Classification with Confidence

- A New Framework of Classification
  \( x \in \mathbb{X}, \ y \in \{0, 1, 5\}, \ (x, y) \sim P(x, y) \)
  
  Want to find a classification rule \( f(x): \mathbb{X} \to \{0, 0.5, 1, 1.5\} \)

  define \( C_0 = \{ x \mid f(x) = 0 \} \)
  \( C_1 = \{ x \mid f(x) = 1 \} \)
  \( C_0 \cup C_1 = \mathbb{X} \)

  \( C_0 = C_0 \cap C_2 \) is where the data could be "classified" to \( \{0, 1, 5\} \), which called "Ambiguity".

  new question becomes a trade-off bw classification accuracy and the portion of ambiguity.
  
  \( \rightarrow \) want to minimize the ambiguity portion w/a control on accuracy, i.e.

  given \( P_0(C_0) = P(x \in C_0 \mid Y = 0) \geq 1 - \xi_0 \)
  \( P_1(C_1) = P(x \in C_1 \mid Y = 1) \geq 1 - \xi_1 \)

  and \( C_0 \cup C_1 = \mathbb{X} \)

  want to minimize

  \( P(C_0 \cap C_1) = \Pi_0 P_0(C_0 \cap C_1) + \Pi_1 P_1(C_0 \cap C_1) \)

  \( \Pi_j = P(Y = j) \ j = 0, 1 \)

- Thm 1 gives a solution to the target problem \((x)\)

  Thm 1 says, \( f(x) < \delta_0 \ \& \ \delta_1 < 1 \) a solution to \((x)\) is given by

  \( C_0 = \{ x \mid \eta(x) \leq t_0(\delta_0) \} \)
  \( C_1 = \{ x \mid \eta(x) \geq t_1(\delta_1) \} \cup C_0 \)

  where to and t1 are chosen s.t.

  \( P_0(\eta(x) \leq t_0(\delta_0)) = 1 - \xi_0 \)
  \( P_1(\eta(x) \geq t_1(\delta_1)) = 1 - \xi_1 \)
Proof:

W.l.o.g., assume $x_0$.

Suppose we have a new classifier $f(x)$ with

$S_0 = \{ x \mid f(x) = 0 \}$, $P_0(S_0) > 1 - d_0$

$S_1 = \{ x \mid f(x) = 1 \}$, $P_0(S_1) > 1 - d_1$

$S_{01} = S_0 \cap S_1$

Want to show $P_0(C_0) \leq P_0(S_{01})$

\[
P_0(C_0) \leq P_0(S_0)
\]

\[
\implies P_0(C_0 \cup C_1) \leq P_0(S_0 \cup C_0) + P_0(S_0 \cup C_1)
\]

\[
\implies P_0(C_0 \cup C_1 \cap S_0) + P_0(C_0 \cup C_1 \cap S_0^c) \leq P_0(C_0 \cup C_1 \cap S_0) + P_0(S_0 \cup C_1)
\]

\[
\implies P_0(C_0 \cap S_0) + P_0(C_0 \cap S_0^c) + P_0(C_1 \cap S_0) + P_0(C_1 \cap S_0^c)
\]

\[
\leq P_0(C_0 \cap S_0) + P_0(C_0 \cap S_0^c) + P_0(S_0 \cup C_1)
\]

\[
\implies P_0(C_0 \cap S_0^c) + P_0(C_0 \cap S_0^c) \leq P_0(S_0 \cap C_0^c) + P_0(S_1 \cap C_0^c)
\]

\[
\implies P_0(C_0 \cap S_0^c) + P_0(C_0 \cap S_0^c) \leq P_0(S_0 \cap C_0^c) + P_0(S_0 \cap C_1^c)
\]

\[
\therefore P_0(C_0) - P_0(S_{01}) = P_0(C_0 \cap S_0^c) + P_0(C_0 \cap S_0^c) - P_0(C_0 \cap S_0^c) - P_0(C_0 \cap S_1^c) = P_0(C_0 \cap S_0^c) - P_0(C_0 \cap S_1^c)
\]
If we consider a discrete \( X \)

\[
\forall x \in \text{CNS}_i^c, \quad \frac{P(Y=1 | X=x)}{P(Y=0 | X=x)} = \frac{\eta}{1-\eta} \geq \frac{t_1(a_i)}{1-t_1(a_i)}
\]

\[
\frac{P(Y=1, X=x)}{P(Y=0, X=x)} = \frac{P(Y=1, X=x)}{P(Y=0, X=x)} \geq \frac{P(Y=1)}{P(Y=0)}
\]

\[
P_0(x) \leq \frac{-t_1}{t_1} \cdot \frac{\pi_1}{\pi_0} P_1(x)
\]

\[
P_0(\text{CNS}_i^c) \leq \frac{1-t_1}{t_1} \cdot \frac{\pi_1}{\pi_0} P_1(\text{CNS}_i^c)
\]

Similarly,

\[
P_0(\text{SINC}_c^c) \leq \frac{1-t_1}{t_1} \cdot \frac{\pi_1}{\pi_0} P_1(\text{SINC}_c^c)
\]

\[
P_0(\text{CO}_i) - P_0(\text{SO}_i) \leq \frac{\pi_1}{\pi_0} \frac{1-t_1}{t_1} \left[ P_1(\text{CNS}_i^c) - P_1(\text{SINC}_c^c) \right]
\]

\[
P_0(\text{CO}_i) - P_0(\text{SO}_i) \leq 0
\]

b/c

\[
P_1(\text{CNS}_i^c) \leq P_1(\text{SINC}_c^c)
\]

\[\text{area}\]

\[\text{SINC}_c^c\]

\[\text{CNS}_i^c\]

\[\text{pi}_c\]

\[P_1(\text{CNS}_i^c) \leq P_1(\text{SINC}_c^c)\]
• If \( \hat{\eta} \) is estimated by \( \hat{\eta} \), then classifications can be estimated as

\[
\hat{C}_0 = \{ x : \hat{\eta}(x) \leq t_0 \} \quad \hat{C}_1 = \{ x : \hat{\eta}(x) > t_1 \} \quad \text{and} \quad \hat{C}_x = \hat{C}_0 \cup \hat{C}_1
\]

Estimate \( t_0 \) and \( t_1 \) depends on how to model the claim.

If order all \( \hat{\eta}(x_i) \)'s, then \( d_0, x_i \) can get controlled by choosing sample quantiles.

- Want the estimate \( \hat{\eta} \) to be good

- If \( G_i \) is cdf of \( \eta(x) \) under \( P_i \)

\[
(\star \star \star) \quad \text{want } b_1 \leq 1 \quad G_i(t_i + b) - G_i(t_i) \leq b_2 \leq 1
\]

\( G_i \) should be smooth at true \( t_i \)'s

• \((\delta_n, \rho_n)\) accuracy

An estimator \( \hat{\eta} \) is \((\delta_n, \rho_n)\) accurate if

\[
P(\| \hat{\eta} - \eta \| > \delta_n) \leq \rho_n
\]

• Thm 2

If \( \hat{\eta} \) is \((\delta_n, \rho_n)\)-accurate and condition \((\star \star \star)\) holds, then for each \( r > 0 \), there exists a positive constant \( c \) s.t.

\[
P_j(\hat{C}_j \subseteq C_j) \leq c \left( \delta_n^b + \left( \frac{\log n}{n} \right)^{\frac{1}{2}} \right)^2, \quad j = 0, 1
\]

With probability at least \( 1 - \rho_n - n^{-r} \)

i.e., Li-panelized logistic regression gives \( \delta_n - \rho_n \) accurate estimation.

\[\text{[Note: Continue on next page...]}\]