

Model diagnostics and information criteria

ISTA 410 / INFO 510: Bayesian Modeling and Inference

U. of Arizona School of Information

March 8, 2021

Last week:

- MCMC!

Today:

- Information theory and predictive accuracy
- Scoring models to avoid overfitting

Note: no class on Wednesday

Some loose ends from MCMC

Gibbs sampler

One of the simplest Markov chain algorithms, but still useful, is called the *Gibbs sampler* (named for, but not created by, statistical physicist Josiah Willard Gibbs).

Also called *alternating conditional sampling*:

1. Start with a parameter vector $(\theta_1, \theta_2, \dots, \theta_d)$
2. In one iteration,
 - 2.1 Choose an ordering of coordinates 1- d
 - 2.2 In each coordinate, update θ_j according to its distribution conditional on the other θ_i (which are held fixed)
3. After these d steps, have a new parameter vector $(\theta'_1, \theta'_2, \dots, \theta'_d)$

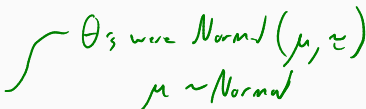
draw a value from
 $p(\theta_j | \theta_1, \theta_2, \dots, \theta_i)$

Gibbs sampler

The Gibbs sampler is useful when the posterior distribution of each θ_i is, conditional on all other parameters, a distribution we can sample directly from

True for, e.g.:

- Hierarchical normal model
- Gaussian mixture models
- Many other models that use conjugate distributions



$\theta_i \text{ were } \text{Normal}(\mu, \tau)$
 $\mu \sim \text{Normal}$

Diagnosing MCMC problems

Three first-line tools for inspecting the result of your Markov chains:

- Trace plots – look for the "fuzzy caterpillar"
- Divergences and pairplots. If you get a lot of divergences, there's trouble. If you get a handful of divergences, there *might* be trouble. Use a pairplot to examine where the divergences are. ✱
- \hat{R} (aka Gelman-Rubin statistic) – used for assessing convergence to the target distribution. Works by comparing between-chain variance to within-chain variance. Should be close to 1 (e.g. $< 1.1, 1.05$). ✱

A slight warning

A slight warning:

- Most diagnostics work by comparing chains or inspecting their local properties
- Global properties of the posterior distribution are “invisible”
- If there is a region of the posterior with substantial probability mass that is hard to access, it may not be reflected in the sample *and you might not see that*
 - Multimodality is a big problem here
 - Even if the sampler reaches all peaks, estimating how much mass is in each peak is tough

Various warnings you might get

- Maximum tree depth warnings: means that NUTS hit a step cap before reaching the U-turn criterion. Most common when priors are flat and parameters are far from 0 – e.g. ordinary linear regression with non-standardized data. Apply weakly regularizing priors.
- Acceptance rate warnings: observed acceptance rate higher or lower than the target. Means the sampler didn't tune the internal parameters quite right – increase the tuning interval (or ignore)
- Low effective sample size: means that new draws are highly correlated with one another. How bad this is depends on what you want out of your model. Means and variances can often be estimated with $ess \approx 200$.

Model selection and information theory

Comparing models

One of the major problems in applied statistics is the problem of choosing between different models, or different sets of predictors.

Two distinct goals:

- Predictive accuracy: a model should produce predictions that agree with observed data
- Causal explanation: a model should inform causal relationships between observed variables

Today's tools focus on the first case. So what makes a prediction "good"? What makes models perform well or poorly?

Recommended reading: Statistical Rethinking ch. 7

BDA ch. 7

Overfitting and underfitting

Two ways a model can fail:

- *overfitting*: a model learned too much from the data; model fit random noise that doesn't generalize to new observations; usually, because of too many parameters
- *underfitting*: a model learned too little from the data; model is unable to fit the true relationships in the data; usually, because of insufficient parameters

In order to make sense of these, we need a way to assess whether the model is accurately fitting the data. For that, we need some information theory.

A little information theory

The main contribution of information theory to statistics is a measurable notion of uncertainty.

What is uncertainty?

- We don't know the value of future observations yet

The main contribution of information theory to statistics is a measurable notion of uncertainty.

What is uncertainty?

- We don't know the value of future observations yet
- However, we know something about them (predictive distribution)
- The more “flat” the probability distribution, the more uncertainty

Uncertainty: example

A simple example of varying uncertainty: weather forecasts.

- On any given day, what's the weather like in Tucson?

Uncertainty: example

A simple example of varying uncertainty: weather forecasts.

- On any given day, what's the weather like in Tucson?
- On any given day, what's the weather like in Seattle?

Uncertainty: example

A simple example of varying uncertainty: weather forecasts.

- On any given day, what's the weather like in Tucson?

sunny, warm

$$p(\text{sunny}) \gg p(\text{cloudy})$$

- On any given day, what's the weather like in Seattle?

cloudy, rainy, cool

$$p(\text{sunny}) \ll p(\text{cloudy})$$

- On any given day, what's the weather like in Chicago?

... could be anything.

$$p(\text{sunny}) \approx p(\text{cloudy})$$

Uncertainty: example

A simple example of varying uncertainty: weather forecasts.

- On any given day, what's the weather like in Tucson?
- On any given day, what's the weather like in Seattle?
- On any given day, what's the weather like in Chicago?

Not much uncertainty for Tucson or Seattle; a lot for Chicago

Information entropy

Measurement for uncertainty: *information entropy*. Introduced by Claude Shannon (1947) at Bell Labs; named for similarity to thermodynamic entropy.

If p is any probability distribution:

$$H(p) = - \sum_i p_i \log_2(p_i)$$

convention, to make this positive.

- The base 2 is a convention, and sets the “units” of uncertainty to “bits”; one bit is the amount of uncertainty in a fair coin flip or yes/no question. Natural log also used sometimes.
- Difficult to interpret in isolation; but comparison across distributions with the same sample space is useful
- Key property: maximized by flat distributions.

Information entropy

Where did this definition come from? Three assumptions/targets:

- The measure of uncertainty should be continuous, so that small *changes don't change output too much.*
- More possible outcomes should mean more uncertainty
- Uncertainty about independent observations should be additive

This determines the function up to a constant (i.e. a unit of measurement)

Entropy and encoding

History: symbol codes

- Goal: encode information (e.g. text messages) into sequences of bits (0/1)
- Assign a bit string (called a code word) to each symbol in the alphabet
- How many bits does each symbol need?

Entropy and encoding

History: symbol codes

- Goal: encode information (e.g. text messages) into sequences of bits (0/1)
- Assign a bit string (called a code word) to each symbol in the alphabet
- How many bits does each symbol need?
- Exploit symbol frequencies: assign shorter code words to more common symbols
- Theoretical minimum *average* length: entropy of the frequency distribution

Kullback-Leibler divergence

Kullback-Leibler (KL) divergence:

$$D_{KL}(p, q) = \sum p_i (\log_2 p_i - \log_2 q_i)$$

H(p_i)

Entropy calculated with log probs of q averaged over p.

Interpretation:

= extra uncertainty from computing with "wrong" p-probabilities q_i

- p is the true outcome distribution
- q is the model predictive distribution
- KL divergence measures incorrectness, in some way

A nice interpretation of KL divergence is as a “potential for surprise.” (The idea of surprise as a measurable quantity is all over information theory.)

Imagine two scenarios:

- You raise a dog in Chicago, and then you move here to Tucson
- You raise a dog in Tucson, and then you move to Chicago

Potential for surprise

The weather in Chicago is highly variable:

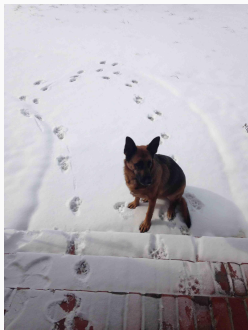
- It's hot and humid in the summer
- It's bitterly cold in the winter
- Sometimes it just oscillates between the two on a daily basis

Your Chicagoan dog has experienced all kinds of weather, and will be comfortable in the heat and the cold

Potential for surprise

As previously noted, the weather in Tucson is pretty consistent.

Your Tucsonan dog, upon moving to Chicago:



Asymmetry in KL divergence

This is reflected by the asymmetry in KL divergence.

City	Tucson	Chicago
p_{hot}	0.95	0.5
p_{cold}	0.05	0.5

$D_{KL}(\text{Tuc}, \text{Chi}) \approx 0.714$

$$D_{KL}(\text{Chi}, \text{Tuc}) \approx 1.198$$

$D_{KL}(\text{Chi}, \text{Tuc}) \approx 1.198$

Asymmetry in KL divergence

Points of statistical interpretation:

- A flat model is closer to a nonflat model than vice versa
- Advantage to simpler models: they have higher entropy
- Maximum-entropy distributions:
 - include the weakest assumptions (example: for a given mean/variance, a normal distribution maximizes entropy)
 - are closest to other distributions

Information criteria for scoring models

Kullback-Leibler divergence

Kullback-Leibler (KL) divergence:

$$D_{KL}(p, q) = \sum p_i (\log_2 p_i - \log_2 q_i)$$

Interpretation:

- p is the true outcome distribution
- q is the model predictive distribution
- KL divergence measures incorrectness, in some way

KL divergence for model comparison

We think of $D_{KL}(p, q)$ as measuring the distance from our model, q , to the truth, p .

Problem: we don't know p and never will!

But this isn't an obstacle for comparing models, because if we have two models q and r , then

$$D_{KL}(p, q) - D_{KL}(p, r) = \sum p_i(\log_2(r_i) - \log_2(q_i))$$

i.e. the $H(p)$ term drops out. We still don't know p_i , but we can estimate this from a sample of observations (because the observations are drawn from p_i); e.g. from a MCMC trace

Log score and deviance

Scoring models using log probabilities:

$$\text{log score} \quad S(q) = \sum \log(q_i)$$

Deviance:

$$D = -2S(q) = -2 \sum \log(q_i)$$

(What's the factor of -2 about?)

In Bayesian world, the posterior isn't one model, it's a distribution of models – so we should average:

$$\text{lppd}(y, \theta) = \sum_i \log \left(\frac{1}{S} \sum_s p(y_i | \theta_s) \right)$$

(log pointwise predictive density)

Out-of-sample prediction error

lppd isn't enough on its own, though, because it only looks inside the sample

- Adding parameters nearly always improves fit within the sample
- Eventually, adding parameters reduces accuracy out of the sample (overfitting)
- How can we predict out-of-sample prediction accuracy?
 - Cross-validation
 - Information criteria

Since we use lppd to estimate the fit of our model, our goal with all of these tools is to estimate what our lppd will be on new data.

In other words, ultimately we are trying to estimate some form of:

$$\text{elpd} = \mathbb{E}(\log p(\tilde{y}|y))$$

the expected log predictive density of a new data point.

In some cases (e.g. with AIC) we'll calculate the expected deviance of a new data set of the same size (can work either way).

Overfitting in action

To demonstrate overfitting, we'll consider a few models fit to fake data.

True data-generating process:

$$y_i \sim \text{Normal}(\mu_i, 1)$$

$$\mu_i = 0.15x_1 - 0.40x_2$$

We'll fit models with the same likelihood and

$$\mu_i = \alpha$$

$$\mu_i = \alpha + \beta_1 x_1$$

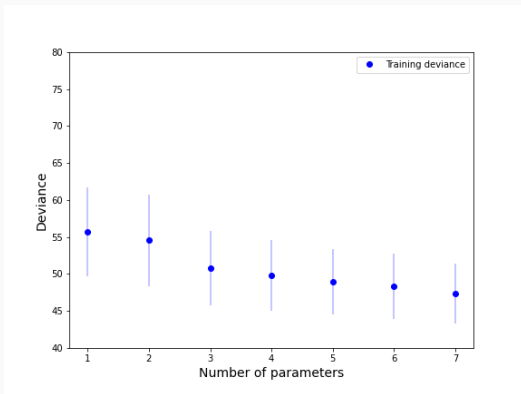
$$\mu_i = \alpha + \beta_1 x_1 + \beta_2 x_2$$

$$\mu_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

...

Overfitting in action

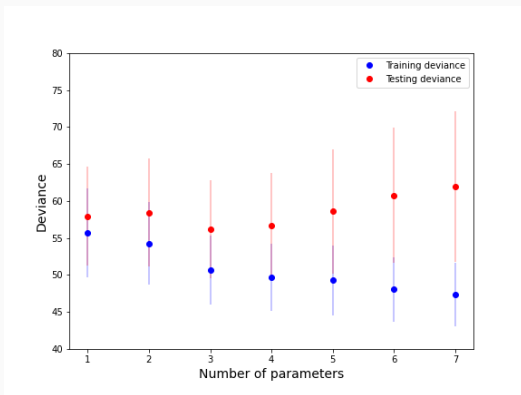
On the training set:



Remember: past 3 parameters, the predictors have no relationship to y in the true data generating process

Overfitting in action

Add in the testing set:



As expected, the additional parameters just make matters worse.

Akaike information criterion (AIC)

AIC: named for Akaike (but he called it “an information criterion”). Attempts to estimate the out-of-sample deviance.

Assuming a point estimate $\hat{\theta}$ for model parameters, calculate the log score and apply a penalty to correct for overfitting:

$$AIC = D_{\text{train}} + 2k$$

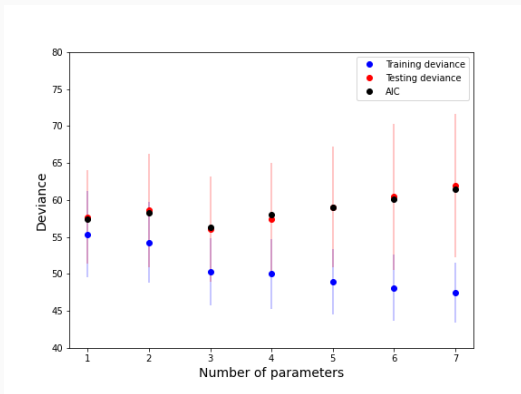
Here D is the log predictive density evaluated at the maximum likelihood estimate, multiplied by -2.

k is the number of parameters. Assumes Gaussian posterior.

Where it comes from: Taylor expansion around the posterior mode.

Overfitting in action

Adding the AIC:



We see that for this model, the AIC is a good estimate of the out-of-sample deviance. So, it is reasonable to choose the model with the smallest AIC.

Today:

- information theory / entropy
- A first information criterion: AIC

Next time:

- Better criteria for estimating out-of-sample error:
 - WAIC – a refinement of AIC
 - PSIS (or LOO-CV) – another estimate of out-of-sample error