STATS 235: Modern Data Analysis
Learning

Babak Shahbaba

Department of Statistics, UCI

Outline

- In this lecture, we discuss several learning methods.
- More specifically, we discuss fitting models within frequentist, Bayesian, and information theory frameworks.
- Throughout this lecture, we will use exponential family as examples.

Frequentist Framework

Likelihood function

- 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,

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

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

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

- For single parameter exponential family,

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

- To find MLE, we use

$$u(\hat{\theta}) = 0$$

Score function and information

- Under some regularity conditions (mainly to make it possible to interchange integration and differentiation), for a given value of θ we have

$$E_\theta[u(\theta)] \quad = \quad 0$$

- As the result

$$var_\theta[u(\theta)] \quad = \quad E[u^2(\theta)] = i(\theta)$$

- $i(\theta)$ is called *Fisher information* about θ given y.

- Under the regularity conditions assumed above,

$$i(\theta) \quad = \quad E[u^2(\theta)] = E[-\frac{\partial^2 L(\theta)}{\partial \theta^2}]$$

Example: Poisson model

- Consider the following Poisson model:

$$
\begin{aligned}
P(y_i|\mu) &= e^{-\mu_i}\mu_i^{y_i}/y_i! \\
&= \exp\{\log(\mu_i)y_i - \mu_i - \log(y_i!)\}
\end{aligned}
$$

  where $\phi_i = g(\mu_i) = \log(\mu_i)$, $S(y_i) = y_i$, $c(\mu_i) = -\mu_i$, and $h(y_i) = -\log(y_i!)$.

- The score function with respect to ϕ_i can be obtained as follows:

$$
\begin{aligned}
u(\phi_i) &= \frac{\partial L(\phi_i)}{\partial \phi_i} \\
&= S(y_i) + \frac{\partial c^*(\phi_i)}{\partial \phi_i} \\
&= y_i - \exp(\phi_i) \\
&= y_i - \mu_i
\end{aligned}
$$

Example: Poisson model

- The total score function based on n observations is

$$u(\phi) = \sum_i y_i - \exp(\phi_i) = \sum_i y_i - \mu_i$$

- As the result, the likelihood equation is:

$$\sum_i y_i - \exp(\hat{\phi}_i) = \sum_i y_i - \hat{\mu}_i = 0$$

- For exponential family models in general, the parameter estimates are obtained by equating sufficient statistics to their expectations.

Example: Poisson model

- In regression analysis, we use the Poisson model when the outcome variable, y, represents counts:

$$y_i|\mu_i \sim \mathrm{Poisson}(\mu_i)$$

- The systematic components are defined as before: $\eta_i = x_i\beta$.

- The usual link function for this model is the log link:

$$g(\mu_i) = \log(\mu_i) = \eta_i$$

- We therefore have

$$\mu_i = \exp(\eta_i) = \exp(x_i\beta)$$

Example: Poisson model

- The likelihood in terms of β can obtained as follows:

$$\begin{aligned} p(y_i|\mu_i) &\propto \prod_i^n \exp(-\mu_i)\mu_i^{y_i} \\ p(y_i|\beta) &\propto \prod_i^n \exp[-\exp(x_i\beta)][\exp(x_i\beta)]^{y_i} \end{aligned}$$

- The variance of $y|x$ in Poisson model depends on the mean and therefore will not be constant

$$var(y_i|x_i) = \mu_i$$

Example: Poisson model

- For Poisson regression model, we are of course interested in regression parameters β.
- Therefore, we would like to write the score function in terms of β.
- To do this, we fist need to specify the link function.
- Suppose we use the log link function

$$g(\mu_i) = \log(\mu_i) = x_i\beta$$

- Since we have $\phi_i = g(\mu_i)$, we can write the link function as follows:

$$\phi_i = \log(\mu_i) = x_i\beta$$

- The link function that transforms the mean to the natural parameter is referred to as the *canonical link*.
- For Poisson model, the log link is the canonical link.

Example: Poisson model

- Using the link function, we can now write the score function in terms of β.

- For the j^{th} element of β, we have

$$\begin{aligned} u(\beta_j) &= \sum_i \frac{\partial L(\beta)}{\partial \beta_j} \\ &= \sum_i \frac{\partial L(\phi)}{\partial \phi_i} \frac{\partial \phi_i}{\partial \beta_j} \\ &= \sum_i [y_i - \exp(x_i\beta)]x_{ij} \end{aligned}$$

- As the result, the likelihood equation in terms of β_j is

$$\sum_i [y_i - \exp(x_i\hat{\beta})]x_{ij} = 0$$

Example: Poisson model

- We can now easily obtain the Fisher information matrix in terms of β.

$$
\begin{aligned}
i(\beta_j \beta k) &= E[-\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_k}] \\
&= E[\sum_i x_{ij} x_{ik} \exp(x_i \beta)] \\
&= \sum_i x_{ij} x_{ik} \exp(x_i \beta)
\end{aligned}
$$

- In a matrix format

$$
i(\beta) = x' w x
$$

where w is a diagonal matrix whose i^{th} element is $\exp(x_i \beta)$.

- Moreover,

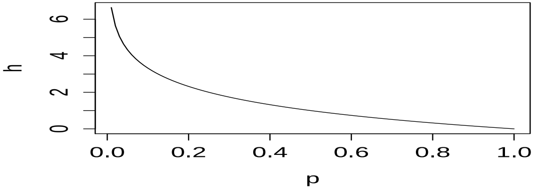
$$
cov(\hat{\beta}) = (x' \hat{w} x)^{-1}
$$

Information Theory

Information content

- Information theory deals with communication problems.
- Shannon: Fundamental problem in information theory is reliable communication over unreliable channels.
- Information content (in bits) of an outcome x is

$$h(X = x) = \log_2 \frac{1}{P(X = x)}$$



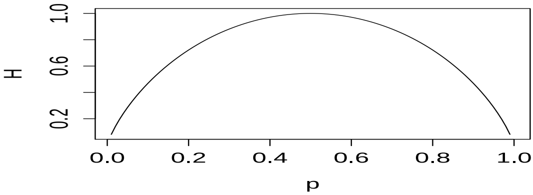
Entropy

- Shannon information content is highest for outcomes with low probability.

- For a set of outcomes, entropy is defined as the average Shannon info:

$$H(X) = \sum_{x} p(x) \log_2 \frac{1}{P(x)}$$

- Suppose there are only two possible outcomes with probabilities p and $1 - p$,



Entropy maximization

- We choose the probability model by maximizing entropy (maxent).
- This would be the “fairest” thing to do: entropy is maximized when p is uniform.
- Additionally, this is based on the philosophy that we should use the least number of assumptions for inference.
- Suppose we want to find the probability distribution $P(x)$ by maximizing entropy subject to K constraints of the following form:

$$\sum_x f_k(x)P(x) = F_k, \qquad k = 1, \ldots, K$$

Entropy maximization

- For example, we might match the moments under the probability model with the moments under the empirical distribution, i.e.,

$$\sum_x xP(x) = \overline{x}$$

- Additionally,

$$\sum P(x) = 1$$

- To maximize entropy subject to above constraints, we define the Lagrangian as follows:

$$\begin{aligned} L(P,\lambda) &= -\sum_x P(x)\log P(x) + \lambda_0(1 - \sum_x P(x)) \\ &\quad + \sum_k \lambda_k(F_k - \sum_x f_k(x)P(x)) \end{aligned}$$

Entropy maximization

- We treat $P(x)$ is a vector of fixed length (as opposed to a function),

$$\frac{\partial L(P,\lambda)}{\partial P(x)} \;=\; -1-\log P(x)-\lambda_0-\sum_k \lambda_k f_k(x)=0$$

- Therefore,

$$P(x)=\frac{1}{z}\exp(-\sum_k \lambda_k f_k(x)), \qquad \text{where } z=\exp(1+\lambda_0)$$

  which has the form of the exponential family.

- Therefore, maxent with the least number of assumptions in terms of moments leads to exponential family models.

Relative entropy

- The relative entropy between two probability distributions $P(x)$ and $Q(x)$ is defined as

$$D_{KL}(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

  which satisfies Gibbs' inequality

$$D_{KL}(P||Q) \geq 0$$

- Note that this is not symmetric in general so $D_{KL}(P||Q)$ is not the same as

$$D_{KL}(Q||P) = \sum_x Q(x) \log \frac{Q(x)}{P(x)}$$

- This also know as Kullback-Leibler divergence

Bayesian Framework

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*.

Bayesian inference

- Bayes' theorem provides a mathematical formula for obtaining $P(\theta|y)$ based on $P(\theta)$ and $P(y|\theta)$:

$$P(\theta|y) = \frac{P(\theta)P(y|\theta)}{P(y)}$$

- Since $P(y)$ does not depend on θ, we can use the following unnormalized form of the posterior distribution:

$$P(\theta|y) \propto P(\theta)P(y|\theta)$$

- This is a simple concept; however, finding $P(\theta|y)$ 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 (given n observations):

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

- 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 $y = (y_1, y_2, ..., y_n)$:

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

- 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,

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

MAP

- When the posterior distribution does not have a closed form, we can use sampling algorithms to obtain the posterior distribution.
- Alternatively, we could simply estimate model parameters by maximizing the posterior distribution.
- The resulting posterior mode, $\hat{\theta}$, is called the maximum a posteriori (MAP).
- This is not recommended in general since it does not capture our uncertainty regarding the estimates.

Decision Theory

Decision theory

- In the Bayesian paradigm, hypothesis testing and model evaluation are special cases of decision problems. In fact many topics such as point estimation and prediction are also discussed in the context of decision theory.
- Decision theory provides a mathematical framework for making decision under uncertainty; that is, when the outcome of an event is not known. We do, however, know what our loss (or gain) would be if any of the possible outcomes occur.
- Decision making is easy in theory, but it may be difficult in practice.

Decision theory

- We use $\mathcal{V}$ to denote the set of all possible values, v, for unknown variables. We refer to $\mathcal{V}$ as the *outcome space*.
- v could be the value of future observations. For example, $\mathcal{V} = \{Head, Tail\}$ when you are tossing a coin.
- Or, it could be the value of a parameter in a model. For example $\mathcal{V} = \mathcal{R}$, when we want to estimate μ, the mean of a normal distribution.

Decision theory

- We present the set of all possible actions, a, as $\mathcal{A}$. We refer to $\mathcal{A}$ as the *action space*.
- If we are predicting the outcome of the next coin toss, $\mathcal{A} = \{Head, Tail\}$.
- If we want to estimate μ (i.e., *point estimation*), our action space would be $\mathcal{A} = \mathcal{R}$.
- For hypothesis testing, we can define our action $\mathcal{A} = \{0, 1\}$, where 0 means do not reject the null hypothesis $H_0 : \mu \leq 0$ and 1 means rejecting it.

Decision theory

- We define *Utility* as a function $u = U(v, a)$ that maps the product of outcome space and action space to a real number $u \in \mathcal{R}$ representing how much we gain if we choose action a and the outcome v occurs.

- Pessimistic people might choose a loss function instead of utility (e.g., negative of utility) representing our loss when we choose action a and the outcome v occurs.

- In the coin tossing experiment, the loss function, $L(v, a)$ can be defined as follows:

  $$L(Head, Head) = L(Tail, Tail) = 0, L(Head, Tail) = L(Tail, Head) = 1$$

- This is known as $0 - 1$ loss function.

Decision rule

- Now, assume that we have observed data y, for example, $y = HHTHTHHT$, which is the outcome from a sequence of coin tossing. Using this data, we want to make a decision about what the outcome of the next toss would be (or what is θ, the probability of head for this coin).

- The tool for making decision is called *decision rule*, and it's denoted as $\delta(y)$. Note that δ is function of data (i.e., y) only.

- For example, we might define our decision rule for guessing what would be the outcome of the next toss as follows:

$$\delta(y) = \begin{cases} \textit{Head} & \text{if the observed fraction of Heads is} \geq 0.5 \\ \textit{Tail} & \text{if the observe fraction of Heads is} < 0.5 \end{cases}$$

Posterior risk

- Posterior risk for a decision rule is

$$r(\delta|y) = \int_{\mathcal{V}} L(v, \delta(y)) P(v|y) dv$$

- Note that we replaced the action a with the decision rule $\delta(y)$ since our action now depends on our decision rule which itself depends on the observed data.

- Also, note that $p(v|y)$ is the posterior predictive probability if v is future observation (i.e., what is the outcome of the next toss), or it is posterior probability if v is the parameter of a model (i.e., μ, the mean of a normal distribution).

Formal Bayes rule

- *The expected loss* principle: In deciding between different rules, choose the one with the smallest posterior risk. That is, take the action according to the rule with the smallest posterior expectation of loss function.

- The resulting rule is called a *formal Bayes rule*.

- Formal Bayes rule: $\delta_0(y)$ is a formal Bayes rule if $r(\delta_0|y) < \infty$ for all y and $r(\delta_0|y) \leq r(\delta|y)$ for all y and δ.

Formal Bayes rule

- In theory, this is all we need to know for all sorts of decision problems (e.g., prediction, point estimation, and hypothesis testing).
- For example, if we have a simple 0-1 loss function and a discrete action space such as the coin tossing example, the formal Bayes rule is choosing the mode of the posterior distribution $P(v|y)$.
- This is the reason we classify objects to the the class with a highest posterior probability when we use a multinomial logistic model.
- A different loss function may result in a different action.

Bayesian estimation

- Many decision problems in statistics deal with estimating the parameter of a probability model (e.g., the mean of a normal model, or the coefficients in a linear regression model), i.e. we have $\mathcal{V} = \theta$.

- A possible loss function is the *squared error loss*: $L(\theta, a) = (\theta - a)^2$.

- In general, the formal Bayes rule for this specific loss function is to choose the mean of the posterior distribution:

$$\begin{aligned} E_{\theta|y}(L(\theta, a)|y) &= E_{\theta|y}((\theta - a)^2|y) = E_{\theta|y}(\theta^2 - 2a\theta + a^2|y) \\ &= E_{\theta|y}(\theta^2|y) - 2aE_{\theta|y}(\theta|y) + a^2 \end{aligned}$$

We take the the derivative with respect to a and set it to zero:

$$-2E(\theta|y) + 2a = 0 \Rightarrow a = E(\theta|y)$$

- This is the reason behind using posterior expectation for point estimate.

Bayesian estimation

- Now suppose we want to use the *absolute error loss* function: $L(\theta, a) = |\theta - a|$
- Therefore, we need to minimize $E_{\theta|y}(|\theta - a|)$
- Using Leibniz's rule,

$$\frac{\partial}{\partial t}\int_{a(t)}^{b(t)} f(x,t)dx = \int_{a(t)}^{b(t)} \frac{\partial}{\partial t} f(x,t)dx - f(a(t),t)a'(t) + f(b(t),t)b'(t)$$

we have

$$\begin{aligned}\frac{\partial}{\partial a}E_{\theta|y}(|\theta - a|) &= \frac{\partial}{\partial a}\int_{-\infty}^{a}(a-\theta)f(\theta|y)d\theta + \frac{\partial}{\partial a}\int_{a}^{\infty}(\theta - a)f(\theta|y)d\theta \\ &= \int_{-\infty}^{a} f(\theta|y)d\theta - \int_{a}^{\infty} f(\theta|y)d\theta\end{aligned}$$

- This is zero when a is set to the median of the posterior distribution.