Worksheet 3: Binary Classification

Name:Aaron Zoll

Due September 27, 2022

Recall for model $\tilde{y} : \mathcal{X} \to \mathbb{R}$, we obtain prediction model with thresholding: for $t \in \mathbb{R}$, we define $\bar{y}_t := \mathbb{1}_{\tilde{y}} \geq t$. Typically, $\tilde{y}(\mathbb{R}) \subset [0,1]$, because we may like to interpret $\tilde{y}(x)$ as (something like) $\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y = 1 \mid x)$, but it isn't strictly necessary.

1. Consider joint probability space $\mathcal{X} \times \mathcal{Y} = \mathbb{R} \times \{0,1\}$ with measure

$$\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}((a,b) \times \{j\}) = \alpha_j \cdot \int_a^b f_j(x)dx,$$

where $\alpha_0 + \alpha_1 = 1$, both nonnegative, and $f_j(x) = \gamma_j e^{-\gamma_j t} \cdot \mathbb{1}_{t \geq 0}$ with $\gamma_j > 0$. Express both marginals $\mathbb{P}_{\mathcal{X}}((a,b))$ and $\mathbb{P}_{\mathcal{Y}}(y = j)$. You may use (properties of) this density for problems 2-4 as well.

Solution: (assume $b > a \geq 0$ and later $t > 0$)

$$\begin{aligned}\mathbb{P}_X((a,b)) &= \mathbb{P}_{\mathcal{X}\times\mathcal{Y}}((a,b) \times \{0,1\}) \\ &= \alpha_0 \int_a^b f_0(x)dx + \alpha_1 \int_a^b f_1(x)dx\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}_y(y = j) &= \mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(\mathbb{R} \times j) \\ &= \alpha_j \int_0^\infty f_j(x)dx \\ &= \alpha_j \int_0^\infty \gamma_j e^{-\gamma_j t} dx \\ &= \alpha_j[-e^{-\gamma_j t}]\Big|_0^\infty \\ &= \alpha_j\end{aligned}$$

2. Show that $\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y = 1 \mid x) = \frac{\alpha_1 f_1(x)}{\alpha_0 f_0(x) + \alpha_1 f_1(x)}$. You may wish to recall the Fundamental Theorem of Calculus, and use continuity of measure (which you may suppose without proof for both arguments of $\mathbb{P}(\cdot \mid \cdot)$).

Solution: Fix $a \in \mathbb{R}$ and note that

$$\begin{aligned}\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y = 1|(a,x)) &= \frac{\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}((a,x) \times \{1\})}{\mathbb{P}_{\mathcal{X}}((a,x))} \\ &= \frac{a_1 \int_a^x f_1(x)dx}{\alpha_0 \int_a^x f_0(x)dx + \alpha_1 \int_a^x f_1(x)dx}\end{aligned}$$

As we take $a \to x$, by using FTC and continuity, we immediately get the result, but more explicitly

$$\begin{aligned}\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y = 1|x) &= \lim_{a \to x} \mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y = 1|(a,x)) \\ &= \mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y = 1| \lim_{a \to x}(a,x)) \\ &= \lim_{a \to x} \frac{a_1 \int_a^x f_1(x)dx}{\alpha_0 \int_a^x f_0(x)dx + \alpha_1 \int_a^x f_1(x)dx} \\ &= \frac{a_1 f_1(x)}{a_0 f_0(x) + a_1 f_1(x)} \text{ by L'Hospital rule + FTC}\end{aligned}$$

3. We may treat input data $x \in \mathcal{X} = \mathbb{R}$ as a score itself. For threshold predictor $\bar{y}_t : \mathcal{X} \to \mathcal{Y}$ defined by $\bar{y}_t(x) := \mathbb{1}_{x \geq t}$, express the true positive rate, false positive rate, and precision of $\bar{y}_t$ defined as

$$\text{tpr(t)} := \mathbb{P}_{\mathcal{X}|\mathcal{Y}}\left(\bar{y}_t = 1 \mid y = 1\right), fpr(t) := \mathbb{P}_{\mathcal{X}|\mathcal{Y}}\left(\bar{y}_t = 1 \mid y = 0\right) \text{ and } \text{prec}(t) := \mathbb{P}_{\mathcal{Y}|\mathcal{X}}\left(y = 1 \mid \bar{y}_t = 1\right).$$

Solution:

$$\begin{aligned} tpr(t) = \mathbb{P}_{\mathcal{X}|\mathcal{Y}}(\bar{y}_t = 1|y = 1) &= \frac{\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(\{\bar{y}_t = 1\} \times \{y = 1\})}{\mathbb{P}_{\mathcal{Y}}(y = 1)} \\ &= \frac{\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(\{x \geq t\} \times \{y = 1\})}{\mathbb{P}_{\mathcal{Y}}(y = 1)} \\ &= \frac{a_1 \int_t^\infty f_1(x)dx}{a_1} \\ &= e^{-\gamma_1 t} \end{aligned}$$

Similarly,

$$\begin{aligned} fpr(t) = \mathbb{P}_{\mathcal{X}|\mathcal{Y}}(\bar{y}_t = 1|y = 0) &= \frac{\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(\{\bar{y}_t = 1\} \times \{y = 0\})}{\mathbb{P}_{\mathcal{Y}}(y = 0)} \\ &= \frac{\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(\{x \geq t\} \times \{y = 0\})}{\mathbb{P}_{\mathcal{Y}}(y = 0)} \\ &= \frac{a_0 \int_t^\infty f_0(x)dx}{a_0} \\ &= e^{-\gamma_0 t} \end{aligned}$$

and

$$\begin{aligned} prec(t) = \mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y = 1|\bar{y}_t = 1) &= \frac{\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(\{\bar{y}_t = 1\} \times \{y = 1\})}{\mathbb{P}_X(\bar{y}_t = 1)} \\ &= \frac{a_1 \int_t^\infty f_1(x)dx}{a_0 \int_t^\infty f_0(x) + a_1 \int_t^\infty f_1(x)dx} \\ &= \frac{a_1 e^{-\gamma_1 t}}{a_0 e^{-\gamma_0 t} + a_1 e^{-\gamma_1 t}} \end{aligned}$$

4. It is typically impossible to maximally satisfy all desired objectives. For example, optimal tpr = 1may be realized for minimal threshold at the expense of inducing undesirable fpr = 1 (the other extreme realizes fpr = 0 at the expense of tpr = 0). Suppose you are given objective function

$$f(t) := \lambda \operatorname{tpr}(t) + (1-\lambda) \operatorname{prec}(t)$$

Explain how you would solve for $t^* = \arg\max_{t\in\mathcal{X}} f(t)$, and compute the steps-as much as you can-to do so. Simplify expressions as much as possible and use 1st order Taylor expansion of exp to obtain candidate approximation for t^*. Interpret whether this is a good approximation and justfy why if so/hypothesize why not if not.

Solution:

$f'(t) = \lambda tpr'(t) + (1-\lambda) prec'(t)$

$$tpr'(t) = -\gamma_1 e^{-\gamma_1 t} \tag{1}$$

$$\begin{aligned} prec'(t) &= \frac{[-\gamma_1\alpha_1 e^{-\gamma_1 t}][\alpha_0 e^{-\gamma_0 t} + \alpha_1 e^{-\gamma_1 t}] - \alpha_1 e^{-\gamma_1 t}[-\gamma_0\alpha_0 e^{-\gamma_0 t} - \gamma_1\alpha_1 e^{-\gamma_1 t}]}{\alpha_0^2 e^{-2\gamma_0 t} + 2\alpha_0\alpha_1 e^{-(\gamma_0+\gamma_1)t} + \alpha_1^2 e^{-2\gamma 1 t}} & (2) \\ &= \frac{(\gamma_0-\gamma_1)\alpha_0\alpha_1 e^{-(\gamma_0+\gamma_1)t}}{\alpha_0^2 e^{-2\gamma_0 t} + 2\alpha_0\alpha_1 e^{-(\gamma_0+\gamma_1)t} + \alpha_1^2 e^{-2\gamma 1 t}} & (3) \end{aligned}$$

Since obtaining $t^* \in \mathcal{X} = \mathbb{R}$, we just need to solve when the derivative is equal to 0, so the denominator doesn't matter all too much (especially because we can ensure it is strictly positive.)
We then get that the numerator for the entirety of $f'(t)$ is

$$-\lambda\left[\gamma_1\alpha_0^2 e^{-(2\gamma_0+\gamma_1)t} + 2\gamma_1\alpha_0\alpha_1 e^{-(\gamma_0+2\gamma_1)t} + \gamma_1\alpha_1^2 e^{-3\gamma_1 t}\right] + (1-\lambda)(\gamma_0-\gamma_1)\alpha_0\alpha_1 e^{-(\gamma_0+\gamma_1)t} \tag{4}$$

Since we are solving this equal to 0, we can divide by $e^{-\gamma_1 t}$ to get the equation (after rearranging and factoring a bit)

$$(1-\lambda)(\gamma_0-\gamma_1)\alpha_0\alpha_1 e^{-\gamma_0 t} = \lambda\gamma_1(\alpha_0 e^{-\gamma_0 t} + \alpha_1 e^{-\gamma_1 t})^2 \tag{5}$$

We can then linearize to get a fairly simple quadratic (and the fact that $\alpha_0 + \alpha_1 = 1$):

$$(1-\lambda)(\gamma_0-\gamma_1)\alpha_0\alpha_1(1-\gamma_0 t) = \lambda\gamma_1(1-(\alpha_0\gamma_0+\alpha_1\gamma_1)t)^2 \tag{6}$$

Expanding this out and doing quadratic formula (please don't make me do this) is going to give us an estimate on t^*. However, this can be pretty bad for two reasons. One, this could give us the argmin, not the argmax as this is just looking for a critical point. Furthermore, while linear approximations may be decent, we are then squaring, and we already modified this a decent amount by removing the daa of the denominator. So while this may be a decent search for a zero, it may give us the "wrong" one, and we may get shifted farther away when looking back at the original function.

5. Define equal error rate (eer) as false positive rate eer $:= \mathbb{P}_{\mathcal{X}|\mathcal{Y}}\left(\bar{y}_{t^e} = 1 \mid y = 0\right)$ at

$$\mathrm{t^e} := \arg\min_{\mathrm{t}\in\tilde{\mathfrak{y}}(\mathbb{R})} |1 - \mathrm{fpr(t)} - \mathrm{tpr(t)}|.$$

Express accuracy $\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}\left(\bar{\mathrm{y}}_{\mathrm{t^e}} = \mathrm{y}\right)$ at eer in terms of only $\mathrm{tpr}\left(\mathrm{t^e}\right)$ and $\mathrm{fpr}\left(\mathrm{t^e}\right)$. Simplify as much as you can.*
Solution:

$$\begin{aligned}
\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(\bar{y}_{t^e} = y) &= \int_{\bar{y}_{t^e}=y} d\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(x,y) \\
&= \int_{\mathcal{Y}}\int_{\bar{y}_{t^e}=y|y} dp_{\mathcal{X}|\mathcal{Y}}(x|y)d\mathbb{P}_{\mathcal{Y}}(y) \\
&= \int_{\mathcal{Y}} \mathbb{P}_{X|Y}(\bar{y}_{t^e} = y|y)d\mathbb{P}_{\mathcal{Y}}(y) \\
&= \mathbb{P}_Y(y=0)\mathbb{P}_{X|Y}(\bar{y}_{t^e} = 0|y=0) + \mathbb{P}_Y(y=1)\mathbb{P}_{X|Y}(\bar{y}_{t^e} = 1|y=1)
\end{aligned}$$

Let $p = \mathbb{P}_Y(y=1)$, then

$$= (1-p)(1 - fpr(t^e)) + p(tpr(t^e))$$

Then because we have simultaneous limits of $\pm\infty$ on the image of $fpr(t)$ *and* $tpr(t)$ we know that the min $|1 - fpr(t) - tpr(t)| = 0$ and so at t^e we have that $1 - fpr(t^e) = tpr(t^e)$
Thus, final expression is

$$\begin{aligned}
&(1-p)tpr(t^e) + ptpr(t^e) && (7) \\
&= tpr(t^e) && (8)
\end{aligned}$$

6. For loss function $\ell_{\tilde{y}}(x,y) := -\log\left(\tilde{y}(x)^y(1-\tilde{y}(x))^{(1-y)}\right)$, show that the optimal model $y^* : \mathcal{X} \to [0,1]$ is calibrated.
   Solution: First note that $\ell_{\tilde{(y)}}(x,y) = -[y\log(\tilde{y}(x)) + (1-y)\log(1-\tilde{y}(x))]$

$$\begin{aligned}
\mathbb{E}(\ell_{\tilde{y}}) &= \int_{\mathcal{X}\times\mathcal{Y}} -[y\log(\tilde{y}(x)) + (1-y)\log(1-\tilde{y}(x))]d\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(x,y) \\
&= \int_{\mathcal{X}}\int_{\mathcal{Y}|\mathcal{X}} -[y\log(\tilde{y}(x)) + (1-y)\log(1-\tilde{y}(x))]d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x)d\mathbb{P}_{\mathcal{Y}}(y)
\end{aligned}$$

Going from global to local opti, we now look at

$$\begin{aligned}
&\int_{\mathcal{Y}|\mathcal{X}} -[y\log(\tilde{y}(x) + (1-y)\log(1-\tilde{y}(x))]d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x) \\
&= -\mathbb{P}_Y(y=0|x)\log(1-\tilde{y}(x)) - \mathbb{P}_Y(y=1|x)\log(\tilde{y}(x)) \\
&= -(1-p(x))\log(1-\tilde{y}(x)) - p(x)\log(\tilde{y}(x))
\end{aligned}$$

Differentiating w.r.t. $(\tilde{y}(x))$ and setting equal to 0

$$\begin{aligned}
0 &= \frac{1-p(x)}{1-\tilde{y}(x)} - \frac{p(x)}{\tilde{y}(x)} \\
&= \frac{\tilde{y}(x) - p(x)\tilde{y}(x) - p(x) + p(x)\tilde{y}(x)}{(1-\tilde{y}(x))\tilde{y}(x)}
\end{aligned}$$

Thus, as long as y isn't constant, we get $\tilde{y}(x) = p(x)$, as desired

7. Suppose model score $\tilde{y} : \mathcal{X} \to [0,1]$ has the relation $\tilde{y}(p) = \sigma(p) = \frac{1}{1+e^{-c(2p-1)}}$ for $p(x) := \mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y = 1 \mid x)$ and c > 0. Define calibrated model $\bar{y} : \mathcal{X} \to [0,1]$ in terms of $\tilde{y}$, i.e. map $\varphi : [0,1] \to [0,1]$ for which $\bar{y} := \varphi \circ \tilde{y} = p$.
Solution: Set $\phi(x) = \frac{1}{2}(\frac{-1}{c}\log(\frac{1}{x} - 1) + 1)$ and we get that

$$\begin{aligned}
\bar{y}(t) &:= \phi \circ \tilde{y} \\
&= \phi(\frac{1}{1+e^{-c(2p-1)}}) \\
&= \frac{1}{2}(\frac{-1}{c}\log(\frac{1}{\frac{1}{1+e^{-c(2p-1)}}} - 1)) + 1) \\
&= \frac{1}{2}(\frac{-1}{c}\log(1 + e^{-c(2p-1)} - 1)) + 1) \\
&= \frac{1}{2}(\frac{-1}{c}\log(e^{-c(2p-1)})) + 1) \\
&= \frac{1}{2}(\frac{-1}{c}(-c(2p-1)) + 1) \\
&= \frac{1}{2}(2p - 1 + 1) \\
&= p
\end{aligned}$$