Recall

- $(\underline{x}_1, y_1), \dots, (\underline{x}_n, y_n)$ with $y_i \in \{0,1\}$.

- Logistic regression:

Model: $$P(Y=1 \mid \underline{x}) = \frac{e^{\beta_0 + \underline{\beta}^T \underline{x}}}{1+e^{\beta_0+\underline{\beta}^T\underline{x}}} \in (0,1)$$

- Estimate $\beta_0, \underline{\beta}$ via maximum likelihood, $\hat{\beta}_0, \hat{\underline{\beta}}$.

- For a new feature $\underline{x}_0$ we predict:

$$\hat{y} = \begin{cases} 1 & \text{if } \frac{e^{\hat{\beta}_0+\hat{\underline{\beta}}^T\underline{x}_0}}{1+e^{\hat{\beta}_0+\hat{\underline{\beta}}^T\underline{x}_0}} > \frac{1}{2} \\ 0 & \text{if } \underbrace{\frac{e^{\hat{\beta}_0+\hat{\underline{\beta}}^T\underline{x}}}{1+e^{\hat{\beta}_0+\hat{\underline{\beta}}^T\underline{x}}}}_{\hat{P}(Y=1 \mid \underline{x}_0)} < \frac{1}{2} \end{cases} \iff \begin{cases} 1 & \text{if } \hat{\beta}_0+\hat{\underline{\beta}}^T\underline{x}_0 > 0 \\ 0 & \text{if } \hat{\beta}_0+\hat{\underline{\beta}}^T\underline{x}_0 < 0 \end{cases}$$

$\hat{\beta}_0 + \hat{\underline{\beta}}^T \underline{x}_0 = 0$ decision boundary

Assessing Quality of Model Fit

Given a set of data $y_1, \dots, y_n$ and predictions $\hat{y}_1, \dots, \hat{y}_n$, how can we assess quality of the fit?

Recall y_i & $\hat{y}_i \in \{0,1\}$, think of 1 = email is spam, 0 = email is okay.

- <u>Error rate</u> $= \frac{1}{n} \sum_{i=1}^{n} \underbrace{\mathbb{1}[y_i \neq \hat{y}_i]}_{\text{indicator}}$ = percent misspecified

$$= \begin{cases} 1 & y_i \neq \hat{y}_i \\ 0 & y_i = \hat{y}_i \end{cases}$$

- Confusion matrix

	True 1s	True 0s
predicted 1s	10	4
predicted 0s	2	12

10 = predicted 10 spams that were indeed spam

predicted 4 spam that were not spam

- sensitivity = % of true positives $(10/12)$
- specificity = % of true negatives $(12/16)$
- positive predictive value = % of predicted 1s correct $(10/14)$
- negative " " = % of predicted 0s correct $(12/14)$

- error rate $= \frac{4+2}{28}$

- False positive rate (FPR) = 1-specificity ($4/16$)

- True positive rate (TPR) = sensitivity ($10/12$)

$\boxed{\text{Note}}$

- FPR = good emails that go to spam

- TPR = spam " " " " "

$\boxed{\text{Goal}}$ Try to minimize FPR and simultaneously maximize TPR

$\boxed{\text{Note}}$ As a baseline comparison, compare against random guessing

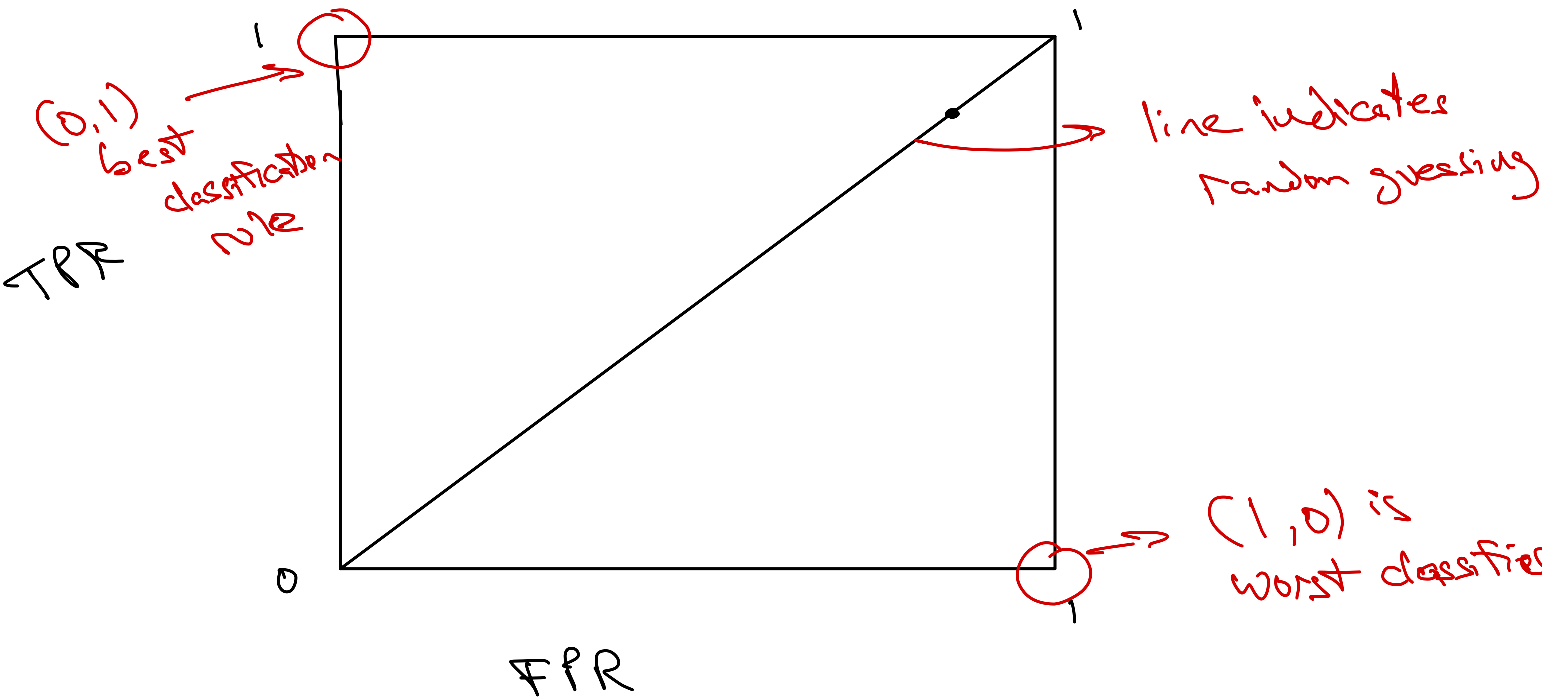
$\boxed{\text{Ex}}$ Dataset w/ 100 positives + 1000 negatives

Guess 1 with 90% probability

	true 1s	true 0s
pred 1s	90	900
pred 0s	10	100

- $FPR = \frac{900}{1000} = 90\%$
- $TPR = \frac{90}{100} = 90\%$

DEF A receives operating characteristic (ROC) graph plots the TPR (y-axis) against FPR (x-axis)



For any given model + classification/decision rule, we get a single point in ROC space.

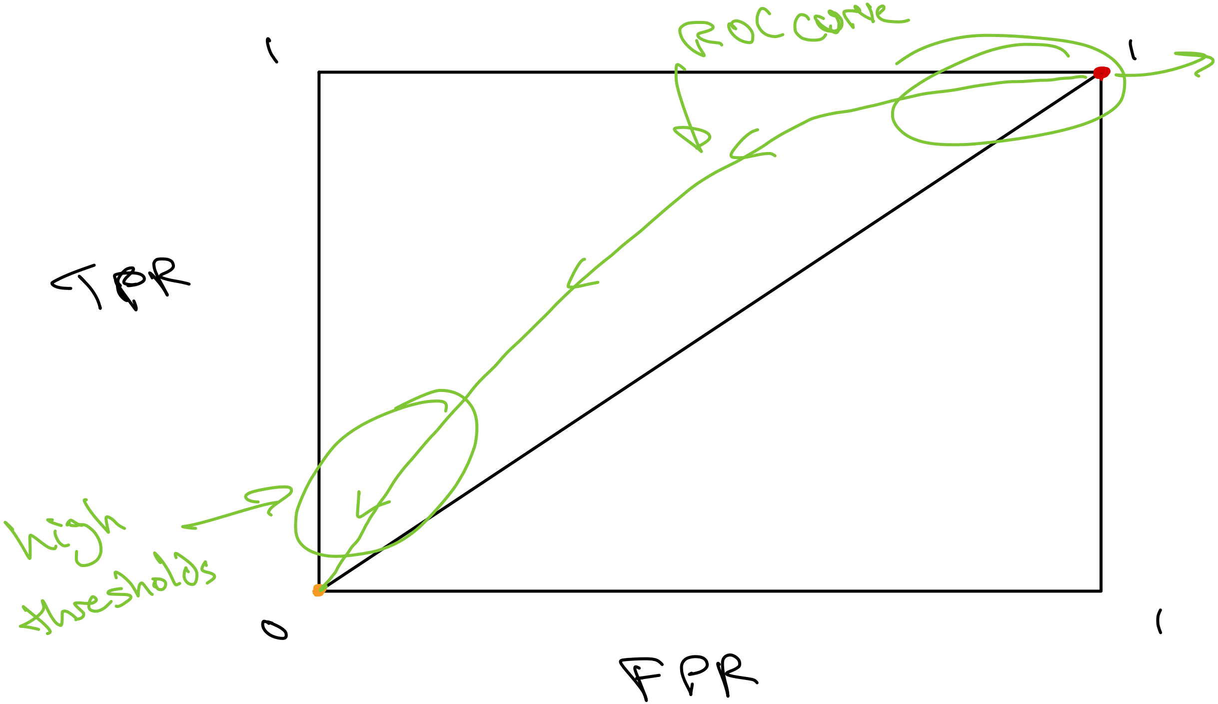
$\boxed{\text{Note}}$

Usual classification rule is $\hat{y} = 1$ if $\hat{p} > \frac{1}{2}$.

What if we vary $\frac{1}{2}$?

Ex: $\hat{y} = 1$ if $\hat{p} > \underline{0.75}$

As we vary the classification threshold from 0 to 1, we sweep out a set of points in ROC space that is called a <u>ROC curve</u>.



- (FPR, TPR) for thres of 0 = (1, 1)
- (FPR, TPR) for thres of 1 = (0, 0)

$\boxed{\text{Note}}$ One numeric summary of quality of a model fit is the <u>area under curve (AUC)</u> $\in [0, 1]$

More goodness of fit diag.

In regression we used RSS as a goodness-of-fit measure, in logistic regression models the equivalent idea is

<u>deviance</u>:

$$G^2 = 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{p}(\underline{x}_i)}\right) + (1-y_i)\log\left(\frac{1-y_i}{1-\hat{p}(\underline{x}_i)}\right)\right]$$

The ith <u>deviance residual</u> is

$$dev_i = \pm\left[-2\left[y_i \log \hat{p}(\underline{x}_i) + (1-y_i)\log(1-\hat{p}(\underline{x}_i))\right]\right]^{1/2}$$

where $\pm$ is $+$ if $y_i \geq \hat{p}(\underline{x}_i)$

Think $y_i = 1$

- $\hat{p} \approx 1 \Rightarrow$ der ≈ 0
- $\hat{p} \approx 0 \Rightarrow$ der big