Prediction

STA 215. Surya Tokdar

Consider data X modeled as $X \sim f(x|\theta)$, $\theta \in \Theta$. Suppose we want to predict an unobserved quantity X^*, which depends on the same parameter θ, based on an observation $X = x$.

Example (Hurricane counts)**.** Based on count data $X = (X_1, \cdots, X_n)$ from n consecutive years, we might be interested in forecasting the number of TCs X_{n+1} in the coming year. Here $X^* = X_{n+1}$ and a reasonable model is $X_1, \cdots, X_n \overset{\text{IID}}{\sim} Poi(\mu)$, $X^* = X_{n+1} \sim Poi(\mu)$ and X and X^* are independent, where $\mu \in (0, \infty)$ is an unknown model parameter.

Example (Hurricant counts (contd.))**.** In the same setting, we might be interested in whether the next year's count exceeds a certain cut-off mark, say 15. In this case the variable of interest is the binary variable X^*, with $X^* = 1$ when $X_{n+1} > 15$ and $X^* = 0$ when $X_{n+1} \leq 15$, where X_{n+1} is the count for the coming year. Borrowing from the description of X and X_{n+1} above, we can describe X and X^* as: $X_1, \cdots, X_n \overset{\text{IID}}{\sim} Poi(\mu)$, $X^* \sim Ber(p(\mu))$ where $p(\mu) = \sum_{k>15} e^{-\mu}\mu^k/k!$, and X and X^* are independent.

Example (Food expenditure)**.** Suppose we collect data from n Duke undergraduates on their (average) weekly expenditure on food $X_1, \cdots, X_n$. We might be interested in predicting $X^* = X_{n+1}$, the (average) amount a (hypothetical) future student is likely to pay on food per week. We can model $X_1, \cdots, X_n, X_{n+1} \overset{\text{IID}}{\sim} N(\mu, \sigma^2)$, with $(\mu, \sigma^2) \in (-\infty, \infty) \times (0, \infty)$ as unknown model parameters.

Example (Food expenditure (contd.))**.** We might also be interested in predicting the difference $X^* = X_{n+1} - X_{n+2}$ in expenditures for two (hypothetical) future students. If we model $X_1, \cdots, X_n, X_{n+1}, X_{n+2} \overset{\text{IID}}{\sim} N(\mu, \sigma^2)$, then we have the following model on X and X^*: $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$, $X^* \sim N(0, 2\sigma^2)$, X and X^* independent.

Example (Highway accident)**.** Data collected in 1973 from $n = 39$ sections of large highways in Minnesota. Measurements include accident rates (Rate_i) (in million vehicle-miles) and road characteristics z_i comprising of (an intercept plus) average daily traffic count (ADT_i, in thousands), truck volume as percentage of total volume (Trks_i), number of access points per mile (Acpt_i) and speed limit (Slim_i, in mph). We may be interested in predicting accident rate Y^* at a different section of a given road characteristic z^*. A reasonable model is: $Y_i = z_i^T\beta + \epsilon_i$, $Y^* = z^{*T}\beta + \epsilon^*$, $\epsilon_1, \cdots, \epsilon_n, \epsilon^* \overset{\text{IID}}{\sim} N(0, \sigma^2)$.

From the above examples it is clear that we are discussing prediction of a variable X^*, given observation on data X in the following context: $X \sim f(x|\theta)$, $X^* \sim f^*(x^*|\theta)$, $\theta \in \Theta^*$, for some collections of pdfs/pmfs $f(x|\theta)$ and $f^*(x^*|\theta)$ indexed by a common parameter $\theta \in \Theta$. Also note that in all of the above examples the model parameters are not real, physical quantities that we could measure if we had more resources (unlike the opinion poll example where the parameter is the actual proportion of supporters, a measurable quantity). For such examples, prediction might be a more useful data

analysis task than inference on the model parameters. This will be highlighted for the linear Gaussian model. Below we discussion classical and Bayesian approaches to prediction.

Bayesian approach

Prediction under a Bayesian formulation is conceptually very straightforward. Suppose we have $X \sim f(x|\theta)$, $X^* \sim f^*(x^*|\theta)$, X and X^* independent, and $\theta \in \Theta$ is assigned a prior $\pi(\theta)$. Once we observe $X = x$, the other variables X^* and θ are jointly described by the conditional pdf

$$h^*(x^*, \theta|x) = f(x^*|\theta)\pi(\theta|x), \quad x^* \in S^*, \theta \in \Theta.$$

This holds because

$$\begin{aligned} h^*(x^*, \theta|x) &= \text{const} \times f(x|\theta)f^*(x^*|\theta)\pi(\theta) \\ &= \text{const} \times \{f(x|\theta)\pi(\theta)\} \times f^*(x^*|\theta) \\ &= \text{const} \times \{\text{const} \times \pi(\theta|x)\} \times f^*(x^*|\theta) \\ &= \text{const} \times f^*(x^*|\theta)\pi(\theta|x). \end{aligned}$$

The last constant term must be 1, because both $h^*(x^*, \theta|x)$ and $f^*(x^*|\theta)\pi(\theta|x)$ are pdfs in (x^*, θ).

The conditional pdf $f^*(x^*|x)$ is now obtained by integrating out θ from $h^*(x^*, \theta|x)$, i..e,

$$f^*(x^*|x) = \int_\Theta f^*(x^*|\theta)\pi(\theta|x)d\theta.$$

Intuitively, the predictive distribution $f^*(x^*|x)$ stands for the following. If we knew θ, we would use $f^*(x^*|\theta)$ to describe X^*. But we do not know θ and our understanding of it is represented by the posterior pdf $\pi(\theta|x)$ given $X = x$. So we must combine our representation of X^* given θ with our representation of θ to get $f^*(x^*|x) = \int_\Theta f^*(x^*|\theta)\pi(\theta|x)d\theta$.

Posterior predictive distribution of future observation for conjugate models

Consider data X and future observation X^* modeled as $X \sim Bin(n, p)$, $X^* \sim Bin(m, p)$, X and X^* are independent, $p \in [0, 1]$ assigned a $Be(a, b)$ prior pdf. Then,

$$f^*(x^*|x) = \int_0^1 \binom{m}{x^*} p^{x^*}(1-p)^{m-x^*}\pi(p|x)dp, \quad x^* \in \{0, 1, \cdots, m\}.$$

But $\pi(p|x) = Be(a' = a + x, b' = b + n - x)$ and so, for any $x^* \in \{0, \cdots, m\}$,

$$\begin{aligned} f^*(x^*|x) &= \int_0^1 \binom{m}{x^*} p^{x^*}(1-p)^{m-x^*} \frac{p^{a'-1}(1-p)^{b'-1}}{B(a',b')} dp \\ &= \binom{m}{x^*} \frac{1}{B(a',b')} \int_0^1 p^{a'+x^*-1}(1-p)^{b'+m-x^*-1} dp \\ &= \binom{m}{x^*} \frac{B(a'+x^*, b'+m-x^*)}{B(a',b')}. \end{aligned}$$

Here we could evaluate the integral $\int_\Theta f^*(x^*|\theta)\pi(\theta|x)d\theta$ because it boils down to evaluating the normalizing constant of a function that is a constant multiple of a beta density. Similar calculations will be possible for any conjugate model (try the Poisson model).

Special calculations for Gaussian linear models

The same applies to a conjugate Gaussian linear model:

$$\begin{aligned} Y_i &= z_i^T \beta + \epsilon_i, i = 1, \cdots, n \\ Y^* &= z^{*T} \beta + \epsilon^* \\ \epsilon_1, \cdots, \epsilon_n, \epsilon^* &\overset{\text{IID}}{\sim} N(0, \sigma^2) \\ \pi(\beta, \sigma^2) &= N_p\chi^{-1}(m_0, K_0, r_0, s_0^2) \\ \text{Or, } \pi(\beta, \sigma^2) &= \text{const}/\sigma^2 \end{aligned}$$

and we can carry out the integration $\int f^*(y^*|\beta, \sigma^2)\pi(\beta, \sigma^2|y)$ analytically. Note that for either choice of the prior, $\pi(\beta, \sigma^2) = N_p\chi^{-2}(m_n, K_n, r_n, s_n^2)$, with formulas for m_n, K_n, r_n and s_n^2 depending on the particular choice.

To see what this predictive distribution looks like, we need to basic results in probability:

RESULT 1. If $W \sim N(a, b^2)$ and $U|(W = w) \sim N(w, c^2)$ then $U \sim N(a, b^2 + c^2)$.

Proof. Clearly $U = W + Z$ where $Z \sim N(0, c^2)$ and is independent of W. But two independent normals add to a normal with means and variances added, therefore $U = W + Z \sim N(a + 0, b^2 + c^2) = N(a, b^2 + c^2)$. □

RESULT 2. If $(W, V) \sim N_p\chi^{-2}(m, K, r, s^2)$ and $U|(W = w, V = v) \sim N(a^T w, v)$ then $T = \frac{U-m}{s\sqrt{1+a^T K^{-1}a}} \sim t(r)$.

Proof. We know $rs^2/V \sim \chi^2(r)$. Think about the description of (U, W) given $V = v$. This is precisely, $U|(W = w) \sim N(a^T w, v)$ and $a^T W \sim N(a^T m, va^T K^{-1}a)$, therefore, by Result 1, still under the condition $V = v$, $U \sim N(m, v(1 + a^T K^{-1}a)) = N(m, v/k_a)$ where $k_a = 1/(1 + a^T K^{-1}a)$. But this description of U given $V = v$, coupled with the description $rs^2/V \sim \chi^2(r)$ means that (U, V) must have $N_1\chi^{-2}(m, k_a, r, s^2)$ distribution. From properties of this distribution $T = \frac{U-m}{s\sqrt{1/k_a}} \sim t(r)$. □

So, for the conjugate (or reference) Gaussian linear model, under the posterior predictive distribution,

$$\frac{Y^* - z^{*T} m_n}{s_n\sqrt{1 + z^{*T} K_n^{-1} z^*}} \sim t(r_n).$$

So, to summarize this posterior-predictive pdf of Y^*, we could report its q-th posterior predictive quantile as

$$z^{*T} m_n + \Phi_{r_n}^{-1}(q) s_n \sqrt{1 + z^{*T} K_n^{-1} z^*},$$

where Φ_r denotes the cdf of $t(r)$. In particular, the posterior-predictive median of Y^* is $z^{*T} m_n$ and the $100(1 - \alpha)\%$ central predictive credible interval for Y^* is $z^{*T} m_n \mp z_{r_n}(\alpha) s_n \sqrt{1 + z^{*T} K_n^{-1} z^*}$.

For the reference prior analysis, $m_n = \hat{\beta}_{\text{LS}}$, $K_n = Z^T Z$, $r_n = n - p$ and $s_n = s_{y|z}$, and hence a $100(1 - \alpha)\%$ posterior predictive interval for Y^* is

$$z^{*T} \hat{\beta}_{\text{LS}} \mp z_{n-p}(\alpha) s_{y|z} \sqrt{1 + z^{*T} (Z^T Z)^{-1} z^*}.$$

Classical approach

The main vehicle of prediction in classical statistics is the so-called plug-in approach. Suppose we obtain an estimate $\hat{\theta}(x)$ of θ from observation $X = x$ (based on ML or other considerations). Then the predictive description of X^* given $X = x$ is the pdf/pmf $\hat{f}^*(x^*|x) = f^*(x^*|\theta = \hat{\theta}(x))$. Although this is a reasonable approach, there is one difficulty. We essentially took the point summary $\hat{\theta}(x)$ to capture all uncertainty about θ. This goes against our intuition of uncertainty associated with statistical modeling that encouraged us to consider interval summaries over point summaries.

This difficulty can be explored formally as follows. Consider the Gaussian linear model as above [but without the prior specification on (β, σ^2)]. The plug-in predictive of Y^* is $N(z^{*T} \hat{\beta}_{\text{LS}}, s_{y|z}^2)$. Based on this pdf, a $100(1 - \alpha)\%$ predictive interval for Y^* is

$$z^{*T} \hat{\beta}_{\text{LS}} \mp z(\alpha) s_{y|z}.$$

But does this interval really guarantee a $(1-\alpha)$ frequentist coverage of Y^*? The actual coverage at any (β, σ^2) is

$$P_{[Y,Y^*|\beta,\sigma^2]}(Y^* \in z^{*T}\hat{\beta}_{\text{LS}} \mp z(\alpha)s_{y|x}) = P_{[Y,Y^*|\beta,\sigma^2]}\left(-z(\alpha) \leq \frac{Y^* - z^{*T}\hat{\beta}_{\text{LS}}}{s_{y|x}} \leq z(\alpha)\right)$$
$$= 2\Phi_{n-p}\left(\frac{z(\alpha)}{\sqrt{1 + z^{*T}(Z^TZ)^{-1}z^*}}\right) - 1$$

because, $Y^* - z^{*T}\hat{\beta}_{\text{LS}} \sim N(0, \sigma^2\{1 + z^{*T}(Z^TZ)^{-1}z^*\})$ and $(n-p)s^2_{y|x}/\sigma^2 \sim \chi^2(n-p)$. This coverage is strictly smaller than $1-\alpha$.

Of course a correct coverage is given by the corrected predictive interval:

$$z^{*T}\hat{\beta}_{\text{LS}} \mp z_{n-p}(\alpha)s_{y|x}\sqrt{1 + z^{*T}(Z^TZ)^{-1}z^*}.$$

However, such fixes are not generally available for non-normal models. Calculating the coverage can be a challenging task. Even normal approximations to the MLE may not salvage the situation, because we also need to account for X^*. However, simulations techniques (as we saw in labs) can be used to approximate coverage probabilities of a given predictive interval procedure.