Logistic Regression

*Logistic Class Probability Estimation

Darren Reger Lecture for Galvanize DSI

Classification

- Given a feature vector X and a qualitative response Y which takes values in set C, build a function C(X) that takes the feature vector X and predicts the qualitative variable Y

- Examples:
  - Customer churn
  - Species extinction
  - Patient outcome
  - Prospective buyer purchase
  - Eye Color
  - Email is spam

We could model the probability of being in a food coma based on how much steak you ate.

FOGO DE CHÃO®
CHURRASCARIA
BRAZILIAN STEAKHOUSE

Odds

Odds are commonly used in gambling, especially horse-racing

- Even odds (1:1) $\longrightarrow p = \frac{1}{1+1} = 0.5$
- Odds are 3:1 for an event $\longrightarrow p = \frac{3}{1+3} = 0.75$
- Long shot: 20:1 against

  $\longrightarrow 1 - p = 1 - \frac{20}{21} = \frac{1}{21} = 0.0476$

Odds Ratios

Given the definition of odds above, the odds ratio is

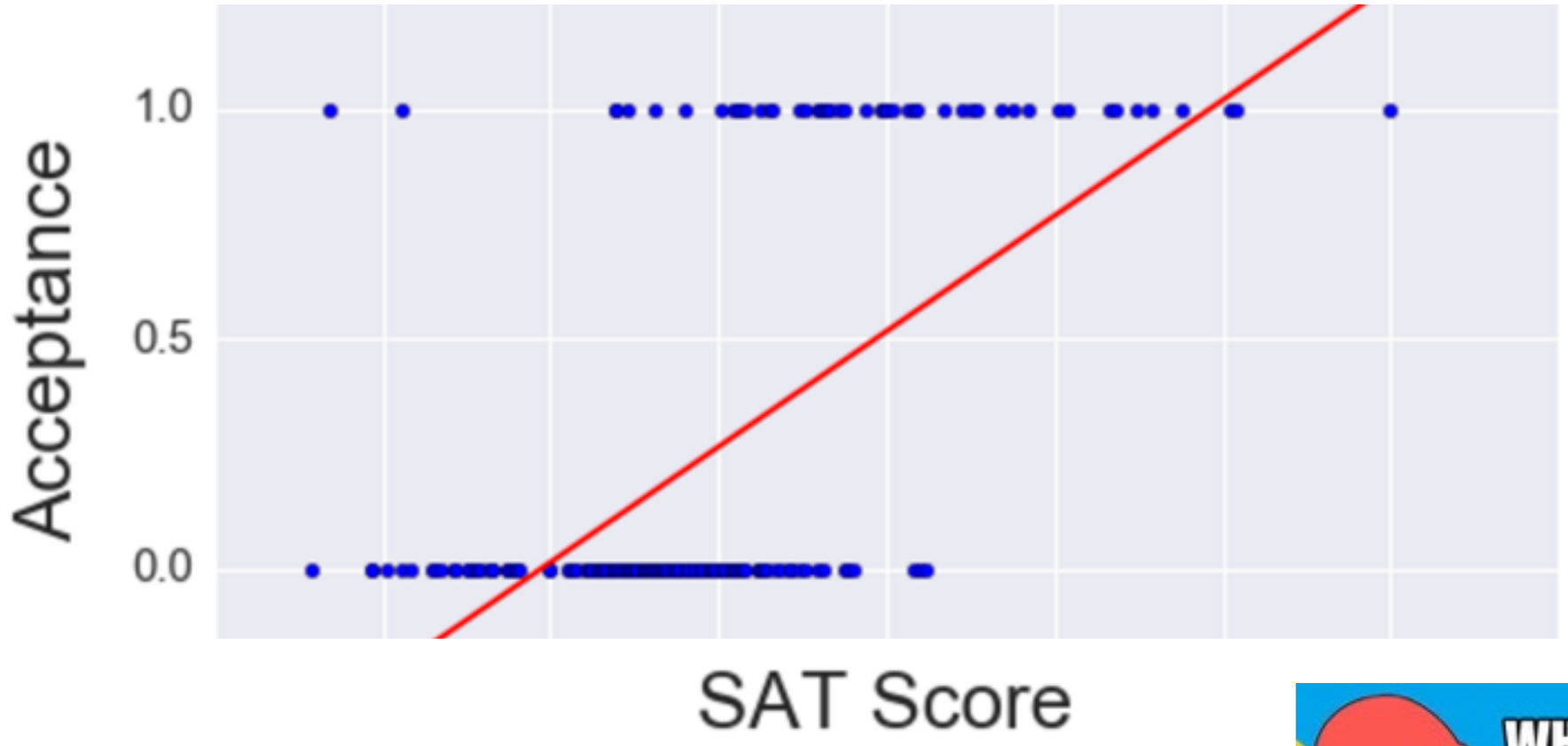
$$OR = \frac{Odds_1}{Odds_2} = \frac{(p_1/(1-p_1))}{(p_2/(1-p_2))}$$

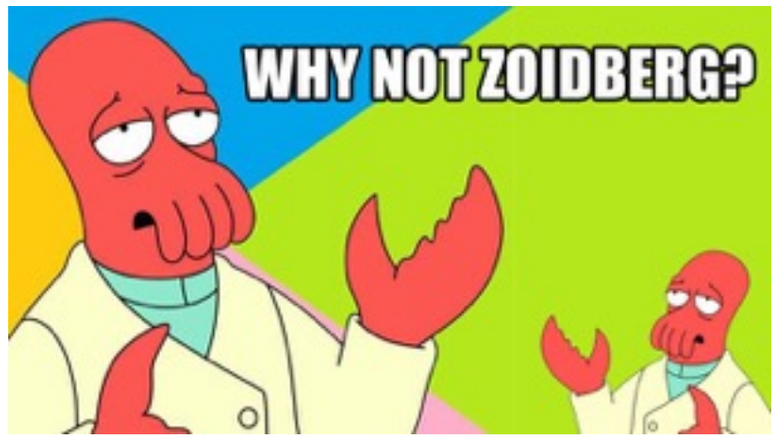
For example, say the probability of a disease in individuals with a certain genetic trait is $p_1 = 0.05$ while in the general population its $p_2 = 0.001$ the resulting odds ratio would be

$$OR = \frac{0.05/0.95}{0.001/0.999} \approx 53$$

This represents a measure of relative risk such that an individual with the genetic trait is 53 time more likely to develop the disease than a randomly chosen person

Why not linear regression?



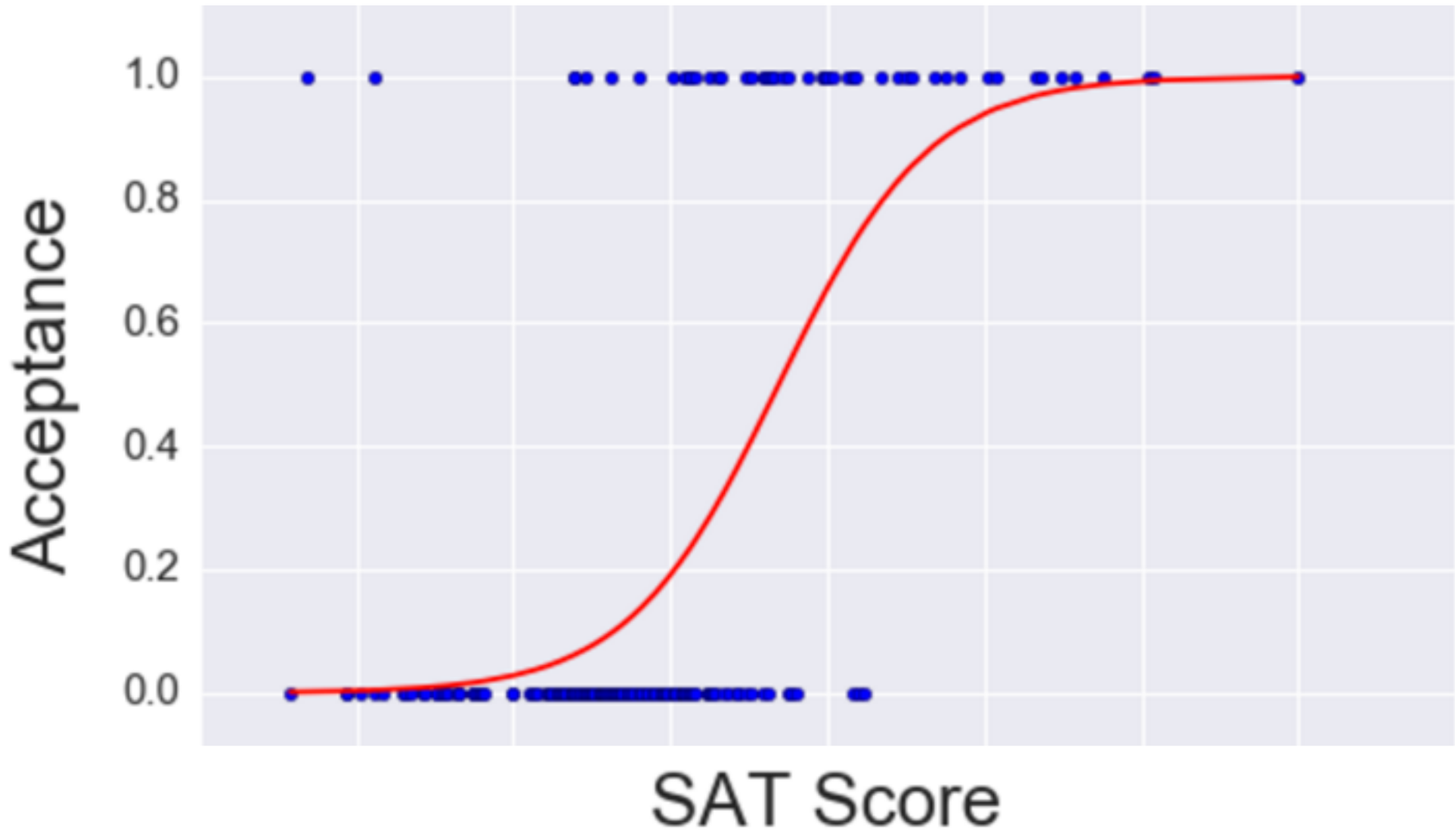


What do we need?

1. Takes continuous input (e.g. -infinity to infinity)
2. Produces output [0,1]
3. Has an intuitive transition
4. Has interpretable coefficients

   (like linear regression)

Logit Function



Logit Function

- The s-shaped curve from the previous slide was created using the logistic function which comes from the sigmoid family.

- This is going to be our “link” function

- How do we turn this into a regression?

$$S(t) = \frac{1}{1 + e^{-t}}$$

How to model

In a **binary classification setting**, the response is **binary**:

$$y_i \begin{cases} 1, \text{ if event occurs} \\ 0, \text{ if event doesn't occur} \end{cases}$$

Each observation is drawn from a **Bernoulli distribution**:

$$y_i \mid X \sim Bernoulli(p)$$

Our **standard** linear model won't work

Generalized Linear Models

It turns out that this is a very general way of addressing this type of problem in regression, and the resulting models are called generalized linear models (GLMs). Logistic regression is just one example of this type of model.

All generalized linear models have the following three characteristics:

1. A probability distribution describing the outcome variable
2. A linear model
   - $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$
3. A link function that relates the linear model to the parameter of the outcome distribution
   - $g(p) = \eta$ or $p = g^{-1}(\eta)$

GLM cont'd

Logistic regression is a GLM used to model a binary categorical variable using numerical and categorical predictors.

We assume a binomial distribution produced the outcome variable and we therefore want to model p the probability of success for a given set of predictors.

To finish specifying the Logistic model we just need to establish a reasonable link function that connects η to p. There are a variety of options but the most commonly used is the logit function.

Logit function

$$logit(p) = \log\left(\frac{p}{1-p}\right), \text{ for } 0 \leq p \leq 1$$

GLM cont'd 2

The logit function takes a value between 0 and 1 and maps it to a value between $-\infty$ and ∞.

Inverse logit (logistic) function

$$g^{-1}(x) = \frac{\exp(x)}{1+\exp(x)} = \frac{1}{1+\exp(-x)}$$

The inverse logit function takes a value between $-\infty$ and ∞ and maps it to a value between 0 and 1.

This formulation also has some use when it comes to interpreting the model as logit can be interpreted as the log odds of a success, more on this later.

GLM Logistic Link

The three GLM criteria give us:

$$y_i \sim \text{Binom}(p_i)$$

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

$$\text{logit}(p) = \eta$$

From which we arrive at,

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i})}{1 + \exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i})}$$

Generalized Linear Models

The parameters of our logistic regression are estimated via **maximum likelihood**. We know that each individual observation follows a Bernoulli distribution:

$$y_i \mid X \sim Bernoulli(p)$$

Given this, we can construct the likelihood of our β matrix as:

$$\mathcal{L}(\beta \mid y) = \prod_{i=1}^{N} p(y_i)^{y_i} \, (1 - p(y_i))^{(1-y_i)}$$

And from there, our log likelihood:

$$\ell = \sum_{i=1}^{N} y_i \log p(y_i) + (1 - y_i) log(1 - p(y_i))$$

Finding the Coefficients

Notation change: p —> h(theta) log—>ln

In practice, we maximize the log likelihood:

$$\ln p(\vec{y}|X;\theta) = \sum_{i=1}^{n} (y_i \ln h_\theta(x_i) + (1 - y_i) \ln(1 - h_\theta(x_i)))$$

Observe how the value of each term varies:

$$y_i = 0 \Rightarrow \lim_{h_\theta(x)\to 0} (y_i \ln h_\theta(x_i) + (1 - y_i) \ln(1 - h_\theta(x_i))) = 0$$

$$y_i = 0 \Rightarrow \lim_{h_\theta(x)\to 1} (y_i \ln h_\theta(x_i) + (1 - y_i) \ln(1 - h_\theta(x_i))) = -\infty$$

$$y_i = 1 \Rightarrow \lim_{h_\theta(x)\to 1} (y_i \ln h_\theta(x_i) + (1 - y_i) \ln(1 - h_\theta(x_i))) = 0$$

$$y_i = 1 \Rightarrow \lim_{h_\theta(x)\to 0} (y_i \ln h_\theta(x_i) + (1 - y_i) \ln(1 - h_\theta(x_i))) = -\infty$$

Finding the Coefficients

The regression coefficients can be estimated using maximum likelihood estimation

Unlike linear regression, no closed form solution exists, therefore an iterative method such as Newton-Rhapson or Gradient Descent is needed

Reasons that the model may not reach convergence

- A large number of features relative to subjects $\longrightarrow$ rule of thumb is at least 10 cases for each explanatory variable
- Multicollinearity
- Sparseness, specifically low cell counts for categorical predictors

Interpreting Coefficients

In linear regression, the $\hat{\beta}$ coefficients can be interpreted directly as the change in y for a 1-unit increase in the explanatory variable

In logistic regression, however, this would represent the change in logit value for a 1-unit increase in the explanatory variable, which is not interpretable

We can however convert the $\hat{\beta}$ coefficient to an estimate of Odds Ratio for a 1-unit increase in the explanatory variable

$$\widehat{OR} = e^{\hat{\beta}}$$

Making Predictions

Once the $\hat{\beta}$ coefficients have been calculated, we can estimate the probabilities of the event occurring ($Y = 1$) for a specific covariate profile (X)

$$\hat{\pi} = \frac{e^{X\hat{\beta}}}{1 + e^{X\hat{\beta}}}$$

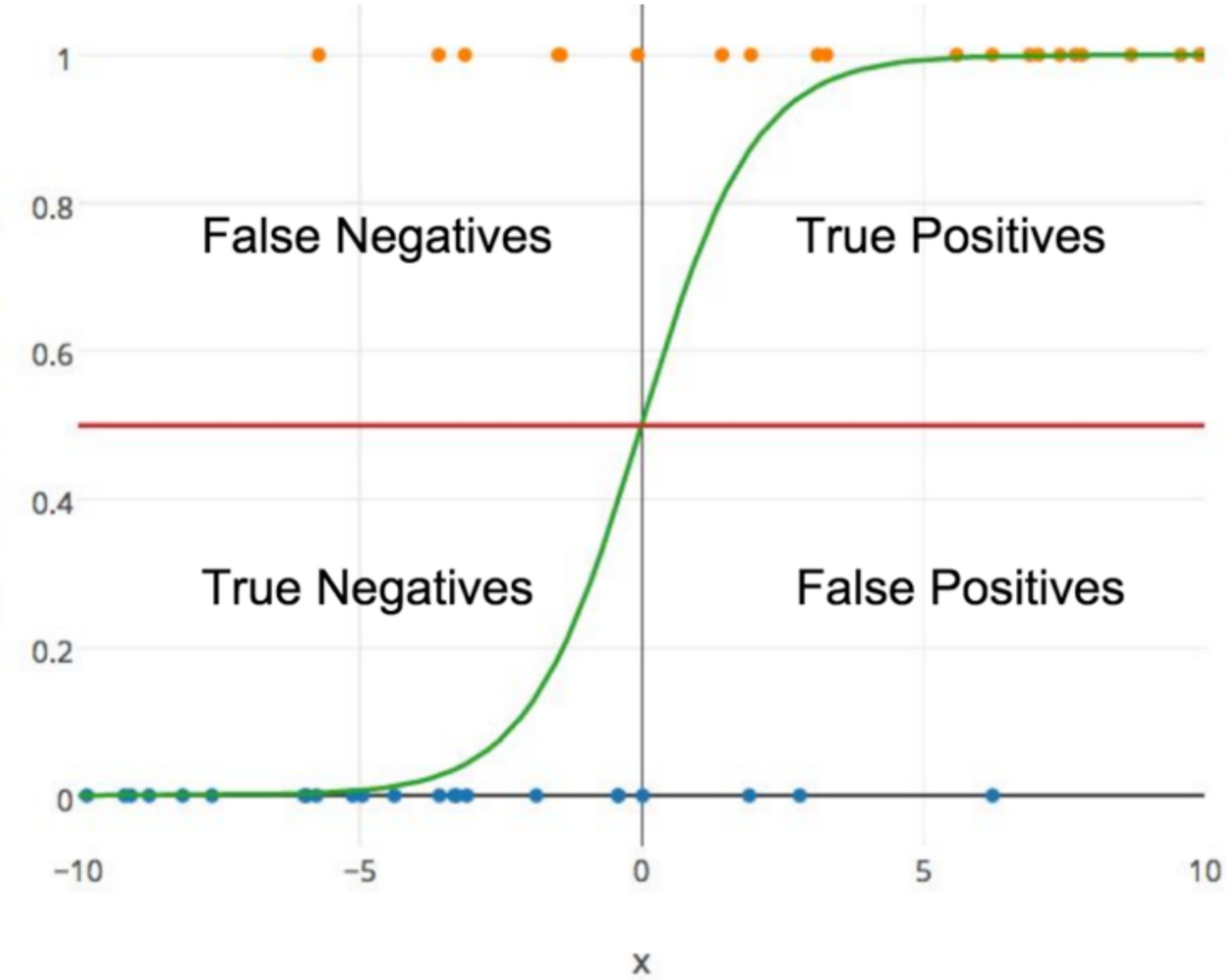
$\hat{\pi}$ is a vector of probabilities for the entire sample.

To find a specific probability π_i, you would find the dot product of the i^{th} row of the X matrix and the vector of $\hat{\beta}$ coefficients

Titanic Data:
With logistic regression it feels like your on top of the world!

ROC Curves & Scoring Metrics

Assessing Model Fit



Comparison with Hypothesis Testing

	H_0 is true	H_0 is false
Accept H_0	Correct Decision ($1-\alpha$)	Type II Error (β)
Reject H_0	Type I Error (α)	Correction Decision ($1-\beta$)

	Predicted Positive	**Predicted Negative**
Actually Positive	True Positives	False Negatives
Actually Negative	False Positives	True Negatives

Questions We Might Have

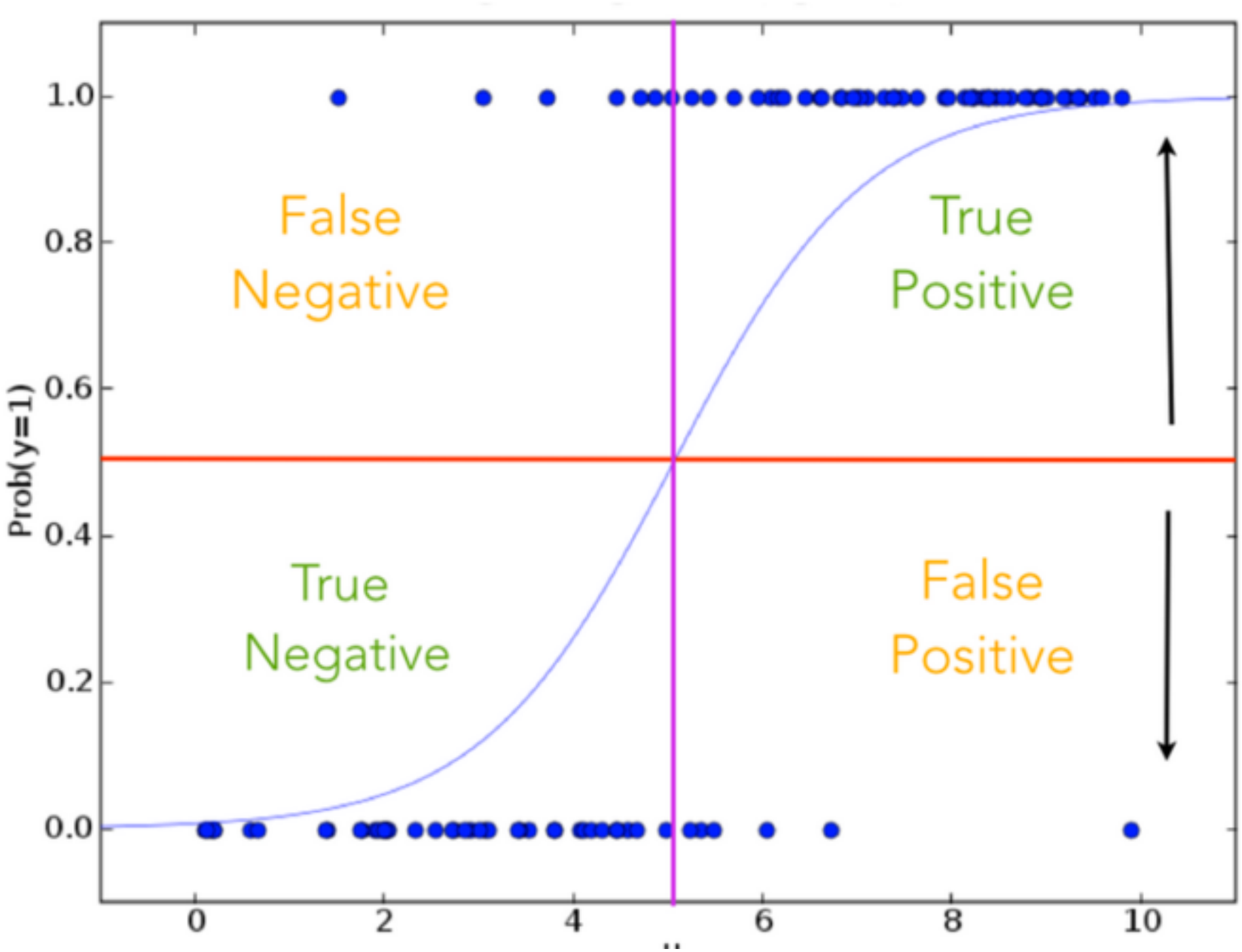
- How often is our classifier correct?
    - ★ Accuracy
- How often is our classifier wrong?
    - ★ Misclassification Rate (1 - Accuracy) aka Error Rate
- When it's actually a yes, how often does it predict yes?
    - ★ True Positive Rate aka Sensitivity/Recall
- When it's actually a no, how often does it predict no?
    - ★ False Positive Rate
- When it predicts yes, how often is it correct?
    - ★ Precision
- How often does a yes actually occur in our sample?
    - ★ Prevalence

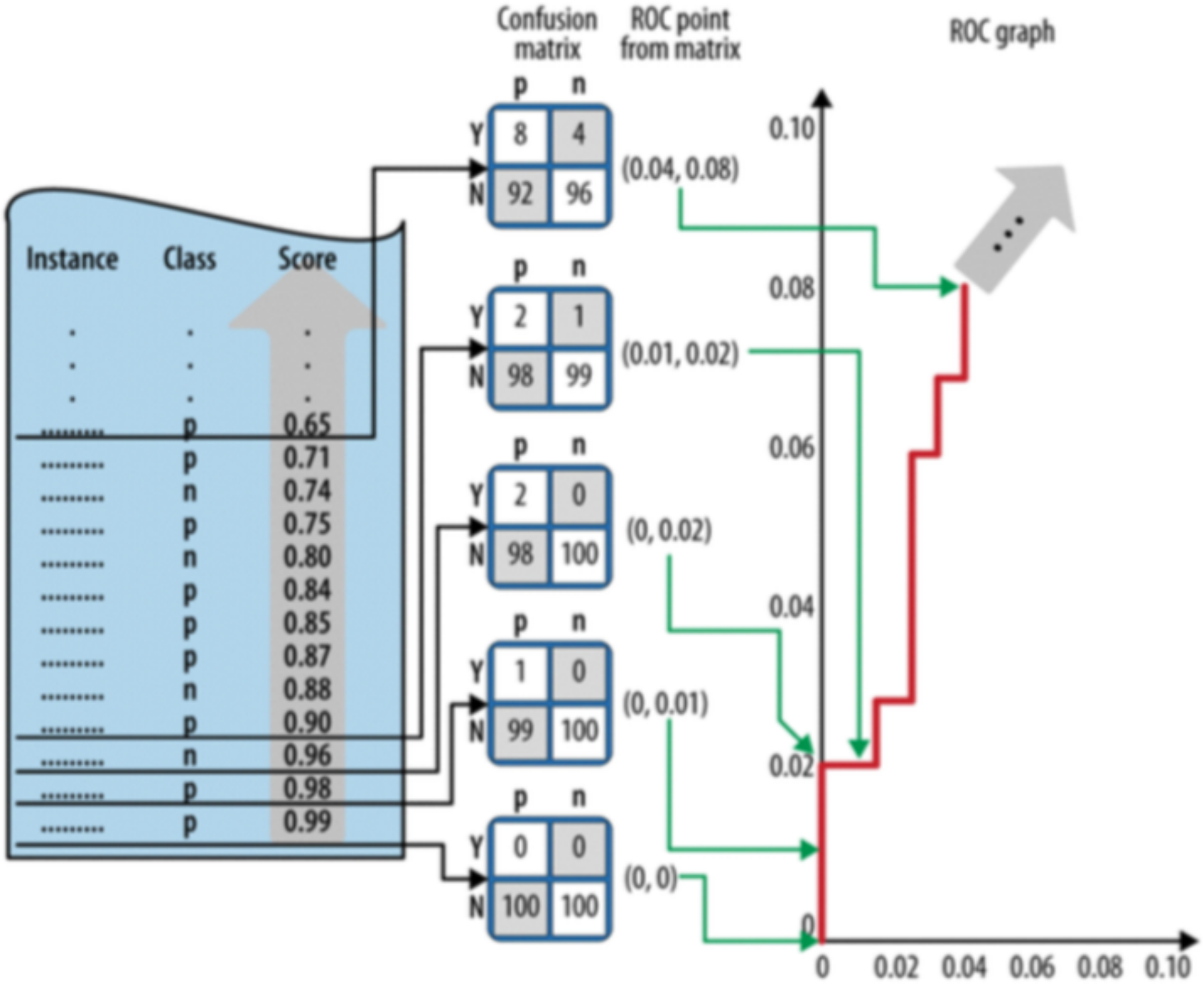
Confusion Matrix

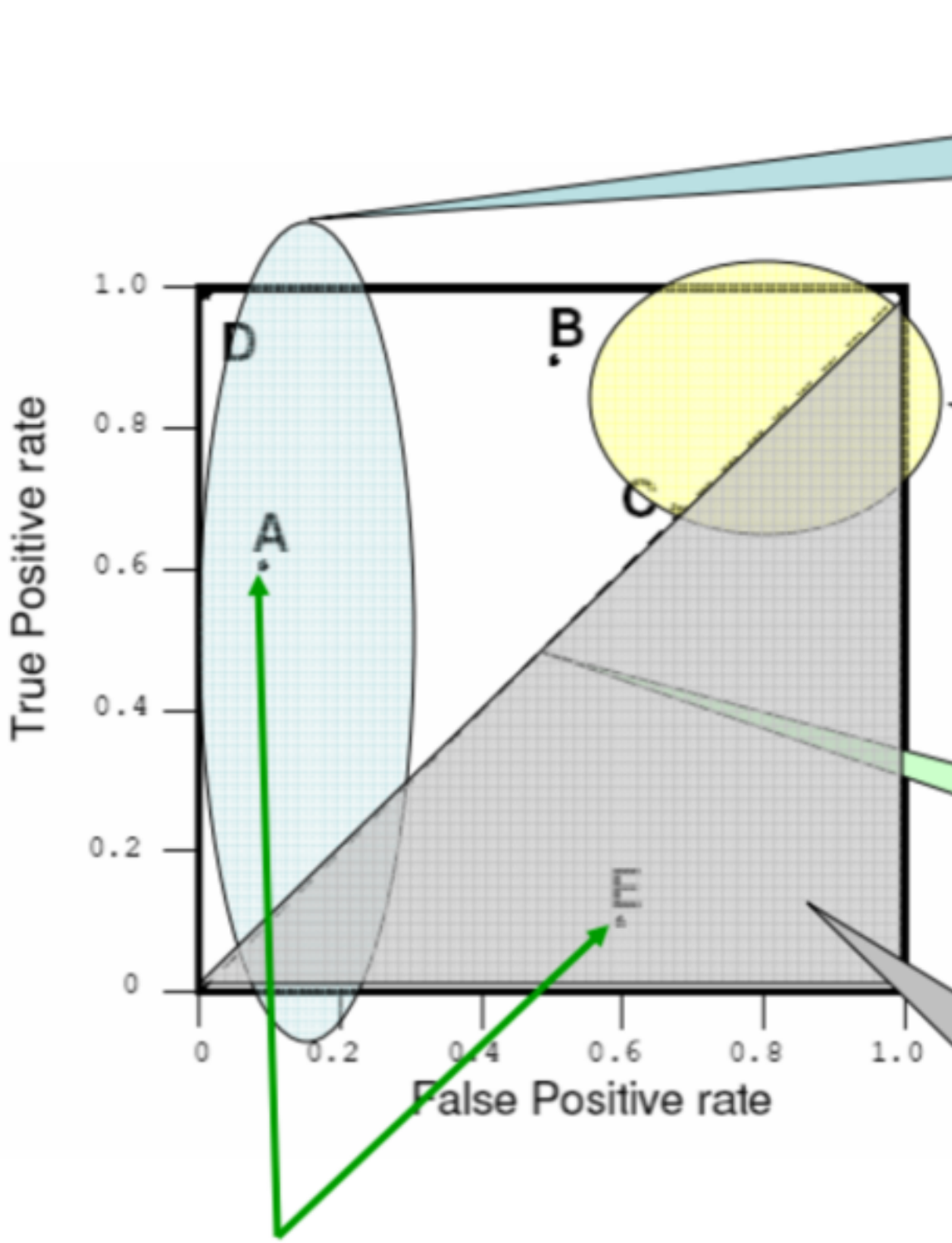
n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

ROC Curve

Think of sliding the purple/red line along the sigmoid function




Confusion matrix
ROC point from matrix
ROC graph
Instance
Class
Score
p
n
Y
N
8
4
92
96
(0.04, 0.08)
2
1
98
99
(0.01, 0.02)
2
0
98
100
(0, 0.02)
1
0
99
100
(0, 0.01)
0
0
100
100
(0, 0)
0.65
0.71
0.74
0.75
0.80
0.84
0.85
0.87
0.88
0.90
0.96
0.98
0.99
0.10
0.08
0.06
0.04
0.02
0
0.02
0.04
0.06
0.08
0.10



Positive predicted only on strong evidence, low FP, low TP

Positive predicted with weak evidence, high TP rate, high FP also

Line y=x
Randomly guessing the class
No model

Worse than random guessing

Note: E (bad) is negation of A

Comparing ROC Curves