Bayesian Data Analysis

2020

Week 6: Generalized linear model, model assessment, and Criticism

Jarno.Vanhatalo@helsinki.fi

The previous lecture

- Hierarchical models
- Exchangeability

Aims of the week

- Generalized linear models
    - Exercise 6.1 Generalized linear model with Binomial observation model
- Model criticism, assessment and comparison
    - Exercise 6.2

Generalized linear models

- Extension of the linear modeling process that allows models to be fit to data that follow probability distributions other than the Normal distribution (such as the Poisson, Binomial, Multinomial etc.).
- Can be used in cases for which the linear relationship between x and E(yIx) is not appropriate.
- For each variate yi, i=1,…,n, with a corresponding set of k explanatory variables xij, there exist a (monotone differentiable) function g, called the link function, such that

- g(E[yiIxi])=β'xi,

- where xi=[xi1,…,xik]' and β'=[β1,…,βk] is a parameter vector.

Generalized linear models

- Specifications in 3 steps:
  - The linear predictor η= β'X, where X=[x1,…,xn] is k×n matrix of explanatory variables and η is a vector of n linear predictor values.
  - The link function g(·) that relates the linear predictor to the mean of the outcome variable: μ=g-1(η)=g-1(β'X)
  - The random component specifying the distribution of the outcome variable y with mean E(yIx)=μ=g-1(β'X)
- Data distribution
- $p(y|x,\beta) = \prod_{i=1}^{n} p(y_i|x,\beta)$
- Interpretation of the model parameters not so straightforward. The linear predictor is used to predict link function g(μ) rather than μ=E(y).

Generalized linear models

- Link functions for different data type:
  - Normal data with mean μ: g(μ)=μ
  - Binomial data with probability p:
  - $g(p) = logit(p) = log\left(\frac{p}{1-p}\right)$
  - also probit link function used in econometrics
  - Poisson data with rate λ: g(λ)=log(λ) (see lecture example)

Logistic link function

- Consider a Binomial model with success probability π
- Logistic regression assumes that the log odds is a function of covariates, e.g.,

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = c + ax_i$$

- Then

$$\pi_i = \frac{1}{1+e^{-(c+ax_i)}}$$

Model assessment

- Checking for problems in the model
  - Are the results sensible?
  - In what respect the model works / does not work
  - Have we done mistakes
- All models are approximations of the reality and not all aspects of the phenomena are included into the model
  - Are the results sensible when compared to information not encoded in the model
- There is no universal method to rule out problems with model
  - Compare to convergence analysis

Model assessment

- Typical uses of a model
  - Explain data
  - Predict
    - Interpolation
    - extrapolation
  - Hypothesis testing

Posterior predictive check

- Internal validation
- Predict replicate data $\tilde{y}$
  - Compare replicates $\tilde{y}$ to observations
  - Good model should be at least consistent with data it is conditioned to
- Pragmatic method to check for
  - The worst problems with model
  - systematic deviations between model and data
- Not a formal method but useful;
- The main problem is that it uses data twice
  - May overfit the data
- <u>Read Chapter 6 from the BDA3!</u>

Sensitivity analysis

- Check how sensitive the results are to model assumptions
  - Prior
  - Likelihood
  - Hierarchical vs. pooled vs. independent
- Sensitivity is a neutral term
  - If your inference is sensitive to aspects of model that you are confident about, you have found a ”real thing”
  - You should be concerned if your inference is sensitive to aspects of model that you are not able to justify well

Predictive assessment

- Interpolation
  - how well model works when predicting ”in the vicinity and between” data points
    - Dose – response exercise
    - Regression exercise
- Extrapolation
  - how well model works when predicting far from data
    - New laboratory in the rat tumor exercise
    - Climate in the next century
- When predicting we rely on assumption that the model works similarly both where
  - we have not been yet
  - we have data from
  - -> ”data generating process does not change”
    ! Remember discussion on exchangeability !

External validation

- Compare model’s predictions to new / external observations
    - Generally used method in science
    - Can be planned to test interpolation and extrapolation
    - Predict something that has not been measured / observed before
        - e.g. Einsteins theory of relativity or Higgs boson

Partial validation

- Training and test set
  - Divide data into training and validation sets
  - Train the model using training set
  - Predict validation set and compare predictions to observations
- Pros / cons

  \+ easy

  \+ rather safe
  - Assessment with similar data as used for training (interpolation)
    - In some cases data can be divided to test extrapolation
  - Sensitive to how data is divided
    - If division introduces structure the validation may give false confirmation / doubts
  - Training is conditional only on subset of data

Cross validation

- Cross validation sets
  - Divide data into k sets
  - Train the model k times using k-1 sets in training each time
  - Each time, predict for the left out set and compare predictions to observations
  - Do the comparison for each of the k sets and calculate average over them
- Pros / cons

  + rather easy if k is small

  + rather safe

  +/- not so sensitive to how data is divided
    - the larger k is the less sensitive to how data is divided
  - Assessment with similar data as used for training (interpolation)
    - In some cases data can be divided to test extrapolation

Bayesian model averaging / hypothesis testing

- Sometimes we may have competing models / hypotheses
  - $M_1, M_2, \ldots, M_m$
- The posterior distribution of a model

$$p(M_1|y) = \frac{p(y|M_1)p(M_1)}{p(y)}$$

Bayesian model averaging (BMA)

$$p(M_1|y) = \frac{p(y|M_1)p(M_1)}{p(y)}$$

- The probabilities $p(M_1|y)$, $p(M_2|y)$, … tell the relative credibility of each of the models considered
  - Model's prior : $p(M_1)$
  - Data: y

Bayesian model averaging (BMA)

$$p(M_1|y) = \frac{p(y|M_1)p(M_1)}{p(y)}$$

- Probabilities are relative within the model set
  - This is not the ”absolute” probability of the model
- BMA prediction averages over the alternative models

$$p(\tilde{y}|y) = p(\tilde{y}|y, M_1)p(M_1|y) + p(\tilde{y}|y, M_2)p(M_2|y) + \ldots$$

  - (usually) better than to choose one individual model

Bayes Factor

$$p(M_1|y) = \frac{p(y|M_1)p(M_1)}{p(y)}$$

- Bayes Factor: e.g.

$$\mathrm{BF} = \frac{p(y|M_1)}{p(y|M_2)}$$

  - The posterior odds of models when uniform prior
- Things to remember
  - practical computation is often hard
  - $p(y|M_1)$ is sensitive to priors of parameters
    - This is model's prior predictive density
    - Priors are part of the hypothesis
      -> remember BF is about hypothesis testing

Model comparison and choice

- Philosophically strict approach
  - The model is your best subjective description of the phenomenon
    - BMA takes into account your uncertainty about different plausible hypotheses
  - Model assessment helps to
    - Locate ”implementation errors”
    - Reveal shortcomings in your hypotheses and, thus, need for new hypotheses (=model)
      - e.g. Earth is a ball, or finding new planets
      - Remember: your inference is conditional to your model(s)
      - Bayesian theory does not say where new models come from

Model comparison and choice

- ”statistical” approach
  - The model is not strict description of your prior understanding of the phenomenon but ”good enough” to
    - Explain data
    - Predict
- Sometimes there is need to choose between models
  - Simpler model may be easier to explain or cheaper to use
    - e.g. in prediction it needs less measured covariates
  - Choose a model whose predictive performance is the best
    - Averaging over models is usually even better
- Model choice is a decision problem

Predictive performance

- Model is used for predicting future observations
  - Dose – response model
  - Weather forecasting
  - Google search
- A set of alternative models from which to choose
  - Often reduces to covariate selection
- Prior predictive performance -> BMA
- Posterior predictive performance
  - How well does the model predict new data given the current data (and the model) ?

Model comparison measures

- What are you using the model for?
  - > what are the important aspects of the model?
- Commonly used (general) validation measures
  - Root mean squared error (RMSE)
  - Log predictive distribution (deviance)
  - Classification accuracy

Root mean squared error

- The question to answer:
    - How well, e.g., the posterior predictive mean $E[\tilde{y}_i|y, M_1]$ predicts the observations

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i - E[\tilde{y}_i|y, M_1])^2}$$

- The smaller the better
- Does not assess the sharpness (uncertainty estimate) of the prediction
- Compare to the estimate of the standard deviation of the data

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

Log predictive density

- The question: How well does our posterior predictive distribution, $p(\tilde{y}_i|y)$, represent the distribution of the unseen data $p(\tilde{y}_i)$?
  - Formally this could be analysed with expected log-predictive density

$$\text{LPD} = \int \log\big(p(\tilde{y}_i|y, M_1)\big)p(\tilde{y}_i)d\tilde{y}_i$$

    which gets its maximum when $p(\tilde{y}_i|y, M_1) = p(\tilde{y}_i)$.
    - Accounts for both location and width (uncertainty) of distribution
  - Since we don't know $p(\tilde{y}_i)$ we approximate the log predictive density with

$$\text{LPD} = \frac{1}{n}\sum_{i=1}^{n} \log\big(p(\tilde{y}_i|y, M_1)\big)$$

    where $\tilde{y}_i$ are observations in the test data.

Model comparison metrics

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i - E[\tilde{y}_i|y, M_1])^2}$$

$$\text{LPD} = \frac{1}{n}\sum_{i=1}^{n}\log\left(p(\tilde{y}_i|y, M_1)\right)$$

- Where do we get samples from the distribution of future data, $\tilde{y}_i$? For example:
  - Partial validation ($\tilde{y}_i$ from test set)
  - Cross validation ($\tilde{y}_i$ from test sets)

Information Criterion

- There are many information criteria introduced in the literature
  - Akaike Information Criterion (AIC)
  - Deviance Information Criterion (DIC)
  - Bayesian Information Criterion (BIC)
- The aim is to compare models' predictive performance (crudely) with

$$(model\ complexity) - (model\ fit)$$

  - The model with minimum ”criterion” is the best
- All information criteria are more or less ad hoc measures

Information Criterion

- The name information criterion originates from ”model fit” term being related to distribution’s entropy
  - Entropy is a measure of information
- Deviance Information Criterion (DIC)
  - Widely used in Bayesian statistics
  - Approximates (under certain assumptions) the log predictive density
  - Easy to calculate and ready made functionalities in JAGS
  - Problems how to interpret and implement with hierarchical models
- I recommend posterior predictive check with cross validation

Model assessment and sensitivity analysis revised

- The purpose is to check for
  - possible misspecifications
  - To which parts of the model the inference is sensitive to
- Error control procedure
  - The model is at best good approximation of the reality
  - You may have made logical or implementation errors

Model updating

- If model assessment or sensitivity analysis show aspects of concern
  - Check your code
  - Check/revise your priors
  - (Try robust observation models)
  - Revise your model assumptions concerning independe, hierarchy, …
  - … is your theory correct

How to construct the model in practice?

- Depends on the problem – there is no rule-of-thump
- Prior information about the phenomenon essential
- Think conditionally and build model gradually via conditional distributions!
  - What are you interested in?
    - e.g. the proportion of black balls θ
  - What are the things your variable of interest is related to and how?
    - Need for more model layers?
  - How are the observations related to the variable of interest
    - the observation model $p(y|\theta)$
  - What prior information do you have on the variable of interest
    - The prior $p(\theta)$
    - Literature, expert knowledge, earlier experiments…
- With complex models division between prior and observation model is not straightforward
  - Conditional thinking helps!

Many sources of error

- ”All models are wrong but some are usefull” (George P. Box)
  - Process/observation model is an approximation
  - Prior distributions are approximations of our knowledge
- Implementation error. We may
  - do the math wrong
  - code the model wrong
  - read the data wrong
  - read results wrong
  - calculate Monte Carlo estimates wrong
  - …

Ways to mitigate errors

- Do every task carefully
- Record your model building and implementation
  - Notes on how model was derived theoretically
  - Comments within code
  - Intuitive names for variables
  - Clean code
- Reserve time for coding
  - Rush is the single largest cause of errors
  - Analysis of data may be as time consuming as the laboratory experiments
- Build your model gradually
  - Start with simple assumptions, pen and paper
  - Build first as simple (small) model as possible
  - Extend your model step by step

Ways to mitigate errors

- Make a personal repository of reliable code
  - Collect functions and scripts that you use frequently
  - Put in repository only code that is well double checked
- Check you results by approximating it with alternative means. The easiest are
  - ”intuitively”
    - does the result make sense
    - Is its implication sensible
  - crudely
    - Are the parameter values in right order of magnitude
  - …
- Posterior predictive check (later)
- Double check - double check – double check …

On prior distributions

- Prior represents the (subjective) state of knowledge
- Setting up a prior is
  - easy if the uncertainty in the knowledge is small (informative prior)
  - hard if the knowledge is uncertain (non-informative prior)
- Prior should always cover all possible parameter values
  - if prior is 0 also posterior is 0
  - if we have lot of data likelihood usually dominates in the posterior
  - if we have small amount of data prior may influence a lot the posterior

On prior distributions

- Informative priors
  - Priors that clearly state modelers beliefs
  - collects all proper distributions (Gaussian, Gamma, Beta,...)
- non-informative priors
  - Priors that try to be non-informative and code analyst's ignorance on the prior information
  - <u>only spuriously non-informative and may be heavily informative in a reparameterized model (so be careful)</u>
  - typical example is a uniform prior
  - usually improper distributions
- Weakly informative priors
  - proper distributions that are robust for prior misspecification
  - e.g. Student-t distribution and mixture of Gaussians
- Conjugate priors
  - contains both informative and non-informative priors
  - specific functional form that depends on the likelihood
    - e.g. Beta distribution for success rate of Binomial
  - Makes practical computation easier
- ...

Next week

- Revision of the course content