STATS 230: Computational Statistics
Some Preliminary Concepts

Babak Shahbaba

Department of Statistics, UCI

Inference in the frequentist framework

- We typically start statistical inference by defining the underlying mechanism that generates data, y, using a probability model, $P(y|\theta)$, which depends on the unknown parameter of interest, θ.
- The likelihood function is defined by plugging-in the observed data in the probability distribution and expressing it as a function of model parameters: $f(\theta; y)$.
- To estimate model parameters, we can find their values such that the probability of the observed data is maximum.
- For this, we maximize the likelihood function with respect to model parameters.
- Of course, it is easier to maximize the log of likelihood function, i.e., $L(\theta) = log(f(\theta))$.

Score function and information

- For single parameter exponential family,

$$L(\theta) \quad = \quad g(\theta)s(y) - c(\theta)$$

- The first derivative of log-likelihood function, $L(\theta)$, is called the *score function*

$$u(\theta) \quad = \quad \frac{\partial L(\theta)}{\partial \theta}$$

- For single parameter exponential family,

$$u(\theta) \quad = \quad s(y)\frac{\partial g(\theta)}{\partial \theta} - \frac{\partial c(\theta)}{\partial \theta}$$

- To find MLE, we set

$$u(\theta) \quad = \quad 0$$

Maximum likelihood estimation

- Under weak regularity conditions, the MLE demonstrates attractive properties as $n \to \infty$: the asymptotic distribution of MLE is normal, MLE is asymptotically consistent and efficient.

- Under some regularity conditions (Rao, 1973), the asymptotic covariance matrix for MLE, $Cov(\hat{\theta})$, is the inverse of *Fisher information matrix*, $i(\theta)$, where the (j, k) element of $i(\theta)$ is

$$Cov[\frac{\partial L(\theta)}{\partial \theta_j}, \frac{\partial L(\theta)}{\partial \theta_k}]$$

  which is equal to the following (assuming that we can take differentiate twice inside integral)

$$E\Big(- \frac{\partial^2 L(\theta)}{\partial \theta_j \partial \theta_k}\Big)$$

Bayesian inference

- In Bayesian statistics, besides specifying a model $P(y|\theta)$ for the observed data, we specify our prior $P(\theta)$ for the model parameters.
- Then, we make probabilistic conclusions regarding the unobserved quantity θ conditional on the observed data y.
- That is, we are interested in $P(\theta|y)$, which is called *posterior distribution*.
- Bayes' theorem provides a mathematical formula for obtaining $P(\theta|y)$ based on $P(\theta)$ and $P(y|\theta)$:

$$\begin{aligned} P(\theta|y) &= \frac{P(\theta)P(y|\theta)}{P(y)} \\ &\propto P(\theta)P(y|\theta) \end{aligned}$$

Bayesian inference

- Inference in the Bayesian framework is based on $P(\theta|y)$.
- For example, we can predict future observations, $\tilde{y}$, given the observed data y:

$$P(\tilde{y}|y) = \int_{\theta} P(\tilde{y}|\theta)P(\theta|y)d\theta$$

- This is a simple concept; however, finding $P(\theta|y)$ in practice is challenging.

Conjugate priors

- In some cases, we can limit our choice of prior to a specific class of distributions such that the posterior distribution has a closed form.
- This is called "conjugacy" and the prior is called a "conjugate" prior.
- Conjugacy is informally defined as a situation where the prior distribution $P(\theta)$ and the corresponding posterior distribution, $P(\theta|y)$ belong to the same distributional family.
- Using conjugate priors makes Bayesian inference easier.

Conjugate priors

- Recall that the exponential family has the following form:

$$P(y|\theta) \propto \exp\{g(\theta)s(y) - nc(\theta)\}$$

- Now if we define the prior as follows:

$$P(\theta) \propto \exp\{g(\theta)\nu - \eta c(\theta)\}$$

- Then the posterior would have a similar form:

$$P(\theta|y) \propto \exp\{g(\theta)(\nu + s(y)) - (n + \eta)c(\theta)\}$$

Poisson model

- Poisson model is another member of exponential family and is commonly used for count data.
- Assume we have observed $y = (y_1, y_2, ..., y_n)$:

$$P(y|\theta) \propto \exp(\log(\theta)\sum y_i - n\theta)$$

- The conjugate prior would have the following form:

$$\begin{aligned} P(\theta) &\propto \exp(\log(\theta)\nu - \eta\theta) \\ &\propto \exp(-\eta\theta)\theta^{\nu} \end{aligned}$$

- Using $P(\theta) \propto \exp(-\beta\theta)\theta^{\alpha-1}$, which is a Gamma$(\alpha, \beta)$ distribution,

$$\theta|y \sim \text{Gamma}(\alpha + \sum_{i=1}^{n} y_i, \beta + n)$$

Computation in statistics

- In general, finding MLE and posterior distribution analytically is difficult.
- Therefore, we almost always rely on computational methods.
- In this course, we will discuss a variety of computational techniques for numerical optimization and integration
- Optimization methods are mainly discussed within the frequentist framework.
- Numerical integration methods are mainly discussed with respect to their application in Bayesian inference.
- We will also discuss a variety of other computational methods (e.g., numerical linear algebra, bootstrap) that are commonly used in modern statistics.

What's next?

- Please read Chapter 1 from "Computational Statistics," by Givens and Hoeting and Chapter II.1 from "Handbook of Computational Statistics," by Gentle et al. to review some basic concepts on computer arithmetic and algorithms as well as background materials in statistics and probability.
- We will start our lectures by numerical linear algebra.
- We then spend several lectures on optimization methods.
- Next, we will discuss a variety of methods for numerical integration and approximation.
- The last several lectures are devoted to some additional topics including bootstrap, EM, and advanced sampling algorithms.