EPSE 581C: Bayesian Methods

Ed Kroc

University of British Columbia

ed.kroc@ubc.ca

November 25, 2019

Today

- Cross-validation
- Information criteria
- Measurement error modelling
- Missing data
- Priors

Return to model building and validation

- Information criteria for model selection
- Cross-validation techniques

Akaike (An) Information Criterion (AIC)

- Every statistical model has an associated AIC.
- Let k be the number of model parameters and let $\widehat{L}$ be the maximum value of the likelihood function (given those parameters); then define the AIC of the model as follows:

$$AIC = 2k - 2\ln(\widehat{L}).$$

- Among a set of candidate models, the one with the *lowest* AIC is the "best", according to Akaike's information criterion.
- Informally, AIC penalizes a model for overfit (too many parameters), and rewards a model for optimizing the likelihood.

Quantifying information content

- Formally, AIC is a *relative comparison* of *information loss*.
- Define the *Kullback-Leibler divergence* (*relative entropy*) between two probability density functions f and g as:

$$KL(f, g) = -\int_{-\infty}^{\infty} f(x) \cdot \log\left(\frac{g(x)}{f(x)}\right) \, dx,$$

  and analogously for probability mass functions p and q as

$$KL(p, q) = -\sum_{x} p(x) \cdot \log\left(\frac{q(x)}{p(x)}\right)$$

- This is the amount of information lost when we use g (or q) to estimate f (or p).
- Notice: if $p = q$, then $KL(p, q) = 0$.

Quantifying information content

- Formally, AIC is a *relative comparison* of *information loss*.
- Suppose the data come from some *true* (unknown) data-generating process given by the function f.
- We could then calculate the information lost when estimating this reality by some other model, with associated likelihood g.
- Of course, we never know f, but Akaike showed that the AIC is a good estimate (under some conditions) of the information lost when using g instead of f.
- Less information loss is good, so we want usually want models that have the smallest AIC (could be a large negative number!)

Akaike (An) Information Criterion (AIC)

- The formal AIC justification is only valid asymptotically (as sample size grows without bound).
- For small sample sizes, one usually uses the modified AIC instead:

$$AICc = AIC + \frac{2k^2 + 2k}{n - k - 1},$$

  where n is sample size and k is the number of model parameters.
- Using information theory, one can show that the AIC will settle on the "best" *predictive* model, under certain conditions.
- AIC (AICc) is always a *relative* measure of model fit; i.e. the exact values of the AIC are always *totally meaningless* on their own.

Bayesian Information Criterion (BIC)

- Every statistical model has an associated BIC (also called SIC for Schwarz).
- Let k be the number of model parameters, n be the sample size, and let $\widehat{L}$ be the maximum value of the likelihood function (given those parameters); then define the AIC of the model as follows:

$$BIC = \ln(n)k - 2\ln(\widehat{L}).$$

- Among a set of candidate models, the one with the *lowest* BIC is the "best", according to Bayesian information criterion.
- Informally, BIC penalizes a model for overfit (too many parameters) proportionally to a function of sample size, and rewards a model for optimizing the likelihood.

Bayesian Information Criterion (BIC)

- Similar information content interpretations/justifications exist for the BIC.
- Using these, one can show that the BIC will settle on the "best" explanatory model, under certain conditions.
- As with the AIC, BIC is always a *relative* measure of model fit.
- BIC makes sense *even if you are performing a frequentist analysis.* Why?
- Informally, one can show that the BIC assumes a uniform prior on all candidate models while the AIC assumes a different prior on the set of candidate models.

Deviance Information Criterion (DIC)

- The DIC is commonly used as a model selection index/techniques when models are being fit via MCMC algorithms.
- The exact formula for DIC is similar to AIC/BIC, but relies on the *effective number of parameters* of a model. This is a theoretical quantity that cannot be directly calculated, and it reflects the MCMC (hypothetical sampling) uncertainty in our model estimates.
- In a certain sense, the DIC generalizes the AIC to situations where we can only approximately estimate our model via MCMC techniques.
- As with all ICs, smaller numbers represent "better" fits, and the exact values of the DIC are meaningless except *relative* to each other for a particular analysis.

Cross-validation

- Cross-validation is yet another extensively used techniques for model selection and validation.
- Tools can be applied in either a frequentist of Bayesian context, but particularly common in the Bayesian framework.
- The simplest kind of cross-validation is *leave-one-out cross validation*

Leave-one-out cross-validation

Proceed as follows:

(1) Uniformly at random, exclude *one* data point from your sample of size n. This is the *validation* set.

(2) Fit your proposed model on the remaining $n - 1$ data points. This is the *training* set.

(3) Quantify how well the fitted model "predicts" the originally excluded data point (perhaps by computing its posterior predictive probability)

(4) Repeat steps (1)–(3) many times, arriving at measure of average predictive fit (average posterior predictive probability).

Can compare different models according to this "cross-validation error" then, finding which one does the best job at predicting the artificially missing data.

Cross-validation

- Obviously not feasible to do cross-validation by-hand.
- Many ways to generalize or alter the cross-validation approach.
  - Could put $p > 1$ points in the validation set.
  - Could exclude, say, 10% of your data points for validation.
- Asymptotically, cross-validation will settle on the "best" predictive model; i.e. cross-validation will settle on the same model as the AIC, with large enough sample sizes.

Model selection/building/validation

- If you have *large* datasets (say, $n > 1,000$) with many potential parametes (say, $k > 50$), then the above tools can be very useful.
- If you are in "small data" situations, then these tools can still be used, but they are very likely to be *not* as useful as simply examining your model residuals.
- The above tools have rich and beautiful mathematical theories that motivate and justify their use; in practice, however, people tend to just use the tools as a way to not have to think critically about which model *they should actually choose*.

Measurement error modelling

- The problem of *measurement error* is a classic one, with a rich history in a variety of applied disciplines.
- Most classical concepts of measurement error propose an additive model that relates the *observed measurement* X^* to the *unobserved true variable of interest* X as follows:

$$X = X^* + E,$$

  where E is some error. We often say that X^* is a *proxy* for X.
- This is the basic model (with additional structures) behind classical test theory and factor analysis.
- This is also the basic model behind classical concepts of measurement error in economics and the health sciences.

Measurement error modelling

- To implement the model $X = X^* + E$ in practice, we generally assume some kind of parametric structure to it; e.g. in factor analysis, X^* is assumed to be normally distributed with mean X.
- This allows one to write down a corresponding *likelihood* for the measurement error model.
- This can then be combined with any other likelihood relating the measurement X to other variables of interest (e.g. regression modelling, ANOVA).

Measurement error modelling

- More specifically, suppose we want to relate variables X and Y, but we only measure X via an error-prone proxy, X^*.
- To do any modelling/inference, we need to account for the measurement error while also proposing a model relating X to Y.
- But we never actually observe X; this would be a big problem for inference, since we cannot then plug into the likelihood $f(y, x)$.
- Instead, we have:

$$\begin{aligned} f(y, x^*) &= \int f(y, x, x^*)\ dx \\ &= \int f(x^* \mid y, x) \cdot f(y \mid x) \cdot f(x)\ dx \end{aligned}$$

Measurement error modelling

- Instead, we have:

$$f(y, x^*) = \int f(y, x, x^*)\ dx$$
$$= \int f(x^* \mid y, x) \cdot f(y \mid x) \cdot f(x)\ dx$$

- We call the first term, $f(x^* \mid y, x)$, the *measurement (error) model* and the second term, $f(y \mid x)$, the *response model*; third term is the *exposure model*.
- The measurement model characterizes our model for $X^* = X + E$. Note: often assume that other variables Y have nothing to do with this model so that $f(x^* \mid y, x) = f(x^* \mid x)$.
- Thus, we reduce to

$$f(y, x^*) = \int f(x^* \mid x) \cdot f(y \mid x) \cdot f(x)\ dx$$

Measurement error modelling

$$f(y, x^*) = \int f(x^* \mid x) \cdot f(y \mid x) \cdot f(x) \, dx$$

- Inference/modelling between X and Y can now be accomplished via information on X^* and Y. Why?
- Notice that we observe everything needed to plug into the density functions.
- Notice that we *integrate out* the dependency on the unobserved X.
- This process should feel very Bayesian!
- Although measurement error modelling doesn't have to be strictly Bayesian, it is very naturally placed inside the general methodology.

Missing data modelling

- Moreover, the Bayesian approach to measurement error modelling yields many important ramifications.
- In particular, the problem of *missing data* can be viewed as a measurement error problem.
- Three main types of missing data (Rubin):
  (1) Missing completely at random (MCAR): some X observations are missing randomly.
  (2) Missing at random (MAR): some X observations are missing randomly *conditional* upon other observed variables.
  (3) Missing not at random (MNAR): missingness in X is informed by other, unknown variables (confounders).

Missing data modelling

- MCAR *never* occurs in practice.
- **All** techniques (e.g. imputation, FIML) for handling missing data assume an MAR structure.
- Using the measurement error framework from before:

$$f(y, x^*) = \int f(x^* \mid x) \cdot f(y \mid x) \cdot f(x) \; dx$$

  if x is *missing*, then can use only x^* and y to bypass it.
- As before, this requires a *model* relating the missing observation x with the complete observations x^*, as well as a *model* relating the missing observation x with the other variables of interest (assuming a MAR structure).
- The most critical thing to remember about ALL missing data techniques: they are all *model-based*. That is, your answers are dependent on the quality of the proposed (missing data) model.

Priors

Four main types of priors:

- Informative priors: contains explicit, empirical information about a parameter.
- Weakly informative priors: contains explicit, theoretical information about a parameter.
- Uninformative (diffuse) priors: attempts to be *indifferent* about prior believability.
- Improper priors: special kinds of improper priors.

Asymptotically, the posterior will be *unaffected* by all these priors, assuming that the true parameter value is not *a priori* excluded from their domain.

Priors

Informative priors:

- Can be *subjective* or *objective*:
  - Subjective prior: informed by expert opinion
  - Objective prior: informed by previous empirical research
- Encodes nontrivial and not-only-theoretical information about a model parameter; e.g. tomorrow's temperature is a normal distribution with mean equal to today's temperature and standard deviation such that 95% of historical temperature data falls within the 95% probability interval of this mean.

Priors

Weakly informative priors:

- Usually objective priors.
- Encodes theoretical information about a parameter; e.g. tomorrow's temperature follows a normal distribution with mean equal to today's temperature and standard deviation such that the 95% probability interval around the mean falls between -40 and +40 Celsius.
- A variety of professional statisticians recommend using default weakly informative priors when good informative priors are not available (e.g. A. Gelman, G. Carlin).

Priors

Uninformative priors:

- Eschew all empirical and theoretical knowledge. Attempt platonicity.
- Also called *not very informative* or *objective* priors.
- Note: I have used "objective" prior to mean something else.
- Lots of mathematics around what consitutes an uninformative prior.
- Simplest example: probability of infection of disease is Uniform[0,1].
- Major dilemma: *is this really objective?* Saying all proportions are equally plausible is still a substantive judgment.
- Problem: what is an "objective" uninformative prior for the amount of rainfall tomorrow?

Priors

Improper priors:

- To solve the previous problem of specifying uninformative priors over infinite domains, one may use improper priors.
- Priors are called "improper" if they do *not* integrate to 1.
  - Recall: every non-negative function must integrate to 1 if it is to encode probabilities (i.e. be a PDF).
  - If a non-negative function does *not* integrate to 1, then it is like a *likelihood*; i.e. it can encode *relative* probabilities/likelihoods, but cannot encode absolute probabilities.
  - Improper priors encode only these relative probabilities; still mathematically sufficient for many Bayesian operations.