STAT 30850 Homework 4 Sarah Adilijiang

Problem 1: Randomized Inference

(a) Given a fixed feature vector X_i: $Y_i \sim \text{Binomial}(n_i, f(X_i))$ where $f(X_i) = h(X_i^T\beta)$

$$\Rightarrow P(Y_i = y_i | X_i) = \binom{n_i}{y_i} f(X_i)^{y_i} (1-f(X_i))^{n_i - y_i}$$

For each i, draw $Z_i \sim \text{Binomial}(Y_i, \lambda)$ where $\lambda \in (0,1)$

$$\Rightarrow P(Z_i = z_i | Y_i = y_i) = \binom{y_i}{z_i} \lambda^{z_i} (1-\lambda)^{y_i - z_i}, \text{ where } z_i \le y_i \le n_i$$

$$\Rightarrow P(Z_i = z_i | X_i) = \sum_{y_i = z_i}^{n_i} P(Z_i = z_i | Y_i = y_i) \cdot P(Y_i = y_i | X_i)$$

$$= \sum_{y_i = z_i}^{n_i} \binom{y_i}{z_i} \lambda^{z_i} (1-\lambda)^{y_i - z_i} \cdot \binom{n_i}{y_i} f(X_i)^{y_i} (1-f(X_i))^{n_i - y_i}$$

$$= \sum_{y_i = z_i}^{n_i} \frac{\cancel{y_i!}}{z_i!(y_i - z_i)!} \lambda^{z_i} (1-\lambda)^{y_i - z_i} \frac{n_i!}{\cancel{y_i!}(n_i - y_i)!} f(X_i)^{y_i - z_i} \cdot f(X_i)^{z_i} \cdot (1-f(X_i))^{n_i - (y_i - z_i) - z_i}$$

$$\underline{\underline{\text{set } m_i = y_i - z_i}} \; \frac{n_i!}{z_i!(n_i - z_i)!} (\lambda f(X_i))^{z_i} \sum_{m_i = 0}^{n_i - z_i} \frac{(n_i - z_i)!}{m!(n_i - z_i - m_i)!} ((1-\lambda) f(X_i))^{m_i} (1-f(X_i))^{n_i - z_i - m_i}$$

$$= \binom{n_i}{z_i} (\lambda f(X_i))^{z_i} [(1-\lambda) f(X_i) + 1 - f(X_i)]^{n_i - z_i}$$

$$= \binom{n_i}{z_i} (\lambda f(X_i))^{z_i} (1 - \lambda f(X_i))^{n_i - z_i}$$

$$\Rightarrow Z_i | X_i \sim \text{Binomial}(n_i, \lambda f(X_i))$$

(b)
$$P(Y_i | X_i, Z_i) = \frac{P(Z_i | Y_i, X_i) \cdot p(X_i, Y_i)}{P(X_i, Z_i)} = \frac{P(Z_i | Y_i) \cdot P(Y_i | X_i) P(X_i)}{P(Z_i | X_i) P(X_i)} = \frac{P(Z_i | Y_i) \cdot P(Y_i | X_i)}{P(Z_i | X_i)}$$

$$= \frac{\binom{y_i}{z_i} \lambda^{z_i} (1-\lambda)^{y_i - z_i} \cdot \binom{n_i}{y_i} f(X_i)^{y_i} (1-f(X_i))^{n_i - y_i}}{\binom{n_i}{z_i} (\lambda f(X_i))^{z_i} (1 - \lambda f(X_i))^{n_i - z_i}}$$

$$= \frac{\binom{n_i}{z_i} (\lambda f(X_i))^{z_i} \cdot \binom{n_i - z_i}{y_i - z_i} ((1-\lambda) f(X_i))^{y_i - z_i} (1-f(X_i))^{n_i - y_i}}{\binom{n_i}{z_i} (\lambda f(X_i))^{z_i} \cdot (1 - \lambda f(X_i))^{n_i - z_i}}$$

$$= \binom{n_i - z_i}{y_i - z_i} \left[\frac{(1-\lambda) f(X_i)}{1 - \lambda f(X_i)}\right]^{y_i - z_i} \left[\frac{1 - f(X_i)}{1 - \lambda f(X_i)}\right]^{n_i - y_i}$$

$$\Rightarrow Y_i | X_i, Z_i \sim \text{Binomial}\left(n_i - z_i, \frac{(1-\lambda) f(X_i)}{1 - \lambda f(X_i)}\right)$$

(c) $Z_i \sim \text{Binomial}(Y_i, \lambda)$, where $\lambda \in (0,1)$ represents the chance of being recorded for the exploratory data

Since $Z_i | X_i \sim \text{Binomial}(n_i, \lambda f(x_i))$ & $Y_i | X_i, Z_i \sim \text{Binomial}\left(n_i - Z_i, \frac{(1-\lambda)f(x_i)}{1-\lambda f(x_i)}\right)$

$$\Rightarrow \begin{cases} \text{the probability } \lambda f(x_i) \text{ is a monotone increasing function of } \lambda \\ \text{the probability } \frac{(1-\lambda)f(x_i)}{1-\lambda f(x_i)} = 1 - \frac{1-f(x_i)}{1-\lambda f(x_i)} \text{ is a monotone decreasing function of } \lambda \end{cases}$$

$\because \lambda \in (0,1) \quad \therefore \quad \lambda f(x_i) \in (0, f(x_i))$ & $\frac{(1-\lambda)f(x_i)}{1-\lambda f(x_i)} \in (f(x_i), 0)$

① when λ is large, i.e. $\lambda \to 1$:

$$\begin{cases} Z_i \sim \text{Binomial}(Y_i, \lambda \to 1) \Rightarrow Z_i \text{ would be close to } Y_i \\ \lambda f(x_i) \to f(x_i) \Rightarrow Z_i | X_i \sim \text{Binomial}(n_i, p \to f(x_i)) \Rightarrow E(Z_i|X_i) \approx n_i f(x_i) \\ \frac{(1-\lambda)f(x_i)}{1-\lambda f(x_i)} \to 0 \Rightarrow Y_i | X_i, Z_i \sim \text{Binomial}(n_i - Z_i, p \to 0) \end{cases}$$

② when λ is small, i.e. $\lambda \to 0$:

$$\begin{cases} Z_i \sim \text{Binomial}(Y_i, \lambda \to 0) \Rightarrow Z_i \text{ would be close to } 0 \\ \lambda f(x_i) \to 0 \Rightarrow Z_i | X_i \sim \text{Binomial}(n_i, p \to 0) \\ \frac{(1-\lambda)f(x_i)}{1-\lambda f(x_i)} \to f(x_i) \Rightarrow Y_i | X_i, Z_i \sim \text{Binomial}(n_i - Z_i, p \to f(x_i)) \Rightarrow E(Y_i|X_i, Z_i) \approx (n_i - Z_i) f(X_i) \approx n_i f(x_i) \end{cases}$$

Note that in both ① & ②, we can estimate $\hat{f}(x_i) = \frac{\hat{Y}_i}{n_i}$ $(i=1,\cdots,n)$ thus get the estimate of function $f(\cdot)$

Problem 2: Conformal Prediction

Assume $(X_1,Y_1), \cdots, (X_n,Y_n), \underline{(X_{n+1},Y_{n+1})}$ are iid

Now we need to look for values y, s.t. in the data set $(X_1,Y_1), \cdots, (X_n,Y_n), \underline{(X_{n+1},y)}$, the last point $\underline{(X_{n+1},y)}$ does not appear "unusual" relative to the others

Thus we train the KNN model on the data set $(X_1,Y_1), \cdots, (X_n,Y_n), \underline{(X_{n+1},y)}$:

For each (X_i,Y_i), suppose $X_{i_1}, \cdots, X_{i_k}$ $(i_1,\cdots,i_k \neq i)$ are k nearest neighbors of X_i $(i=1,\cdots,n,\underline{n+1})$

$$\Rightarrow \hat{Y}_i = \frac{1}{k}\sum_{l=1}^{k} Y_{i_l} \text{ , where } Y_{i_l} = \begin{cases} Y_{i_l} & \text{if } i_l = 1,\cdots,n \\ y & \text{if } i_l = n+1 \end{cases}$$

$$\Rightarrow \text{Residuals} = R_i^y = \begin{cases} |Y_i - \hat{Y}_i| & \text{if } i=1,\cdots,n \\ |y - \hat{Y}_{n+1}| & \text{if } i=n+1 \end{cases}$$

$\Rightarrow$ The Prediction Interval for Y_{n+1} is :

$$\{y \in R \ : \ R_{n+1}^y \leq \lceil(1-\alpha)(n+1)\rceil\text{-th smallest value of } R_1^y, \cdots, R_n^y, R_{n+1}^y\}$$