Optimizing F-Measures: A Tale of Two Approaches

Nan Ye YENAN@COMP.NUS.EDU.SG
Department of Computer Science, National University of Singapore, Singapore 117417

Kian Ming A. Chai CKIANMIN@DSO.ORG.SG
DSO National Laboratories, Singapore 118230

Wee Sun Lee LEEWS@COMP.NUS.EDU.SG
Department of Computer Science, National University of Singapore, Singapore 117417

Hai Leong Chieu CHAILEON@DSO.ORG.SG
DSO National Laboratories, Singapore 118230

Appendix. Proofs

We shall often drop θ from the notations whenever there is no ambiguity.

Lemma 1. *For any $\epsilon > 0$, $\lim_{n\to\infty} \mathrm{P}(|F_{\beta,n}(\theta) - F_\beta(\theta)| < \epsilon) = 1$.*

Proof for Lemma 1. By the law of large numbers, for any $\epsilon_1 > 0$, $\eta > 0$, there exists an N (depending on ϵ_1 and η only) such that for all $n > N$, for any i, j

$$\mathrm{P}(|p_{ij,n} - p_{ij}| < \epsilon_1) > 1 - \eta/3, \tag{1}$$

Note that only $p_{ij,n}$ is a random variable in the above inequality. Using the union bound, it follows that with probability at least $1-\eta$, the following hold simultaneously,

$$|p_{11,n} - p_{11}| < \epsilon_1, |p_{10,n} - p_{10}| < \epsilon_1, |p_{01,n} - p_{01}| < \epsilon_1$$

Let $a = (1+\beta^2)p_{11}$, $b = \beta^2\pi_1 + p_{11} + p_{01}$, $\epsilon_1 = \frac{b\epsilon/(1+\beta^2)}{\frac{2a}{b}+2\epsilon+1}$, then when the above inequalities hold simultaneously, it is easy to verify that $2(1+\beta^2)\epsilon_1 < b$, and

$$\begin{aligned}
\frac{a}{b} - \epsilon &\le \frac{a - (1+\beta^2)\epsilon_1}{b + 2(1+\beta^2)\epsilon_1} \\
&< \frac{(1+\beta^2)p_{11,n}}{\beta^2(p_{11,n} + p_{10,n}) + p_{10,n} + p_{01,n}} \\
\frac{a}{b} + \epsilon &\ge \frac{a + (1+\beta^2)\epsilon_1}{b - 2(1+\beta^2)\epsilon_1} \\
&> \frac{(1+\beta^2)p_{11,n}}{\beta^2(p_{11,n} + p_{10,n}) + p_{10,n} + p_{01,n}}
\end{aligned}$$

That is, $F_\beta(\theta) - \epsilon < F_{\beta,n}(\theta) < F_\beta(\theta) + \epsilon$.

Hence for any $\epsilon > 0$, $\eta > 0$, there exists N such that for all $n > N$, $\mathrm{P}(|F_{\beta,n}(\theta) - F_\beta(\theta)| < \epsilon) > 1 - \eta$. □

Lemma 2. *Let $r(n,\eta) = \sqrt{\frac{1}{2n}\ln\frac{6}{\eta}}$. When $r(n,\eta) < \frac{\beta^2\pi_1}{2(1+\beta^2)}$, then with probability at least $1-\eta$, $|F_{\beta,n}(\theta) - F_\beta(\theta)| < \frac{3(1+\beta^2)r(n,\eta)}{\beta^2\pi_1 - 2(1+\beta^2)r(n,\eta)}$.*

Proof for Lemma 2. Let $\eta = 6e^{-2n\epsilon_1^2}$, then $\epsilon_1 = r(n,\eta)$. Using Hoeffding's inequality, for any i, j,

$$\mathrm{P}(|p_{ij,n} - p_{ij}| < \epsilon_1) \quad > 1 - \eta/3 \tag{2}$$

Let $\epsilon_1 = \frac{\beta^2}{1+\beta^2}\frac{\pi_1\epsilon}{3+2\epsilon}$, then $\epsilon = \frac{3(1+\beta^2)\epsilon_1}{\beta^2\pi_1 - 2(1+\beta^2)\epsilon_1} = \frac{3(1+\beta^2)r(n,\eta)}{\beta^2\pi_1 - 2(1+\beta^2)r(n,\eta)}$. From $\beta^2\pi_1 \le b$ and $\frac{a}{b} \le 1$, it follows that $\epsilon_1 \le \frac{b\epsilon/(1+\beta^2)}{\frac{2a}{b}+2\epsilon+1}$. Similarly as in the proof for Proposition 1, we have $\mathrm{P}(|F_{\beta,n}(\theta) - F_\beta(\theta)| < \epsilon) > 1 - \eta$. □

Lemma 2 leads to the following sample complexity: for $\epsilon, \eta > 0$, for $n > \frac{1}{2}(\frac{\beta^2}{1+\beta^2}\frac{\pi_1\epsilon}{3+2\epsilon})^{-2}\ln\frac{6}{\eta}$, with probablity at least $1-\eta$, $|F_{\beta,n}(\theta) - F_\beta(\theta)| < \epsilon$.

The above bounds are not the tightest. For example, Lemma 2 still holds when $\frac{3(1+\beta^2)r(n,\eta)}{\beta^2\pi_1 - 2(1+\beta^2)r(n,\eta)}$ is replaced by the tighter bound $\frac{(1+\beta^2)(2F_\beta(\theta)+1)r(n,\eta)}{\beta^2\pi_1 + p_1(\theta) - 2(1+\beta^2)r(n,\eta)}$, where $p_1(\theta)$ is the probability that θ classifies an instance as positive. In practice, the tighter bound is not useful for estimating the performance of a classifier, because it contains the terms $F_\beta(\theta)$ and $p_1(\theta)$. For the same reason, the tighter bound is also not useful in the uniform convergence that we seek next.

Theorem 3. *Let* $\Theta \subseteq X \mapsto Y$, $d = VC(\Theta)$, $\theta^* = \arg\max_{\theta\in\Theta} F_\beta(\theta)$, *and* $\theta_n = \arg\max_{\theta\in\Theta} F_{\beta,n}(\theta)$. *Let* $\bar{r}(n,\eta) = \sqrt{\frac{1}{n}(\ln\frac{12}{\eta} + d\ln\frac{2en}{d})}$. *If* n *is such that* $\bar{r}(n,\eta) < \frac{\beta^2\pi_1}{2(1+\beta^2)}$, *then with probability at least* $1-\eta$, $F_\beta(\theta_n) > F_\beta(\theta^*) - \frac{6(1+\beta^2)\bar{r}(n,\eta)}{\beta^2\pi_1 - 2(1+\beta^2)\bar{r}(n,\eta)}$.

Proof for Theorem 3. Let $\eta = 12e^{d\ln\frac{2en}{d} - n\epsilon_1^2}$, then $\epsilon_1 = \bar{r}(n,\eta)$. Note that the VC dimension for class consisiting of loss functions of the form $\mathrm{I}(y = i \wedge \theta(x) = j)$ is the same as that for Θ, and the same remark applies for the the class consisting of loss functions of the form $\mathrm{I}(\theta(x) = y)$. By (3.3) in (Vapnik, 1995), for any i, j

$$\mathrm{P}(\sup_\theta |p_{ij,n}(\theta) - p_{ij}(\theta)| < \epsilon_1) \quad > 1 - \eta/3 \tag{3}$$

By the union bound, with probability at least $1-\eta$, the inequalities $\sup_\theta |p_{11,n}(\theta) - p_{11}(\theta)| < \epsilon_1$, $\sup_\theta |p_{10,n}(\theta) - p_{10}(\theta)| < \epsilon_1$, $\sup_\theta |p_{01,n} - p_{01}| < \epsilon_1$, hold simultaneously. Let $\epsilon_1 = \frac{\beta^2}{1+\beta^2}\frac{\pi_1\epsilon}{3+2\epsilon}$, then following the proof of Lemma 2,

$$\begin{aligned}
& F_\beta(\theta_n) - F_\beta(\theta^*) \\
= \;& F_\beta(\theta_n) - F_{\beta,n}(\theta_n) + F_{\beta,n}(\theta_n) - F_\beta(\theta^*) \\
\geq \;& F_\beta(\theta_n) - F_{\beta,n}(\theta_n) + F_{\beta,n}(\theta^*) - F_\beta(\theta^*) \\
\geq \;& -2\epsilon = -\frac{6(1+\beta^2)\bar{r}(n,\eta)}{\beta^2\pi_1 - 2(1+\beta^2)\bar{r}(n,\eta)}
\end{aligned}$$

□

Theorem 4. *For any classifier* θ, $F_\beta(\theta) \leq F_\beta(t^*)$.

Proof for Theorem 4. Let θ be an arbitrary classifier. If $\theta \notin \mathcal{T} \cup \mathcal{T}'$, then when all $x \in X$ are mapped to the number axis using $x \to P(1|x)$, there must be some set B of negative instances which break the positive instances into two sets A and C. Formally, there exist disjoint subsets A, B and C of X such that

$$\begin{gathered}
A \cup C = \{x : \theta(x) = 1\} \\
\theta(B) = \{0\} \\
\sup_{x\in A} P(1|x) \leq \inf_{x\in B} P(1|x) \leq \sup_{x\in B} P(1|x) \leq \inf_{x\in C} P(1|x).
\end{gathered}$$

Without loss of generality we assume $P(A), P(B), P(C) > 0$. Let $a = P(A)$, $x = \mathrm{E}(P(1|X)\big|X \in A)$, $b = P(B)$, $y = \mathrm{E}(P(1|X)\big|X \in B)$, and $c = P(C)$, $z = \mathrm{E}(P(1|X)\big|X \in C)$, then $x \leq y \leq z$. Note that the expectation is taken with respect to X. Let θ_B and θ_C be the same as θ except that $\theta_B(B) = \{1\}$ and $\theta_C(A) = \{0\}$. Thus we have $F_\beta(\theta) = \frac{(1+\beta^2)(ax+cz)}{\beta^2\pi_1+a+c}$, $F_\beta(\theta_B) = \frac{(1+\beta^2)(ax+by+cz)}{\beta^2\pi_1+a+b+c}$, and $F_\beta(\theta_C) = \frac{(1+\beta^2)cz}{\beta^2\pi_1+c}$.

We show that either $F_\beta(\theta_B) \geq F_\beta(\theta)$ or $F_\beta(\theta_C) \geq F_\beta(\theta)$. Assume otherwise, then $F_\beta(\theta) > F_\beta(\theta_B)$, which implies that $ax + cz > (\beta^2\pi_1 + a + c)y$. In addition, $F_\beta(\theta) > F_\beta(\theta_C)$, which implies that $(\beta^2\pi_1 + c)x > cz$. Thus $ax + cz > (\beta^2\pi_1 + c)x + ax > cz + ax$, a contradiction. Hence it follows that we can convert θ to a classifier θ' such that $\theta' \in \mathcal{T} \cup \mathcal{T}'$, and $F_\beta(\theta) \leq F_\beta(\theta') \leq F_\beta(t^*)$. □

Theorem 5. *A rank-preserving function is an optimal score function.*

Proof for Theorem 5. Immediate from Theorem 4. □

Theorem 6. *For any classifier* θ, *any* $\epsilon, \eta > 0$, *there exists* $N_{\beta,\epsilon,\eta}$ *such that for all* $n > N_{\beta,\epsilon,\eta}$, *with probability at least* $1-\eta$, $|\mathrm{E}[F_\beta(\theta(\mathbf{x}), \mathbf{y})] - F_\beta(\theta)| < \epsilon$.

Proof for Theorem 6. This follows closely the proof for Lemma 7. □

Lemma 7. *For any* $\epsilon, \eta > 0$, *there exists* $N_{\beta,\epsilon,\eta}$ *such that for all* $n > N_{\beta,\epsilon,\eta}$, *with probability at least* $1-\eta$, *for all* $\delta \in [0,1]$, $|\mathrm{E}[F_\beta(\mathrm{I}_\delta(\mathbf{x}), \mathbf{y})] - F_\beta(\mathrm{I}_\delta)| < \epsilon$.

Proof for Lemma 7. $p_i(\delta) = \mathrm{E}(\mathrm{I}(\mathrm{I}_\delta(X) = i))$ denotes the probability that an observation is predicted to be in class i, and $p_{j|i}(\delta) = \mathrm{E}(P(j|X)\big|\mathrm{I}_\delta(x) = i)$ denotes the probability that an observation predicted to be in class i is actually in class j. Let $n_i(\delta) = \sum_k \mathrm{I}(\mathrm{I}_\delta(x_k) = i)$, $n_{ji}(\delta) = \sum_k \mathrm{I}(y_k = j \wedge \mathrm{I}_\delta(x_k) = i)$, then $\tilde{p}_i(\delta) = \frac{n_i}{n}$ and $\tilde{p}_{j|i}(\delta) = \frac{n_{ji}(\delta)}{n_i(\delta)}$ are empirical estimates for $p_i(\delta)$ and $p_{j|i}(\delta)$ respectively. We will also need to use $\tilde{p}'_{j|i}(\delta) = \frac{1}{n_i}\sum_i P(j|x)\mathrm{I}(\mathrm{I}_\delta(x) = i)$ as the empirical estimate of $p_{j|i}(\delta)$ based on $\mathbf{x}$ only. Note that $\tilde{p}_i(\delta)$'s and $\tilde{p}'_{j|i}(\delta)$'s are random variables depending on $\mathbf{x}$ only, and $\tilde{p}_{j|i}(\delta)$'s are random variables depending on $\mathbf{x}$ and $\mathbf{y}$. In the following, we shall drop δ from the notations as long as there is no ambiguity. Let $F_\beta(\delta)$ denote the F_β-measure of $\mathrm{I}_\delta(x)$. We have

$$F_\beta(\delta) = \frac{(1+\beta^2)p_1 p_{1|1}}{\beta^2(p_1 p_{1|1} + p_0 p_{1|0}) + p_1} \tag{4}$$

$$F_\beta(\mathrm{I}_\delta(\mathbf{x}), \mathbf{y}) = \frac{(1+\beta^2)\tilde{p}_1 \tilde{p}_{1|1}}{\beta^2(\tilde{p}_1 \tilde{p}_{1|1} + \tilde{p}_0 \tilde{p}_{1|0}) + \tilde{p}_1} \tag{5}$$

The main idea of the proof is to first show that

(a) there is high probability that $\mathbf{x}$ gives good estimates for $p_i(\delta)$'s and $p_{1|i}(\delta)$'s for all δ, and then show that

(b) for such $\mathbf{x}$, there is high probability that $\mathbf{x}, \mathbf{y}$ give good estimates for $p_i(\delta)$'s and $p_{1|i}(\delta)$'s, thus

(c) $F_\beta(\mathrm{I}_\delta(\mathbf{x}), \mathbf{y})$ has high probability of being close to $F_\beta(\delta)$, and its expectation is close to $F_\beta(\delta)$ as a consequence.

(a) We first show that for any $t > 0$, with probability at least $1 - 12e^{\ln(2en) - nt^4}$, we have for all δ, for all i,

$$|\tilde{p}_i(\delta) - p_i(\delta)| \leq t^2, |\tilde{p}_i(\delta)\tilde{p}'_{1|i}(\delta) - p_i(\delta)p_{1|i}(\delta)| \leq t^2 \quad (6)$$

To see this, consider a fixed i. Let $f_\delta(x) = \mathrm{I}(\mathrm{I}_\delta(x) = i)$, $\mathcal{F} = \{f_\delta : 0 \leq \delta \leq 1\}$, $g_\delta(x) = \mathrm{I}(\mathrm{I}_\delta(x) = i)P(1|x)$, and $\mathcal{G} = \{g_\delta : 0 \leq \delta \leq 1\}$. Note that the expected value and empirical average of f_δ and g_δ are $p_i(\delta)$, $\tilde{p}_i(\delta)$, $p_i(\delta)p_{1|i}(\delta)$ and $\tilde{p}_i(\delta)\tilde{p}'_{1|i}(\delta)$ respectively. In addition, both $\mathcal{F}$ and $\mathcal{G}$ have VC dimension 1. Thus, by Inequality (3.3) and (3.10) in (Vapnik, 1995), each of the following hold with probability at least $1 - 4e^{\ln(2en) - nt^4}$,

$$\forall\delta[|\tilde{p}_1(\delta) - p_1(\delta)| \leq t^2]) \quad (7)$$

$$\forall\delta[|\tilde{p}_i(\delta)\tilde{p}'_{1|i}(\delta) - p_i(\delta)p_{1|i}(\delta)| \leq t^2] \quad (8)$$

Now observing that $|\tilde{p}_1(\delta) - p_1(\delta)| \leq t^2$ implies $|\tilde{p}_0(\delta) - p_0(\delta)| \leq t^2$, and applying the union bound, then with probability at least $1 - 12e^{\ln(2en) - nt^4}$, (6) holds.

(b) Consider a fixed $\mathbf{x}$ satisfying that for some δ, for all i, $|\tilde{p}_i(\delta) - p_i(\delta)| \leq t^2$ and $|\tilde{p}_i(\delta)\tilde{p}'_{1|i}(\delta) - p_i(\delta)p_{1|i}(\delta)| \leq t^2$, we show that if $t < 1$, then with probability at least $1 - 4e^{2nt^3}$,

$$\forall i|\tilde{p}_i(\delta)\tilde{p}_{1|i}(\delta) - p_i(\delta)p_{1|i}(\delta)| \leq 5t \quad (9)$$

Consider a fixed i. If $p_i \leq 2t$, then

$$|\tilde{p}_i\tilde{p}_{1|i} - p_ip_{1|i}| \leq \tilde{p}_i\tilde{p}_{1|i} + p_ip_{1|i} \leq \tilde{p}_i + p_i \leq 5t$$

If $p_i > 2t$, then $|\tilde{p}'_{1|i} - p_{1|i}| \leq t$, [1] and we also have $\tilde{p}_i > 2t - t^2 > t$, that is $n_i > nt$. Note that $\tilde{p}_{1|i}$ is of the form $\frac{1}{n_i}\sum_{i=1}^{n_i} I_i$ where the I_i's are independent binary random variables, and the expected value of $\tilde{p}_{1|i}$ is $\tilde{p}'_{1|x}$, then applying Hoeffding's inequality, with probability at least $1 - 2e^{-2nt \cdot t^2}$, we have $|\tilde{p}_{1|i} - \tilde{p}'_{1|i}| \leq t$. When $p_i > 2t$, $|\tilde{p}_i - p_i| \leq t^2 < t$, and $|\tilde{p}_{1|i} - \tilde{p}'_{1|i}| \leq t$, we have

$$\begin{aligned}
\tilde{p}_i\tilde{p}_{1|i} - p_ip_{1|i} &\geq (p_i - t)(p_{1|i} - 2t) - p_ip_{1|i} \\
&\geq 2t^2 - 2p_it - p_{1|i}t \geq -5t \\
\tilde{p}_i\tilde{p}_{1|i} - \tilde{p}_i\tilde{p}_{1|i} &\leq (p_i + t)(p_{1|i} + 2t) - p_ip_{1|i} \\
&\leq 2p_it + p_{1|i}t + 2t^2 \leq 5t
\end{aligned}$$

[1] This can be seen by observing that if $\tilde{p}'_{1|i} - p_{1|i} > t$, then $\tilde{p}_i\tilde{p}'_{1|i} - p_ip_{1|i} \geq p_i(\tilde{p}'_{1|i} - p_{1|i}) - |\tilde{p}_i - p_i| > 2t \cdot t - t^2 = t^2$, a contradiction. Similarly, the other case can be shown to be impossible.

That is, $|\tilde{p}_i\tilde{p}_{1|i} - p_ip_{1|i}| \leq 5t$. Combining the above argument, we see that (9) holds with probability at least $1 - 4e^{2nt^3}$.

(c) If for some δ, $\mathbf{x}$ satisfies $|\tilde{p}_i - p_i| \leq t^2 < t$ and $\mathbf{x}, \mathbf{y}$ satisfies (9), then by eq. 5,

$$\begin{aligned}
F_\beta(\mathrm{I}_\delta(\mathbf{x}), \mathbf{y}) &\geq \frac{(1+\beta^2)(p_1p_{1|1} - 5t)}{\beta^2(p_1p_{1|1} + 5t + p_0p_{1|0} + 5t) + p_1 + t} \\
&\geq F_\beta(\delta) - \gamma_1 t
\end{aligned}$$

where γ_1 is some positive constant that depends on β and π_1 only. The last inequality can be seen by noting that for $a, b, d, t \geq 0$, $c > 0$, we have $\frac{a-bt}{c+dt} \geq \frac{a}{c} - \frac{ad+bc}{c^2}t$, and observing that in this case $a = (1+\beta^2)p_1p_{1|1} \leq (1+\beta^2)\pi_1$, $b = 5 + 5\beta^2$, $c = \beta^2\pi_1 + p_1 \geq \beta^2\pi_1$, and $d = 10\beta^2 + 1$.

Similarly, if $t < \frac{1}{2}\frac{\beta^2\pi_1}{10\beta^2+1}$, then

$$\begin{aligned}
F_\beta(\mathrm{I}_\delta(\mathbf{x}), \mathbf{y}) &\leq \frac{(1+\beta^2)(p_1p_{1|1} + 5t)}{\beta^2(p_1p_{1|1} - 5t + p_0p_{1|0} - 5t) + p_1 - t} \\
&\leq F_\beta(\delta) + \gamma_2 t
\end{aligned}$$

where γ_2 is some positive constant that depends on β and π_1 only. The last inequality can be seen by noting that for $a, b, d \geq 0$, $c > 0$, $c > 2dt$, we have $\frac{a+bt}{c-dt} \leq \frac{a}{c} + 2\frac{ad+bc}{c^2}t$, and observing that in this case $a = (1+\beta^2)p_1p_{1|1} \leq (1+\beta^2)\pi_1$, $b = 5 + 5\beta^2$, $c = \beta^2\pi_1 + p_1 \geq \beta^2\pi_1$, $d = 10\beta^2 + 1$, and $c > 2dt$.

Now it follows that for an $\mathbf{x}$ satisfying (6), then for any $\delta \in [0, 1]$, for any $t < \frac{1}{2}\frac{\beta^2\pi_1}{10\beta^2+1}$, with probability at least $1 - 4e^{-nt^3}$, $|F_\beta(\mathrm{I}_\delta(\mathbf{x}), \mathbf{y}) - F_\beta(\delta)| \leq \max(\gamma_1, \gamma_2)t$. Hence

$$|\mathrm{E}[F_\beta(\mathrm{I}_\delta(\mathbf{x}), \mathbf{y})] - F_\beta(\delta)| \leq 4e^{-nt^3} \cdot 1 + \max(\gamma_1, \gamma_2)t$$

For any $\epsilon > 0$, further restrict t to be the maximum satisfying $t \leq \frac{\epsilon}{2\max(\gamma_1,\gamma_2)}$, and let this value be denoted by t_0, then t_0 depends on β, ϵ (and π_1). Now the second term in the above inequality is less than $\epsilon/2$. The first term is monotonically decreasing in n and converges to 0 as $n \to \infty$. Now take $N_{\beta,\epsilon,\eta}$ to be the smallest number such that for $n = N_{\beta,\epsilon,\eta}$, the first term is less than $\epsilon/2$, and $12e^{\ln(2en) - nt^4} < \eta$, then for any $n > N_{\beta,\epsilon,\eta}$, with probability at least $1 - \eta$, $|\mathrm{E}_{\mathbf{y} \sim P(\cdot|\mathbf{x})}[F_\beta(\mathrm{I}_\delta(\mathbf{x}), \mathbf{y})] - F_\beta(\delta)| < \epsilon$. □

Theorem 8. *Let $\mathbf{s}^*(\mathbf{x}) = \max_{\mathbf{s}} \mathrm{E}[F_\beta(\mathbf{s}, \mathbf{y})]$, with $\mathbf{s}$ satisfying $\{P(1|x_i) \mid s_i = 1\} \cap \{P(1|x_i) \mid s_i = 0\} = \emptyset$. Let $t^* = \arg\max_{t \in \mathcal{T}} F_\beta(t)$. Then for any $\epsilon, \eta > 0$,*
(a) There exists $N_{\beta,\epsilon,\eta}$ such that for all $n > N_{\beta,\epsilon,\eta}$, with probability at least $1 - \eta$, $\mathrm{E}[F_\beta(t^*(\mathbf{x}), \mathbf{y})] \leq \mathrm{E}(F_\beta(\mathbf{s}^*(\mathbf{x}), \mathbf{y})) < \mathrm{E}[F_\beta(t^*(\mathbf{x}), \mathbf{y})] + \epsilon$.

(b) There exists $N_{\beta,\epsilon,\eta}$ such that for all $n > N_{\beta,\epsilon,\eta}$, with probability at least $1-\eta$, $|F_\beta(t^*(\mathbf{x}), \mathbf{y}) - F_\beta(\mathbf{s}^*(\mathbf{x}), \mathbf{y}))| < \epsilon$.

Proof for Theorem 8. (a) By Lemma 7, when $n > N_{\beta,\frac{\epsilon}{2},\eta}$, with probability at least $1-\eta$, $\mathbf{x}$ satisfies that for all δ, $|\mathrm{E}_{\mathbf{y}\sim P(\cdot|\mathbf{x})}[F_\beta(\mathrm{I}_\delta(\mathbf{x}), \mathbf{y})] - F_\beta(\delta)| < \epsilon/2$. Consider such an $\mathbf{x}$. The lower bound is clear because $\mathbf{s} = \mathrm{I}_{\delta^*}$ satisfies $\{P(1|x_i) : s_i = 1\} \cap \{P(1|x_i) : s_i = 0\} = \emptyset$. For the upper bound, by Theorem 9 and the definition of $\mathbf{s}^*(\mathbf{x})$, we have $\mathbf{s}^*(\mathbf{x}) = \mathrm{I}_{\delta'}(\mathbf{x})$ for some δ'. Thus $\mathrm{E}[F_\beta(\mathbf{s}^*(\mathbf{x}), \mathbf{y})] < F_\beta(\delta') + \epsilon/2 < F_\beta(\delta') + \epsilon/2 \le F_\beta(\delta^*) + \epsilon/2 < \mathrm{E}[F_\beta(\mathrm{I}_{\delta^*}(\mathbf{x}), \mathbf{y})] + \epsilon$.

(b) From the proof for Lemma 7, for any $t > 0$, with probability at least $1 - 12e^{\ln(2en) - nt^4}$, we have for all δ, for all i, $\mathbf{x}$ satisfies (6), that is,

$$|\tilde{p}_i(\delta) - p_i(\delta)| \le t^2, |\tilde{p}_i(\delta)\tilde{p}'_{1|i}(\delta) - p_i(\delta)p_{1|i}(\delta)| \le t^2$$

In addition, if $t < \frac{1}{2}\frac{\beta^2\pi_1}{10\beta^2+1}$, then for such $\mathbf{x}$, for any δ, with probability at least $1 - 4e^{2nt^3}$,

$$|F_\beta(\mathrm{I}_\delta(\mathbf{x}), \mathbf{y}) - F_\beta(\delta)| < \gamma t$$

where γ is a constant depending on ϵ (and π_1). Note that there exists δ' such that $\mathrm{I}_{\delta'}(\mathbf{x}) = \mathbf{s}^*(\mathbf{x})$. Using the union bound, with probability at least $1 - 8e^{-2nt^3}$,

$$\begin{aligned} |F_\beta(\mathrm{I}_{\delta'}(\mathbf{x}), \mathbf{y}) - F_\beta(\delta')| &< \gamma t \\ |F_\beta(\mathrm{I}_{\delta^*}(\mathbf{x}), \mathbf{y}) - F_\beta(\delta^*)| &< \gamma t \end{aligned} \tag{10}$$

Hence we have

$$\mathrm{E}(F_\beta(\mathrm{I}_{\delta'}(\mathbf{x}), \mathbf{y}) \le (1 - 8e^{-2nt^3})(F_\beta(\delta') + \gamma t) + 8e^{-2nt^3}$$
$$\mathrm{E}(F_\beta(\mathrm{I}_{\delta^*}(\mathbf{x}), \mathbf{y}) \ge (1 - 8e^{-2nt^3})(F_\beta(\delta^*) - \gamma t)$$

Combining the above two inequalities with $\mathrm{E}(F_\beta(\mathrm{I}_{\delta'}(\mathbf{x}), \mathbf{y}) \ge \mathrm{E}(F_\beta(\mathrm{I}_{\delta*}(\mathbf{x}), \mathbf{y})$, we have

$$F_\beta(\delta^*) - F_\beta(\delta') \le 2\gamma t + \frac{8e^{-2nt^3}}{1 - 8e^{-2nt^3}}$$

For those $\mathbf{y}$ satisfying (10), we have

$$\begin{aligned} &|F_\beta(\mathrm{I}_{\delta'}(\mathbf{x}), \mathbf{y}) - F_\beta(\mathrm{I}_{\delta^*}(\mathbf{x}), \mathbf{y})| \\ &= |F_\beta(\mathrm{I}_{\delta'}(\mathbf{x}, \mathbf{y}) - F_\beta(\delta')| + |F_\beta(\delta') - F_\beta(\delta^*)| \\ &\qquad + |F_\beta(\delta^*) - F_\beta(\mathrm{I}_{\delta^*}(\mathbf{x}), \mathbf{y})| \\ &< 4\gamma t + \frac{8e^{-2nt^3}}{1 - 8e^{-2nt^3}} \end{aligned}$$

Combining the above argument, we have with probability at least $(1 - 12e^{\ln(2en) - nt^4})(1 - 8e^{-2nt^3})$ that $|F_\beta(\mathbf{s}^*(\mathbf{x}), \mathbf{y}) - F_\beta(t^*(\mathbf{x}), \mathbf{y})| < 4\gamma t + \frac{8e^{-2nt^3}}{1-8e^{-2nt^3}}$.

Now choose $t = \frac{\epsilon}{8\gamma}$, then for sufficiently large n, we can guarantee that with probability at least $1 - \eta$, $|F_\beta(\mathbf{s}^*(\mathbf{x}), \mathbf{y}) - F_\beta(t^*(\mathbf{x}), \mathbf{y})| < \epsilon$. □

Theorem 9. *(Probability Ranking Principle for F-measure, Lewis 1995) Suppose $\mathbf{s}^*$ is a maximizer of $\mathrm{E}(F_\beta(\mathbf{s}, \mathbf{y}))$. Then $\min\{p_i \mid s^*_i = 1\}$ is not less than $\max\{p_i \mid s^*_i = 0\}$.*

References

Lewis, D.D. Evaluating and optimizing autonomous text classification systems. In *SIGIR*, pp. 246–254, 1995.

Vapnik, V.N. *The nature of statistical learning theory.* Springer, 1995.