confusion table

$\hat{p} \rightarrow \hat{y} \qquad \hat{y} = \mathbb{1}_{\hat{p} \geq p_{th}}$

		$\hat{y}$		
		0	1	
y	0	TN	FP	N
	1	FN	TP	P
		PN	PP	n

these three quantities are static i.e. not a function of your model

Sensitivity = Recall := TP / P $\in [0,1]$

Specificity := TN / N $\in [0,1]$

tradeoff

False Positive Rate (FPR) := FP / N = 1 - Specificity

Consider a "dumb" probability estimation model g_pr(xvec) that produces random guesses from U(0,1). $\hat{y} = \mathbb{1}_{g_{pr}(\vec{x}) \geq p_{th}}$

What does the confusion table for this dumb model look like?

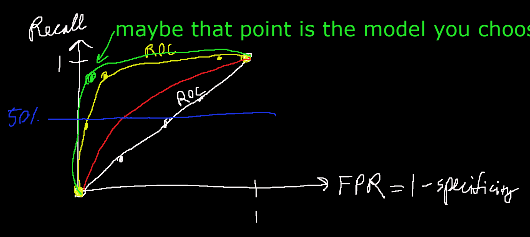
$\hat{y}=0$ | $\hat{y}=1$; 0, p_{th}, 1

		$\hat{y}$		
		0	1	
y	0	$TN = p_{th} N$	$FP = (1-p_{th}) N$	N
	1	$FN = p_{th} P$	$TP = (1-p_{th}) P$	P
		$PN = p_{th} n$	$PP = (1-p_{th}) n$	n

$$Recall = \frac{TP}{P} = \frac{(1-p_{th})P}{P} = 1 - p_{th}$$

$$FPR = \frac{FP}{N} = \frac{(1-p_{th})N}{N} = 1 - p_{th}$$

Let's plot recall vs FPR for all values of p_th in [0, 1]:



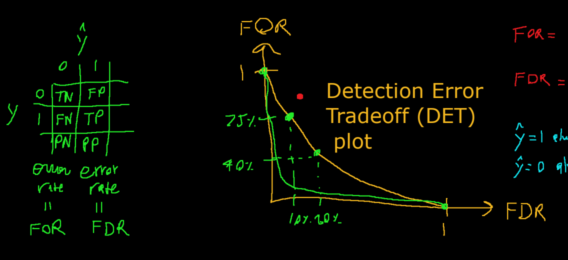
In the dumb model, y is independent of p-hat. In a smarter model, then g_pr is connected to y thus you can have high recall with high specificity. If
g_pr = yhat = 1 always => recall = 1 but specificity = 0 and
g_pr = yhat = 0 always => specificity = 1 but recall = 0.
Tracing out the performance (recall and specificity) for all p_th in [0,1] gives you the yellow line which is named the "receiver-operator curve" (ROC). The ROC can be used to compare probability estimation models by calculating the area under the curve (AUC) also called the "c-statistic":

$$AUC = \int_0^1 ROC \, dFPR \in [0,1]$$

An AUC > 1/2 has predictive power. AUC's closer and closer to 1 indicate better and better models. AUC is kind of a scoring rule (e.g. Brier).

AUC is a metric that gauges the overall fit of a probability estimation model. It is not a metric that gauges performance of a classification model (i.e. one of the point on the ROC curve).

My opinion is that ROC (and AUC) is silly in a prediction context since when you use your model to predict the future, you don't know what the y's are nor P and N. What's critical is your prediction errors (in the columns: FDR vs FOR).



$$FOR = \frac{FN}{PN} = \frac{\cancel{p_{th}} P}{\cancel{p_{th}} n} = p_y, \text{ const}$$

$$FDR = \frac{FP}{PP} = \frac{(1-p_{th})N}{(1-p_{th})n} = 1 - p_y, \text{ const}$$

$\hat{y} = 1$ always $\Rightarrow FDR = 1 - p_y$

$\hat{y} = 0$ always $\Rightarrow FOR = p_y$

Thus, there's no nice "reference line"

The DET is traced out by variying p_th in [0,1].

Once again, your classification model will be one point on the DET curve (just like it's one point on the ROC curve) but I can visually see the tradeoff of the two errors that are critical to prediction.

Bias-Variance Tradeoff in Regression Modeling, $y \in \mathbb{R}$

$y = f(\vec{x}) + \delta$, and xvec is a constant

Let's assume (I) δ is a realization from Δ a r.v. which is mean-independent from xvec and has expectation 0 i.e.

$$E[\Delta \mid \vec{x}] = 0 \Longrightarrow E[\Delta] = 0$$

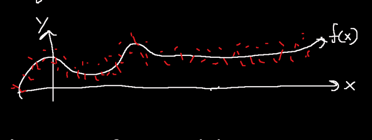
Makes sense since we defined f to be a function that extracts all useful information from x regarding y. This means y is also a realization from r.v. Y:

$$Y = f(\vec{x}) + \Delta \Rightarrow E[Y \mid X = \vec{x}] = E[f(\vec{x}) + \Delta \mid X = \vec{x}] = f(\vec{x}) + E[\Delta \mid X = \vec{x}] = f(\vec{x})$$

In textbooks, f is called the "conditional expectation function" (CEF).

Let's assume (II) homoskedasticity i.e. constant variance.

$$\sigma^2 = Var[\Delta \mid X = \vec{x}] = E[\Delta^2 \mid X = \vec{x}] - E[\Delta \mid X = \vec{x}]^2 = E[\Delta^2 \mid X = \vec{x}]$$



Let's say we fit a model using training data D to get model g:

$$y = g + e = g + (f - g) + \delta \Rightarrow e = f - g + \delta$$

$e = y - g$

$$\overset{I,II}{\Longrightarrow} Y = g + (f - g) + \Delta \Rightarrow E = f - g + \Delta$$

Let's say we now predict for the next observation x_*.

$$Y_* = g(x_*) + (f(x_*) - g(x_*)) + \Delta_* \Rightarrow E_* = f - g + \Delta_*$$

$$Bias(\vec{x}_*) := E[Y_* - g(\vec{x}_*) \mid \vec{x}_*] = E[E_* \mid \vec{x}_*] = E[f - g + \Delta_* \mid \vec{x}_*]$$

$$= f - g + E[\Delta_* \mid \vec{x}_*] \overset{I}{=} f(\vec{x}_*) - g(\vec{x}_*)$$

$$MSE(\vec{x}_*) = E[(Y_* - g(\vec{x}_*))^2 \mid \vec{x}_*] = E[(Y_* - f(\vec{x}_*))^2 \mid \vec{x}_*] = E[\Delta_*^2 \mid \vec{x}_*] = \sigma^2$$

mean squared error

let $g(\vec{x}_*) = f(\vec{x}_*)$ i.e. no bias

If our model is "perfect" i.e. no misspecification nor estimation error, then the MSE is σ^2, the irreducible squared error due to ignorance. If our model is not perfect, $g(\vec{x}) \neq f(\vec{x})$ then we show MSE >= σ^2

$g(\vec{x}) \neq f(\vec{x})$

$Y_* = f(x_*) + \Delta_*$

$$= E[Y_*^2 - 2Y_* g(\vec{x}_*) + g(\vec{x})^2 \mid \vec{x}_*] = E[(f + \Delta_*)^2] - 2g(\vec{x}_*) E[f + \Delta_*] + g(\vec{x})^2$$

$$= f^2 + 2f \cancelto{0}{E[\Delta_*]} + \cancelto{\sigma^2}{E[\Delta_*^2]} \; 2gf + g^2 = \sigma^2 + f^2 - 2gf + g^2$$

$$= \sigma^2 + (f - g)^2 = \sigma^2 + Bias(\vec{x}_*)^2 > \sigma^2$$

This calculation above assumed a fixed dataset D with fixed observations {<xvec_1,y_1>, ... <xvec_n,y_n>} and one fixed xstar. Now, let's consider random dataset D's. In these random datasets, Y_1, ..., Y_n are rvs (but xvec_1,..., xvec_n are still constants) due to Δ_1, ..., Δ_n being rvs (i.e. δ_1, ..., δ_n are realizations creating the random y_1, ..., y_n).