One-against-all

Daniel Hsu (COMS 4771)

**Theorem.** Let \( \hat{\eta}_1, \ldots, \hat{\eta}_K : \mathcal{X} \to [0, 1] \) be estimates of conditional probability functions \( x \mapsto \mathbb{P}(Y = k \mid X = x) \) for \( k = 1, \ldots, K \), and let
\[
\epsilon := \mathbb{E} \left[ \max_{k=1, \ldots, K} \left| \hat{\eta}_k(X) - \mathbb{P}(Y = k \mid X) \right| \right].
\]

Let \( \hat{f} : \mathcal{X} \to \{1, \ldots, K\} \) be the one-against-all classifier based on \( \hat{\eta}_1, \ldots, \hat{\eta}_K \), i.e.,
\[
\hat{f}(x) = \arg \max_{k=1, \ldots, K} \hat{\eta}_k(x), \quad x \in \mathcal{X},
\]
(with ties broken arbitrarily), and let \( f^* : \mathcal{X} \to \{1, \ldots, K\} \) be the Bayes optimal classifier. Then
\[
\mathbb{P}(\hat{f}(X) \neq Y) - \mathbb{P}(f^*(X) \neq Y) \leq 2\epsilon.
\]

**Proof.** Fix \( x \in \mathcal{X} \), \( y^* := f^*(x) \), and \( \hat{y} := \hat{f}(x) \). Let \( \eta_k(x) := \mathbb{P}(Y = k \mid X = x) \) for all \( k = 1, \ldots, K \). Then
\[
\mathbb{P}(\hat{f}(X) \neq Y \mid X = x) - \mathbb{P}(f^*(X) \neq Y \mid X = x) = \sum_{k \neq y} \eta_k(x) - \sum_{k \neq y^*} \eta_k(x)
= \eta_{y^*}(x) - \eta_{y}(x)
= \underbrace{\hat{\eta}_{y^*}(x) - \hat{\eta}_{y}(x)}_{\leq 0} + \eta_{y^*}(x) - \hat{\eta}_{y^*}(x) + \hat{\eta}_{y}(x) - \eta_{y}(x)
\leq 2 \max_{k=1, \ldots, K} \left| \hat{\eta}_k(x) - \eta_k(x) \right|.
\]

Therefore, taking expectations with respect to \( X \),
\[
\mathbb{P}(\hat{f}(X) \neq Y) - \mathbb{P}(f^*(X) \neq Y) \leq 2 \cdot \mathbb{E} \left[ \max_{k=1, \ldots, K} \left| \hat{\eta}_k(X) - \eta_k(X) \right| \right].
\]

The bound on the excess risk is tight. To see this, suppose for a given \( x \in \mathcal{X} \) (with \( y^* = f^*(x) \) and \( \hat{y} = \hat{f}(x) \)), we have \( \hat{\eta}_{y^*}(x) = \hat{\eta}_{y}(x) - \delta \), but \( \eta_{y^*}(x) = \eta_{y^*}(x) + \epsilon \) and \( \eta_{y}(x) = \eta_{y}(x) - \epsilon \). Then
\[
\eta_{y^*}(x) - \eta_{y}(x) = (\hat{\eta}_{y^*}(x) + \epsilon) - (\hat{\eta}_{y}(x) - \epsilon)
= 2\epsilon - \delta
\]
which tends to \( 2\epsilon \) as \( \delta \to 0 \).