One-against-all

Daniel Hsu (COMS 4771)

Theorem. Let $\hat{\eta}_1, \dots, \hat{\eta}_K \colon \mathcal{X} \to [0,1]$ be estimates of conditional probability functions $x \mapsto \mathbb{P}(Y = k \mid X = x)$ for $k = 1, \dots, K$, and let

$$\epsilon := \mathbb{E}\left[\max_{k=1,\dots,K} \left|\hat{\eta}_k(X) - \mathbb{P}(Y = k \mid X)\right|\right].$$

Let $\hat{f} \colon \mathcal{X} \to \{1, \dots, K\}$ be the one-against-all classifier based on $\hat{\eta}_1, \dots, \hat{\eta}_K$, i.e.,

$$\hat{f}(x) = \operatorname*{arg\,max}_{k=1,\dots,K} \hat{\eta}_k(x), \quad x \in \mathcal{X},$$

(with ties broken arbitrarily), and let $f^\star \colon \mathcal{X} \to \{1, \dots, K\}$ be the Bayes optimal classifier. Then

$$\mathbb{P}(\hat{f}(X) \neq Y) - \mathbb{P}(f^\star(X) \neq Y) \leq 2\epsilon.$$

Proof. Fix $x \in \mathcal{X}$, $y^\star := f^\star(x)$, and $\hat{y} := \hat{f}(x)$. Let $\eta_k(x) := \mathbb{P}(Y = k \mid X = x)$ for all $k = 1, \dots, K$. Then

$$\begin{aligned}
\mathbb{P}(\hat{f}(X) \neq Y \mid X = x) - \mathbb{P}(f^\star(X) \neq Y \mid X = x) &= \sum_{k \neq \hat{y}} \eta_k(x) - \sum_{k \neq y^\star} \eta_k(x) \\
&= \eta_{y^\star}(x) - \eta_{\hat{y}}(x) \\
&= \underbrace{\hat{\eta}_{y^\star}(x) - \hat{\eta}_{\hat{y}}(x)}_{\leq 0} + \eta_{y^\star}(x) - \hat{\eta}_{y^\star}(x) + \hat{\eta}_{\hat{y}}(x) - \eta_{\hat{y}}(x) \\
&\leq 2 \max_{k=1,\dots,K} |\hat{\eta}_k(x) - \eta_k(x)|.
\end{aligned}$$

Therefore, taking expectations with respect to X,

$$\mathbb{P}(\hat{f}(X) \neq Y) - \mathbb{P}(f^\star(X) \neq Y) \leq 2 \cdot \mathbb{E}\left[\max_{k=1,\dots,K} \left|\hat{\eta}_k(X) - \eta_k(X)\right|\right].$$

∎

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

$$\begin{aligned}
\eta_{y^\star}(x) - \eta_{\hat{y}}(x) &= (\hat{\eta}_{y^\star(x)} + \epsilon) - (\hat{\eta}_{\hat{y}}(x) - \epsilon) \\
&= 2\epsilon - \delta
\end{aligned}$$

which tends to 2ϵ as $\delta \to 0$.