Lec 21 7/30/18 Math 390.3

The model $y = t(\vec{x})$

But there's stuff you don't know, so it "appears" as:

$$Y \sim \text{Bern}(f_{pr}(\vec{x}))$$

i.e. Given the same $\vec{x}$, sometimes $y=1$ and sometimes $y=0$.

So if we care about $P(Y=1|X)$ our target is $f_{pr}(\vec{x})$.

But $f_{pr}(\vec{x})$ can be arbitrarily complicated so we try to approximate it using a function in class $\mathcal{H}$. We picked

"$\mathcal{H}_{pr}$" $\mathcal{H} = \left\{ \frac{e^{\vec{w}\cdot\vec{x}}}{1+e^{\vec{w}\cdot\vec{x}}} : \vec{w} \in \mathbb{R}^{p+1} \right\}$

Our A was to maximize likelihood assuming each Y_i ind. and we found:

$$\vec{b} = \underset{\vec{w}}{\text{argmin}} \left\{ \sum_{i=1}^{n} \ln\left(1+e^{-z_i \vec{w}\cdot\vec{x}_i}\right) \right\} \quad \text{where } z_i = 2y_i - 1$$

convenience

looks like $\sum e_i$ ← my error

where $e_i = \ln\left(1+e^{-z_i\vec{w}\cdot\vec{x}_i}\right) > 0$ ← why?

why does this make sense?

if $y_i = 1$ $\quad e_i = \ln\left(1+e^{-\vec{w}\cdot\vec{x}_i}\right)$

if $y_i = 0$ $\quad e_i = \ln\left(1+e^{\vec{w}\cdot\vec{x}_i}\right)$

if $\vec{w}\cdot\vec{x}_i$ is large positive ⇒ error small

if ... negative large ⇒ error small

$\frac{d}{d\vec{w}}\left[\sum e_i\right] \overset{\text{set}}{=} 0$ does not have my analytic sol.

~~Bad no analytic solution exists...~~ must use numerical methods just like the SVM algorithm (which min. sum hinge loss)

How to use?

Okay ... $\hat{p}^* := g(\vec{x}^*) = \left(1+e^{-\vec{b}\cdot\vec{x}^*}\right)^{-1} = \hat{P}(Y_i=1 \mid \vec{X}_i)$

prob estimate for new case, $\vec{x}^*$

The lin. term $\vec{b}\cdot\vec{x}$ is buried inside. Does it have meaning?

$$\Rightarrow 1-\hat{p} = \left(1+e^{\vec{b}\cdot\vec{x}^*}\right)^{-1} = P(Y_i=0 \mid \vec{X}_i)$$

$$\Rightarrow \underbrace{\frac{\hat{p}}{1-\hat{p}}}_{\hat{Odds}(Y=1|x)} = \frac{1+e^{\vec{b}\cdot\vec{x}}}{1+e^{-\vec{b}\cdot\vec{x}}} \cdot \frac{e^{\vec{b}\cdot\vec{x}}}{e^{\vec{b}\cdot\vec{x}}} = e^{\vec{b}\cdot\vec{x}} \frac{\left(1+e^{\vec{b}\cdot\vec{x}}\right)}{\left(1+e^{\vec{b}\cdot\vec{x}}\right)}$$

Interpretation? With increase in X_1 means log-odds of $Y=1$ by b_1!

$$\Rightarrow \ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \vec{b}\cdot\vec{x} = b_0 + b_1 x_1 + \ldots + b_p x_p$$

DEMO

"log-Odds$(Y|\vec{x})$" if you know how to read them, that's cool

very negative $\Rightarrow$ low prob.
very positive $\Rightarrow$ high prob.
near zero $\Rightarrow$ 50% prob.

Validating prob. models. What is the best prob. model you can create? f_{pr}. So we should validate against f_{pr}. Can we? No! We have to validate $\hat{p}$ vs. y!

This is very awkward! Enter the theory of "scoring functions" or "scoring rules".

Let $S(\hat{p}_i, y_i)$ be the scoring rule for obs. i

A "proper scoring rule" has the following property

$$\forall i \quad f_{pr}(\vec{x}_i) = \underset{\hat{p}}{\operatorname{argmax}} \left\{ S(\hat{p}_i, y_i) \right\}$$

the expected scoring rule is maximized if you use the true prob.

Popular scoring rules are the "log scoring rule":

$$S_i = y_i \ln(\hat{p}_i) + (1-y_i)\ln(1-\hat{p}_i)$$

evaluated

$$\Rightarrow \text{avg. score} = \frac{1}{n}\sum_{i=1}^{n} S_i$$

If $y_i = 1$, this is when $\hat{p} \approx 1$

If $y_i = 0$, " " " $\hat{p} \approx 0$

- Brier score (1950)

$$S_i = -(y_i - \hat{p}_i)^2$$

Does the avg. score mean anything in the problem? No... just as a means to validate and compare models.

To validate, you need to think in Brier scores e.g. Very opaque!

What's another way to validate? Use prob. est. to do classification! Then validate the classification!

How to turn a $\hat{p}$ into a $\hat{y}$? ⇒ Use the probability model to do binary classification

If 4 is predicted to be 1 w.p. 90% ⇒ $\hat{y}=1$

" " " " " " " " " w.p 10% ⇒ $\hat{y}=0$

" " " " " " " " " w.p 51% ⇒ $\hat{y}=1$

What are you doing?

$$\hat{y}_i := \mathbb{1}_{\hat{p}_i \geq 0.5} \quad \text{(HARD threshold rule)}$$

What if you do

$$\hat{y}_i := \mathbb{1}_{\hat{p}_i \geq 0.9} \; ?$$

this makes it very difficult to get $\hat{y}_i = 1$!

This means you are being very conservative to predict $\hat{y}_i = 1$.

What happens to the confusion table? FP↓ & FN↑

You are saying $\text{cost}_{FP} > \text{cost}_{FN}$! "Asymmetric cost classifiers/models"!

Generally

$$\hat{y}_i := \mathbb{1}_{\hat{p}_i \geq p_{th}}$$ ← the threshold

One prob model ⇒ Infinite classification models

WEIRD

Obviously $p_{th} \in (0,1)$! Each value gives a new classifier!!

Why not look at all of them or nearly all of them.

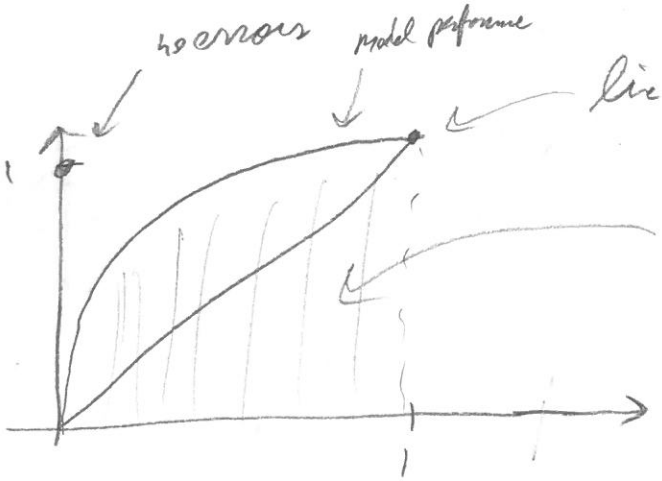
ρ_{th}	TP	TN	FP	FN	Precision	Recall	FOR	FPR
0.01								
0.02								
0.03 ⇒ pick $\Delta\rho_{th}$								
⋮								
0.99								

all can be derived

curve

Now let's plot performance of all possible models

Receiver Operating Characteristic Curve (ROC)



If $> 0.5 \Rightarrow$ majority of models perform better than chance.

False Positive Rate (FPR) ← another metric we didn't discuss yet

Before we go further... let's consider the case

If you use the null models, you classify positives w/ prob. ρ_{th}

because the null models can be $Y_n \sim \text{Bern}(\rho_{th})$ where $\rho_{th} \in (0,1)$.

If ρ_Y is $P(Y=1)$ the marginal base rate, then the following is the confusion table:

true (Y) \ predicted (Ŷ)	0	1	
0	$(1-\rho_{th})(1-\rho_Y)$ TN	$\rho_{th}(1-\rho_Y)$ FP	$1-\rho_Y$ #N
1	$(1-\rho_{th})\rho_Y$ FN	$\rho_{th}\rho_Y$ TP	ρ_Y #P
	#PN $1-\rho_{th}$	#PP ρ_{th}	1

multiply by h

$$FPR = \frac{\rho_{th}(1-\rho_Y)h}{(1-\rho_Y)} = \rho_{th} \quad \left(= \frac{FP}{\#N}\right)$$

$$\text{recall} = \frac{\rho_{th}\rho_Y h}{\rho_Y h} = \rho_{th} \quad \left(= \frac{TP}{\#P}\right)$$