Supplemental material: How many faces can it recognize? Performance extrapolation for multi-class classification

Charles Zheng, Rakesh Achanta and Yuval Benjamini

May 2, 2016

1 Proofs

Theorem 2.1. *(i) an OVO recognition systems equipped with a continuous binary classifier is separable; (ii) an OVA recognition systems equipped with a continuous binary scoring rule is separable; (iii) a kNN recognition system with fixed neighborhood size $\alpha \in (0, 1)$ is separable; (iv) multinomial logistic regression recognition systems are separable; (v) LDA recognition systems are separable.*

Theorem 3.1. *Let U be defined as the random variable*

$$U = u(F, Y)$$

for X, Y drawn from $p(x, y) = p(x)p(y|x)$, and $\hat{F}(X) = \frac{1}{r_1}\sum_{j=1}^{r_1} \delta Y^j$ with $Y^i \stackrel{iid}{\sim} p(y|X)$ Then $p_k = \boldsymbol{E}[U^{k-1}]$.

Proof. Write $q^{(i)}(y) = \mathcal{Q}(\hat{F}_i, 0, y)$, and let $Y^{(i),*} \sim p(y|X^{(i)}$ for $i = 1, \ldots, k$. Note that by using conditioning and conditional independence, p_k can be written

$$\begin{aligned}
p_k &= \mathbf{E}\left[\frac{1}{k}\sum_{i=1}^{k}\Pr[q^{(i)}(Y^{(i),*}) > \max_{j\neq i} q^{(j)}(Y^{(i),*})]\right] \\
&= \mathbf{E}\left[\Pr[q^{(1)}(Y^{(1),*}) > \max_{j\neq 1} q^{(j)}(Y^{(1),*})]\right] \\
&= \mathbf{E}[\Pr[q^{(1)}(Y^{(1),*}) > \max_{j\neq 1} q^{(j)}(Y^{(1),*})|Y^{(1),*}, \hat{F}_1]] \\
&= \mathbf{E}[\Pr[\cap_{j>1} q^{(1)}(Y^{(1),*}) > q^{(j)}(Y^{(1),*})|Y^{(1),*}, \hat{F}_1]] \\
&= \mathbf{E}[\prod_{j>1}\Pr[q^{(1)}(Y^{(1),*}) > q^{(j)}(Y^{(1),*})|Y^{(1),*}, \hat{F}_1]] \\
&= \mathbf{E}[\Pr[q^{(1)}(Y^{(1),*}) > q^{(2)}(Y^{(1),*})|Y^{(1),*}, \hat{F}_1]^{k-1}] \\
&= \mathbf{E}[u(\hat{F}_1, Y^{(1),*})^{k-1}] = \mathbf{E}[U^{k-1}].
\end{aligned}$$

□

Theorem 3.2. *Let U be defined as in Theorem 2.1, and let ν denote the law of U. Then, for any probability distribution ν' on $[0,1]$, one can construct a joint distribution $p(x,y)$ and a scoring rule $\mathcal{Q}$ such that $\nu = \nu'$.*

Proof. Let X and Y have the degenerate joint distribution $X = Y \sim Unif[0,1]$. Let G be the cdf of ν, $G(x) = \int_0^x d\nu(x)$, and let $H(u) = \sup_x\{G(x) \leq u\}/$. Define $\mathcal{Q}$ by

$$\mathcal{Q}(F,0,y) = \begin{cases} 0 & \text{if } \mu(F) > y + H(y) \\ 0 & \text{if } y + H(y) > 1 \text{ and } \mu(F) \in [H(y) - y, y] \\ 1 + \mu(F) - y & \text{if } \mu(F) \in [y, y + H(y)] \\ 1 + y + \mu(F) & \text{if } \mu(F) + H(y) > 1 \text{ and } \mu(F) \in [0, H(y) - y]. \end{cases}$$

One can verify that $u(\hat{F}(X), X) = H(X)$. Therefore, the cdf of U is equal to G, as needed. □

Theorem 3.3. *Let U be defined as in Theorem 2.2, and let ν denote the law of U. Suppose (X,Y) has a density $p(x,y)$ with respect to Lebesgue measure on $\mathcal{X} \times \mathcal{Y}$, and with probability one, $\mathcal{Q}(y, 0, \hat{F}(X))$ satisfies the property of monotonicity*

$$p(y|x) > p(y'|x) \textit{ implies } \mathcal{Q}(\hat{F}(X), 0, y) > \mathcal{Q}(\hat{F}(X), 0, y')$$

and the property of tie-breaking, then μ has a density $\eta(u)$ on $[0,1]$ which is monotonic in u.

Proof. Choose $0 < u < v < 1$ and $0 < \delta < \min(u, 1 - v, v - u)$. For $x \in \mathcal{X}$, define the set

$$\underline{J}_x = \{y \in \mathcal{Y} : \int_{\mathcal{Y}} I(p(y|x) > p(w|x))p(w|x)dw \in [u - \delta, u + \delta]\}$$

and

$$\bar{J}_x = \{y \in \mathcal{Y} : \int_{\mathcal{Y}} I(p(y|x) > p(w|x))p(w|x)dw \in [v - \delta, v + \delta]\}$$

One can verify that for all $x \in \mathcal{X}$,

$$\int_{\underline{J}_x} p(y|x)dy \leq \int_{\bar{J}_x} p(y|x)dy.$$

Yet, since

$$\Pr[U \in [u - \delta, u + \delta]] = \Pr[\cup_{\mathcal{X}} x \times \underline{J}_x]$$
$$\Pr[U \in [v - \delta, v + \delta]] = \Pr[\cup_{\mathcal{X}} x \times \underline{J}_x].$$

we obtain

$$\Pr[U \in [u - \delta, u + \delta]] \leq \Pr[U \in [v - \delta, v + \delta]].$$

Taking $\delta \to 0$, we conclude the theorem. $\square$