MIDTERM REVIEW

Lec 1:

Introduction to modeling, MLE and MAP

Example of Covid 19 test

Sensitivity – true positive rate

Specificity – true negative

Confidence vs. credible intervals

Beta-binomial model – derived posterior distribution for the rate parameter

Lec 2

Showed example of a model that estimates difference in rates using beta-binomial model

Linear Gaussian model

- We introduced a formula sheet for Gaussian Distribution
- we derived p(mu|x) when the variance is known: $x_i = \mu_i + \epsilon_i$
- Example of estimating the download speed of internet
- Showed Gamma->Normal->Gamma model (Lec2 Explanations)

Bayesian linear regression – showed model where the likelihood and the prior, computed posterior and posterior predictive distribution – presenting lambda as a regularizer

Performance metrics for regression: r^2, sum of square residuals, explained sum of squared, Non-linear regression was just mentioned: Page 12 Lec 2

Lec 3

Logistic regression – introduced sigmoid function, showed that it becomes softmax when we have more than 2 clases, showed the formula for the gradient of the cost finction

Sampling: Rejection sampling, Importance sampling

Markov chain: transfer function, detailed balance, 2 examples – one discrete Markov chain and whan with continuous with time series (Lec 3 page 5)

MCMC:

- Algorithms: Metropolis Hastings – proposal distribution q, accept-reject

- Gibbs sampling: sampling in a cyclic order – needs to come up with all conditional distributions
- We said that these probabilities are easier to derive if we look at the probabilistic graphical model (Lec 3, page 3 of the second handout)
- Burn-in, thinning, autocorrelation function, effective sampling size, Gelman rubin statistics

Stochastic gradient descent:

https://www.cse.iitk.ac.in/users/piyush/courses/ml_autumn18/material/771_A18_lec8_print.pdf

Langevin dynamics: combined MCMC and stochastic gradient

- Hamiltonian MC and NUTS are just mentioned as combinations of MCMC and stochastic gradient

Lec 4

Variational inference:

- Kullback–Leibler divergence (minimize), Evidence lower bound (maximize), mean-field variational Bayes
- Estimating parameters of q that will maximize the ELGO for one-dimensional Gaussian example
- Combining variational inverence with gradient descent: Computing gradients of expectations
- Black box variational inference

Mixture models: latent variables, Gaussian mixture model (unsupervised learning)

Hierarchical models; pooling, shrinking

- Hierarchical binomial model, Seven school model (Normal)

Lec 5

Model checking: posterior predictive check (page 8)

- Example: speed of light measurements, beta-binomial model – is data IID

Model selection: log likelihood+penalty: AIC, DIC, WAIC, Bayes factors

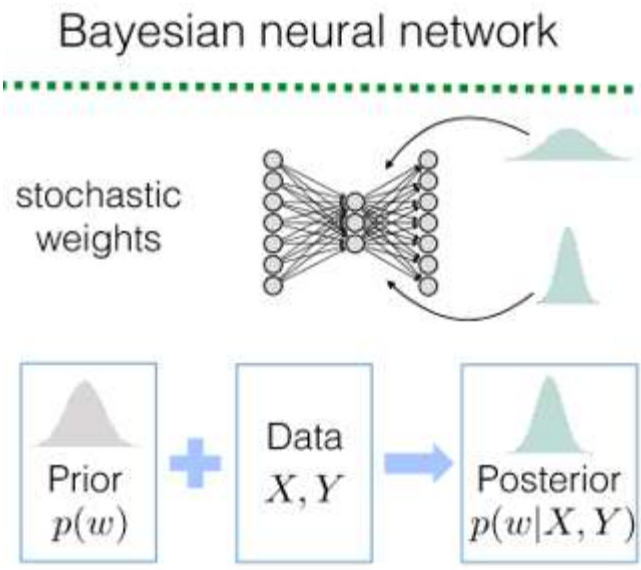
- Leave one out
- Bayesian averaging
- Derived AIC for linear regression, showed how to compute WAIC

Introducing errors in both x and y variables

Bayesian neural networks

QUESTIONS

1. Measurements of the number of units per package are done in 5 factories. All factories produce the same kind of units. We are interested in modeling this system and then answering some questions related to uncertainty.
    a. Perform preliminary data analysis by drawing the boxplot per location. Obtain mean and variance of units per location and total mean and variance.
    b. Define non-hieararchical model in which you would ignore the location information and chose appropriate values for priors. Implement it in PyMC3.
    c. Define hierarchical model and chose appropriate values for priors. Implement it in PYMC3. You can use Poisson distribution for the likelihood. Poisson distribution has parameter lambda- you can use Gamma distribution as a prior for it. Parameter lambda can have different values per location.
    d. Perform Gelman-Rubin and Autocorrelation check for both models
    e. Draw residuals for hierarchical model (both per location as well as for all locations – total 6 graphs) and see if they look uncorrelated and normal.
    f. Compute Waic for both hierarchical and non-hierarchical models and compare them.
    g. Perform posterior predictive simulation and then answer the following questions: What is the probability that the number of units per package in any location to be greater than 15? What is the probability that the next package produced in location 1 will be less than 7.
2. Show that the coeficient Beta_1 in linear regresion decreases is case measurement noise is not taken into account
3. Bayesian neural network is shown below. How could one compute the prediction y* given x*?



4. Run Metropolis Hastings sampling algorithm of size 5000 to sample from the following posterior distribution:

$$p(\theta) = 0.6 * e^{-\theta^2/2} + 0.4 * (\frac{1}{2} e^{-\frac{1}{2*2^2}(\theta-3)^2})$$

Analyze the following:

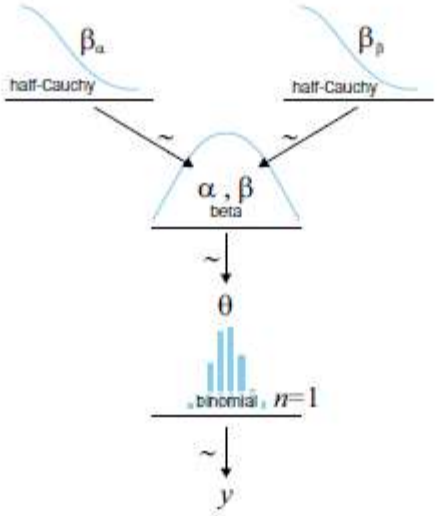
1. Autocorellation plot
2. Gelman-rubin
3. HPD
4. Effective size

5. Water samples are tested at 3 different locations 30 times. We are provided with the number of samples that are OK.

The model is essentially the same one we use for the coin problem, except that now we have to specify the hyper-priors that will influence the beta-prior:

$$\alpha \sim HalfCauchy(\beta_\alpha)$$
$$\beta \sim HalfCauchy(\beta_\beta)$$
$$\theta \sim Beta(\alpha, \beta)$$
$$y \sim Bern(\theta)$$

Using Kruschke diagrams, it is evident that this new model has one additional level compared to all previous models:



The number of samples and the estimated means in the hierarchical model are given below. Explain results.

G_samples	Theta (mean)
18, 18, 18	*0.6, 0.6, 0.6*
3, 3, 3	*0.1, 0.1, 0.1*
18, 3, 3	*0.53, 0.14, 0.14*

6.

Assume that we have some parameter $w \sim \mathcal{N}(w_0, \sigma^2 \mathbf{1})$ drawn from a d-dimensional Normal distribution. Moreover, assume that we perform regression with additive Normal noise ϵ. That is, assume that we have

$$y = \langle w, \phi(x) \rangle + \xi \text{ where } \xi \sim \mathcal{N}(0, \tau^2) \text{ and } w \sim \mathcal{N}(w_0, \sigma^2 \mathbf{1}) \text{ and } \phi(x) \in \mathbb{R}^d.$$

Assume that we observe m pairs (x_i, y_i). Show that the posterior distribution is normal

$$p\left(w | (x_1, y_1), \ldots (x_m, y_m), \tau^2, \sigma^2\right)$$