

Beta($\frac{1}{2}$, $\frac{1}{2}$)



principle of stable estimation



U(left, right)

to obtain a LI prior, always choose 'left' & 'right' to avoid

inappropriate function of lik. fn

$$p(y_{new} \mid \underset{\sim}{y}\,[SM][PM]\,B) \quad (\star) \quad \overset{\text{(hard)}}{=}$$

k $\underset{\sim}{\theta}_i$ in $[SM]$

$$= \int p(y_{new}, \underset{\sim}{\theta} \mid \underset{\sim}{y}\,[SM][PM]\,B)\, d\underset{\sim}{\theta} \quad (\star\star)$$

$p(y_{new} \mid \underset{\sim}{\theta} \cdot [PM]\,B)$ (easy)

sampling distribution for y_{new}

$$= \int p(y_{new} \mid \underset{\sim}{\theta}\ \cancel{\underset{\sim}{y}}\ [SM][PM]\,B)\cdot$$

$$\text{post. dist.} \rightarrow p(\underset{\sim}{\theta} \mid \underset{\sim}{y}\,[SM][PM]\,B)\, d\underset{\sim}{\theta} \quad (\star)(\star)$$

---

how to MCMC sample from this predictive distribution

$(\star)(\star)$ is actually telling a mixture story in disguise:

$(\star)$ is a weighted average of [sampling distribution given $\underset{\sim}{\theta}$] weighted by the post. dist for $\underset{\sim}{\theta}$

how to make random draws from ⊛

① make a random draw of $\underset{\sim}{\theta}$ from the posterior, obtaining $\underset{\sim}{\theta}^*$

② make a random draw from the sampling dist. using the $\underset{\sim}{\theta}^*$ value from ①