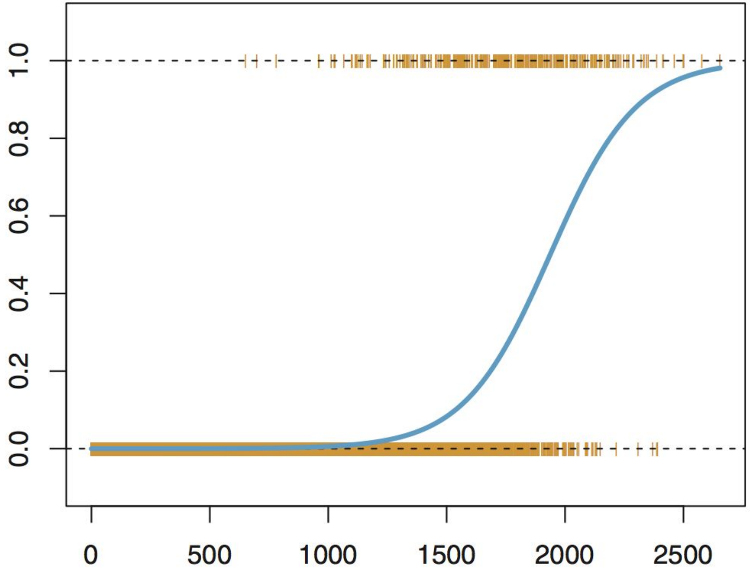
Logistic Regression 1/2

Classification, metrics and ROC curves

DSI, jf.omhover, Dec 6, 2016



Logistic Regression 1/2

Classification, metrics and ROC curves

DSI, jf.omhover, Dec 6, 2016

OBJECTIVES (morning)

- **Relate** Regression to Classification in the context of supervised learning
- **Compare** Logistic Regression to Linear Regression
- **Define** and **compute** metrics for evaluating classifiers

OBJECTIVES (afternoon)

- **Describe** the process for computing parameter values in LogReg
- **Use** the parameters of a LogReg model to **compute** the class of an obverstion

Supervised Learning

Learning / Estimating FUNCTIONS based on examples

Reality VS Model : assumptions and learning

REALITY

OBJECTIVE:
descriptive
predictive
normative
...

MODEL

	type	income	education	prestige
accountant	prof	62	86	82
pilot	prof	72	76	83
architect	prof	75	92	90
author	prof	55	90	76
chemist	prof	64	86	90
minister	prof	21	84	87
professor	prof	64	93	93
dentist	prof	80	100	90
reporter	wc	67	87	52
engineer	prof	72	86	88
undertaker	prof	42	74	57
lawyer	prof	76	98	89

data

$$(x_1, y_1)$$

...

$$(x_n, y_n)$$

$x \quad y$

$$y = f(x) + \epsilon$$

take a function as an assumption

$$\sum (y_i - \hat{f}(x_i))^2$$

COST FUNCTION

$$\hat{y} = \hat{f}(x)$$

Estimator of the function

Linear Regression

REALITY

	type	income	education	prestige
accountant	prof	62	86	82
pilot	prof	72	76	83
architect	prof	75	92	90
author	prof	55	90	76
chemist	prof	64	86	90
minister	prof	21	84	87
professor	prof	64	93	93
dentist	prof	80	100	90
reporter	wc	67	87	52
engineer	prof	72	86	88
undertaker	prof	42	74	57
lawyer	prof	76	98	89

data

$$(x_1, y_1)$$

...

$$(x_n, y_n)$$

$x \quad y$

OBJECTIVE:
descriptive
predictive
normative
...

MODEL

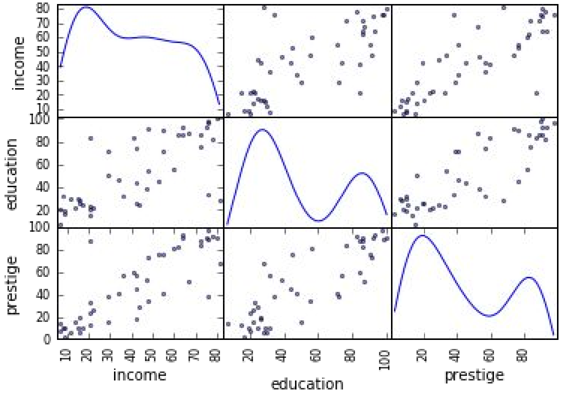
$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

We make the assumption that we have a linear relation

Linear Regression - General Process

REALITY

1) Having a data sample
Observing an underlying behavior

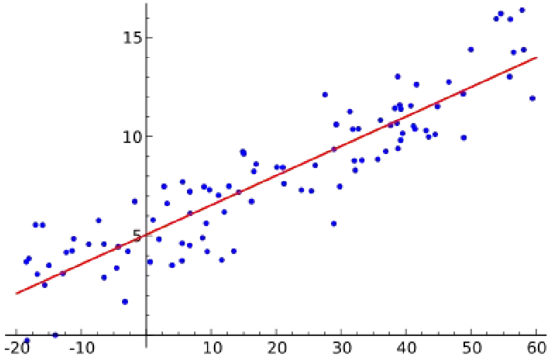


MODEL

2) Make an assumption
on the <u>model</u> underlying the data

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

linear relation
(+ assumptions)



3) <u>**Find**</u> the instance of the model
that <u>**fits**</u> with data sample

Multi-Linear Regression

COST FUNCTION (Residual Sum of Squares)

$$RSS(\beta) = (y - X\beta)^T(y - X\beta)$$

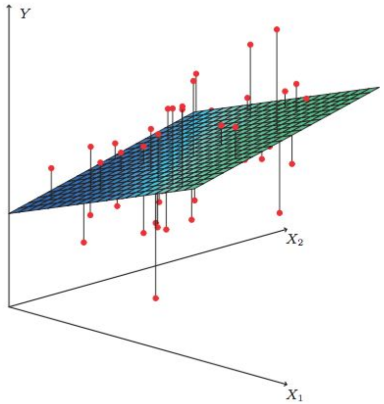
O.L.S.

REALITY

MODEL

DATA

$$X = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p-1} \\ 1 & x_{2,1} & \cdots & x_{2,p-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p-1} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = y$$



model class

$$y \approx X\beta$$

PROBLEM

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$$

$$\hat{y} = X\hat{\beta}$$

model instance
estimator
parameters

SOLUTION

$$\boxed{\hat{\beta} = (X^TX)^{-1}X^Ty}$$

Figure from [ISL]

Classification

Learning / Estimating “models of classes” based on examples

Reality vs Model : assumptions and learning

REALITY

$$X = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p-1} \\ 1 & x_{2,1} & \cdots & x_{2,p-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p-1} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = y$$

OBJECTIVE:
descriptive
predictive
normative
...

$$\Sigma(y_i - \hat{f}(x_i))^2$$

COST FUNCTION

MODEL

$$y = f(x) + \epsilon$$

take a function as an assumption

$$\hat{y} = \hat{f}(x)$$

Estimator of the function

Mapping // Classification algorithms

Logistic Regression

k-NN

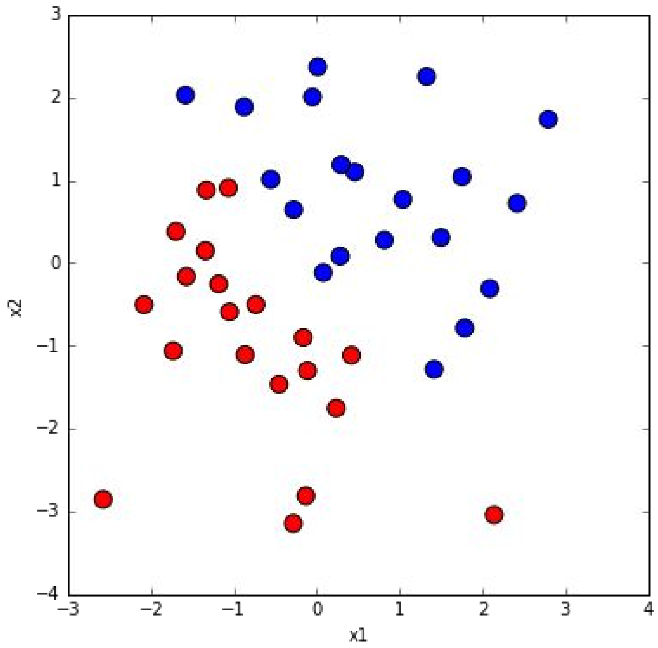
Decision Trees

Random Forest, Boosting

Support Vector Machines (SVM)

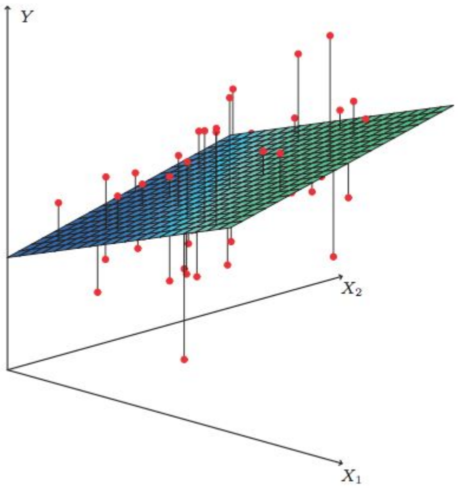
Neural Networks

...

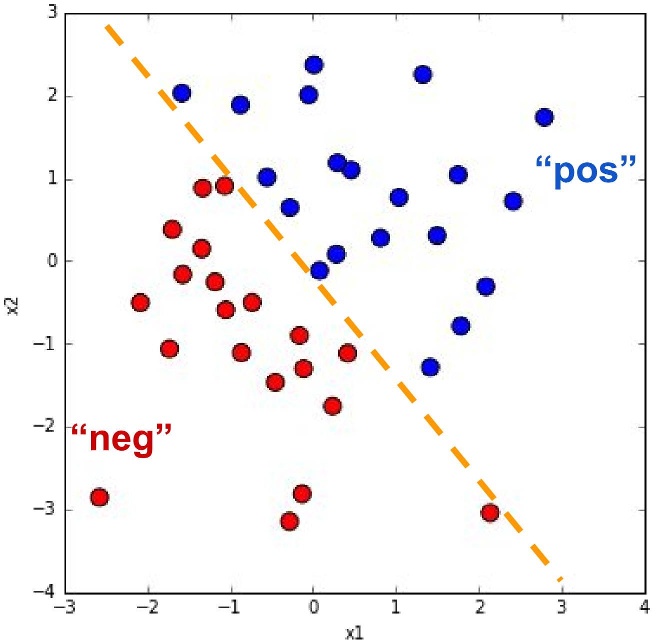


Regression vs Classification

Quantitative response y in R



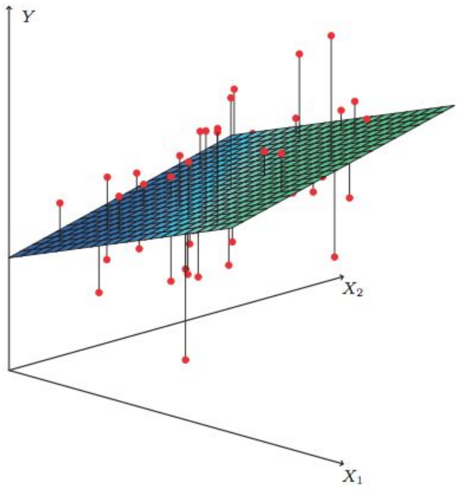
Categorial response y



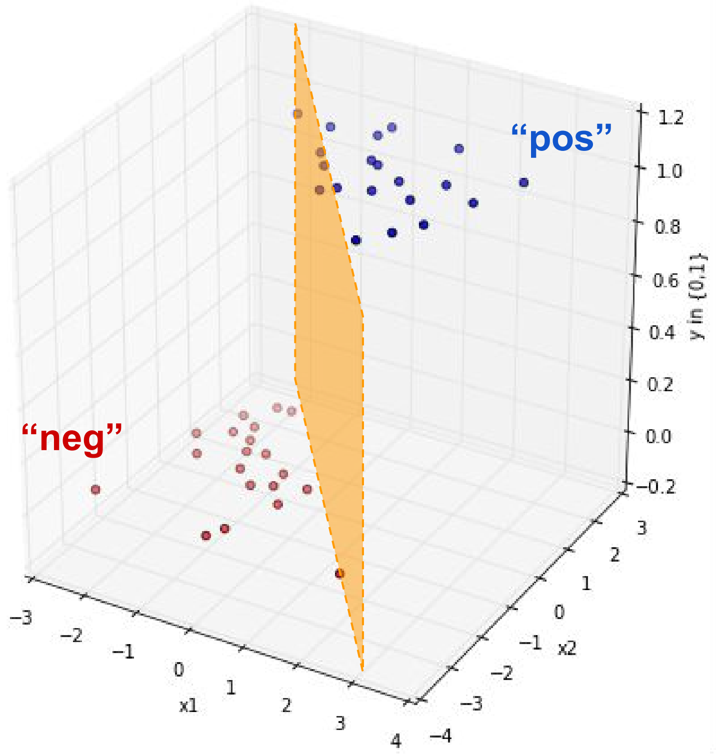
Assigning y = 0 to neg, y = 1 to pos

Regression vs Classification

Quantitative response y in R

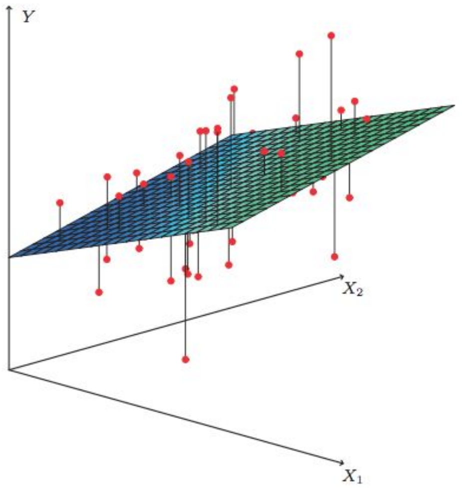


Categorial response y in {0,1}

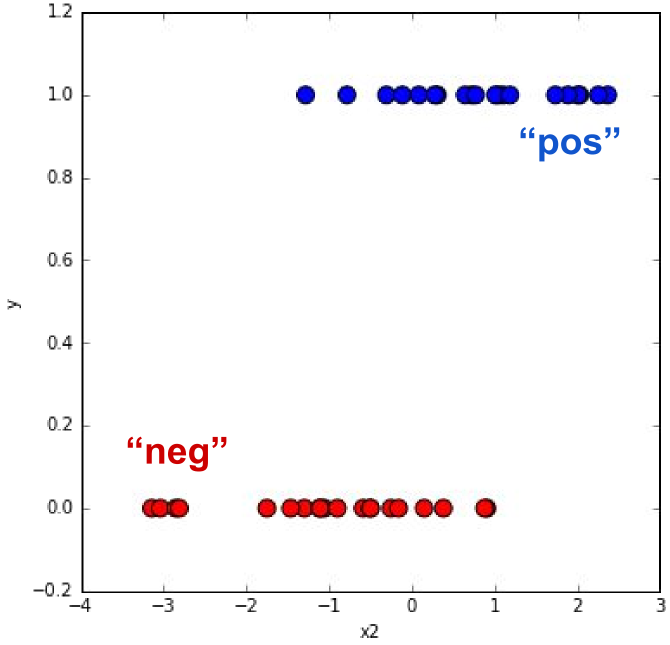


Regression vs Classification

Quantitative response y in R

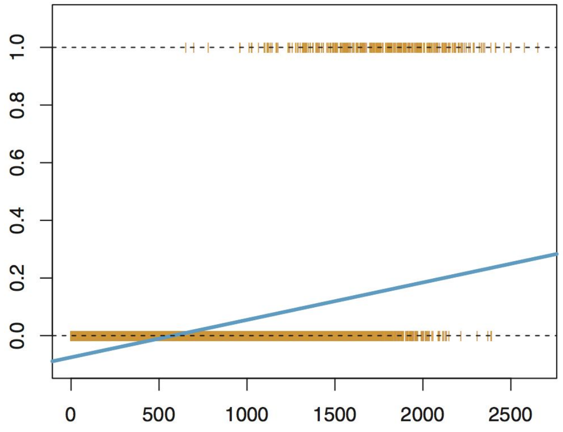


Categorial response y in {0,1}



Trying to apply LinReg to y

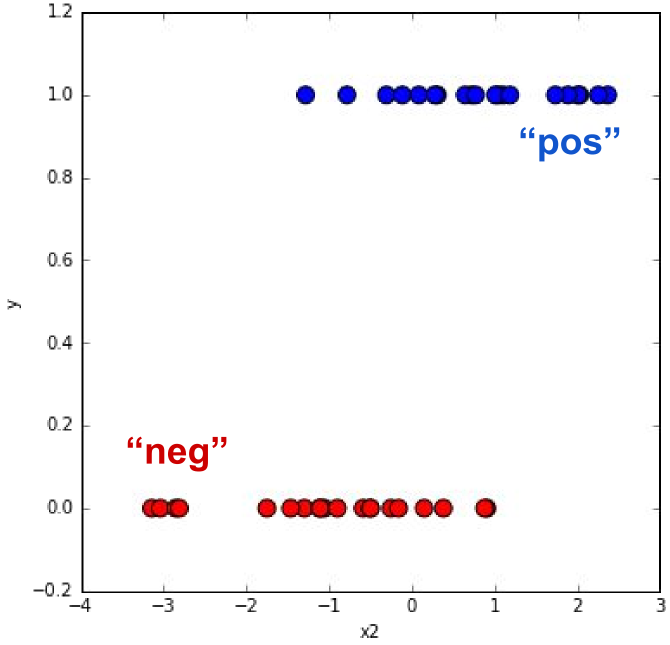
Quantitative response y in R



$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Negative probabilities ?
How to cut-off ?

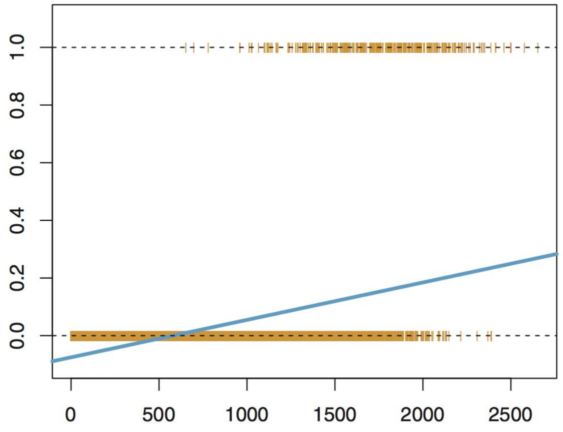
Categorial response y in {0,1}



LogReg as model of probability

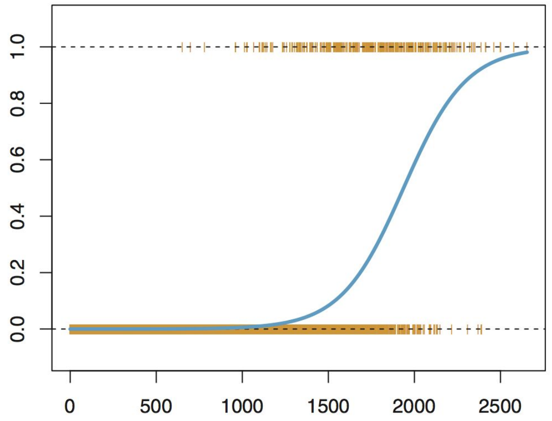
Quantitative response y in R

Categorial response y in {0,1}



$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Negative probabilities ?
How to cut-off ?



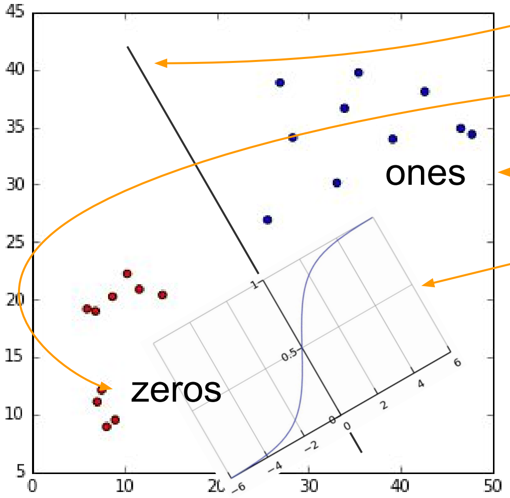
$$p(X) = h(\beta_0 + \beta_1 . x_1 + \cdots + +\beta_p . x_p)$$

Idea : model probability of being positive
as a function of a linear model

LogReg in a nutshell

REALITY

MODEL

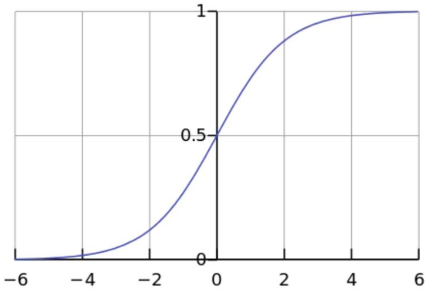


$$p(X) = h(\beta_0 + \beta_1 . x_1 + \cdots + +\beta_p . x_p)$$

It (badly) translates as :
computes the probability of being in one of the two classes
depending on of the side and distance of the plan

$$h : \mathbb{R} \rightarrow [0, 1]$$

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



Mapping // Classification algorithms

Logistic Regression

k-NN

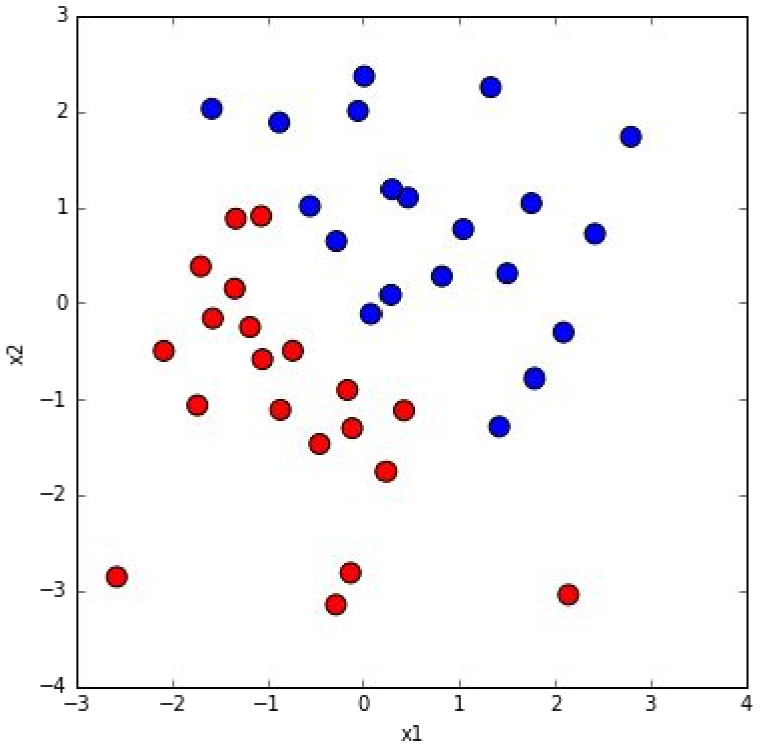
Decision Trees

Random Forest, Boosting

Support Vector Machines (SVM)

Neural Networks

...



How to evaluate a classifier ?

Conforming a classifier to the actual response

x_1	x_2	x_3	x_4	y
...	...	...	...	1
...	...	...	...	0
...	...	...	...	0
...	...	...	...	1
...	...	...	...	1
...	...	...	...	1
…	...	...	...	0
...	...	...	...	1

$\hat{y} = \hat{f}(x)$

1	0.95
0	0.21
1	0.55
0	0.43
1	0.77
0	0.44
0	0.15
1	0.81

$\hat{y} = \hat{f}(x)$

y		Pred P	Pred N	
	Actual P	3	2	P = 5
	Actual N	1	2	N = 3
		P*	N*	

Confusion Matrix

$$\hat{y} = \hat{f}(x)$$

y		Pred P	Pred N	
	Actual P	True Positive	False Negative	P = 5
	Actual N	False Positive	True Negative	N = 3
		P*	N*	

Confusion Matrix - Metrics

The proportion of <u>observations</u> that are <u>correctly</u> classified ?

Accuracy :

The proportion of <u>positives</u> that are <u>correctly</u> identified as such ?

True Pos Rate :

(aka recall, sensitivity)

The proportion of <u>negatives</u> that are <u>correctly</u> identified as such

True Neg Rate :

(aka specificity)

$$\hat{y} = \hat{f}(x)$$

y		Pred P	Pred N	
	Actual P	True Positive	False Negative	P = 5
	Actual N	False Positive	True Negative	N = 3
		P*	N*	

Confusion Matrix - Metrics

The proportion of <u>observations</u> that are <u>correctly</u> classified ?

Accuracy : (TN + TP) / (N + P)

The proportion of <u>positives</u> that are <u>correctly</u> identified as such ?

True Pos Rate : TP / P

(aka recall, sensitivity)

The proportion of <u>negatives</u> that are <u>correctly</u> identified as such

True Neg Rate : TN / N

(aka specificity)

$$\hat{y} = \hat{f}(x)$$

y	Pred P	Pred N	
Actual P	True Positive	False Negative	P = 5
Actual N	False Positive	True Negative	N = 3
	P*	N*	

Confusion Matrix - Metrics

The proportion of <u>observations</u> that are
<u>NOT correctly</u> classified ?

Error rate :

The proportion of <u>positives</u> that are
<u>NOT correctly</u> identified as such ?

False Neg Rate :

(aka fall-out)

The proportion of <u>negatives</u> that are
<u>NOT correctly</u> identified as such

False Pos Rate :

(aka 1-specificity)

$$\hat{y} = \hat{f}(x)$$

y

	Pred P	Pred N	
Actual P	True Positive	False Negative	P = 5
Actual N	False Positive	True Negative	N = 3
	P*	N*	

Confusion Matrix - Metrics

The proportion of <u>observations</u> that are <u>NOT correctly</u> classified ?

Error rate : (FN + FP) / (N + P)

The proportion of <u>positives</u> that are <u>NOT correctly</u> identified as such ?

False Neg Rate : FN / P

(aka fall-out)

The proportion of <u>negatives</u> that are <u>NOT correctly</u> identified as such

False Pos Rate : FP / N

(aka 1-specificity)

$$\hat{y} = \hat{f}(x)$$

y		Pred P	Pred N	
	Actual P	True Positive	False Negative	P = 5
	Actual N	False Positive	True Negative	N = 3
		P*	N*	

Confusion Matrix - Metrics

The proportion of <u>actual positives</u>
<u>in those identified</u> as such ?

Precision : TP / (FP + TP)

The proportion of <u>positives</u> that are
<u>correctly</u> identified as such ?

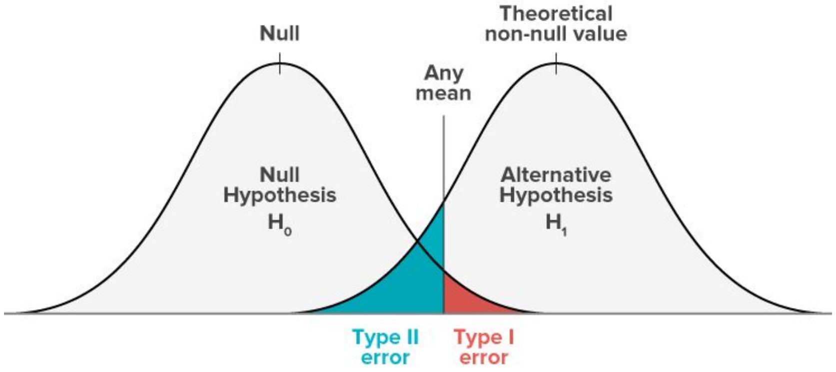
Recall : TP / P

(aka TPR, sensitivity)

$$\hat{y} = \hat{f}(x)$$

y	Pred P	Pred N	
Actual P	True Positive	False Negative	P = 5
Actual N	False Positive	True Negative	N = 3
	P*	N*	

Confusion Matrix - type I and type II error



$$\hat{y} = \hat{f}(x)$$

y

	Pred P	Pred N	
Actual P	good	Type II error	P = 5
Actual N	Type I error	good	N = 3
	P*	N*	

Using response probabilities

x_1	x_2	x_3	x_4	y
...	...	...	...	1
...	...	...	...	0
...	...	...	...	0
...	...	...	...	1
...	...	...	...	1
...	...	...	...	1
…	...	...	...	0
...	...	...	...	1

P > 0.5	
1	0.95
0	0.21
1	0.55
0	0.43
1	0.77
0	0.44
0	0.15
1	0.81

$$\hat{y} = \hat{f}(x)$$

y		Pred P	Pred N	
	Actual P	True Positive	False Negative	P = 5
	Actual N	False Positive	True Negative	N = 3
		P*	N*	

Cut-offs on probabilities

x_1	x_2	x_3	x_4	y
...	...	...	...	**1**
...	...	...	...	**0**
...	...	...	...	**0**
...	...	...	...	**1**
...	...	...	...	**1**
...	...	...	...	**1**
...	...	...	...	**0**
...	...	...	...	**1**

P > 0.5

1	0.95
0	0.21
1	0.55
0	0.43
1	0.77
0	0.44
0	0.15
1	0.81

P > 0.6

1	0.95
0	0.21
0	**0.55**
0	0.43
1	0.77
0	0.44
0	0.15
1	0.81

P > 0.7

1	0.95
0	0.21
0	0.55
0	0.43
1	0.77
0	0.44
0	0.15
1	0.81

P > 0.8

1	0.95
0	0.21
0	0.55
0	0.43
0	**0.77**
0	0.44
0	0.15
1	0.81

P > 0.9

1	0.95
0	0.21
0	0.55
0	0.43
0	0.77
0	0.44
0	0.15
0	**0.81**

Those are sure ones ! => low FPR ! (high precision)
But we miss so many ones ! => low TPR ! (low recall)

Cut-offs on probabilities

x_1	x_2	x_3	x_4	y
...	...	...	...	1
...	...	...	...	0
...	...	...	...	0
...	...	...	...	1
...	...	...	...	1
...	...	...	...	1
...	...	...	...	0
...	...	...	...	1

P > 0.5

1	0.95
0	0.21
1	0.55
0	0.43
1	0.77
0	0.44
0	0.15
1	0.81

P > 0.4

1	0.95
0	0.21
1	0.55
1	**0.43**
1	0.77
1	**0.44**
0	0.15
1	0.81

P > 0.3

1	0.95
0	0.21
1	0.55
1	0.43
1	0.77
1	0.44
0	0.15
1	0.81

P > 0.2

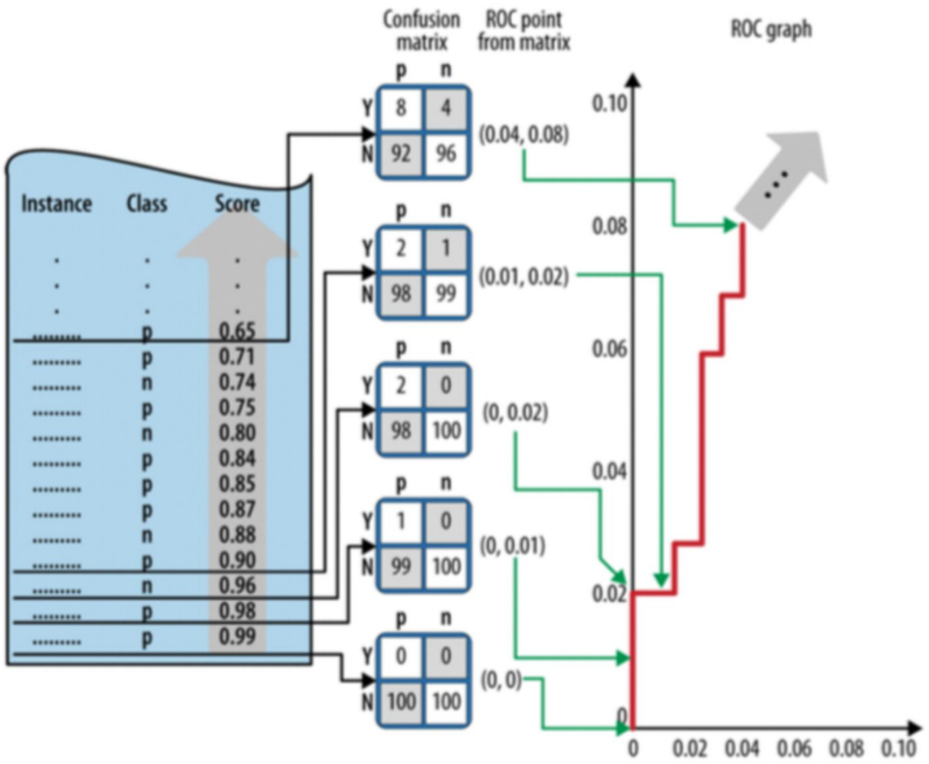
1	0.95
1	**0.21**
1	0.55
1	0.43
1	0.77
1	0.44
0	0.15
1	0.81

P > 0.1

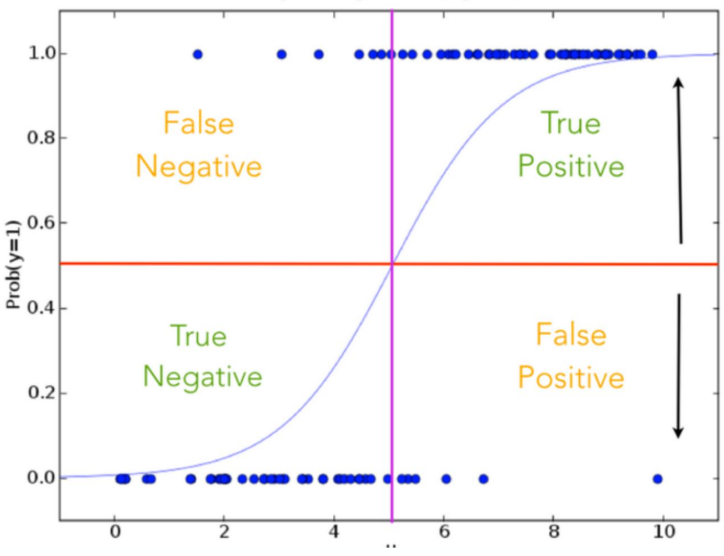
1	0.95
1	0.21
1	0.55
1	0.43
1	0.77
1	0.44
1	0.15
1	0.81

We have so many FP ! => high FPR ! (low precision)
But we capture all the ones ! => high TPR ! (high recall)

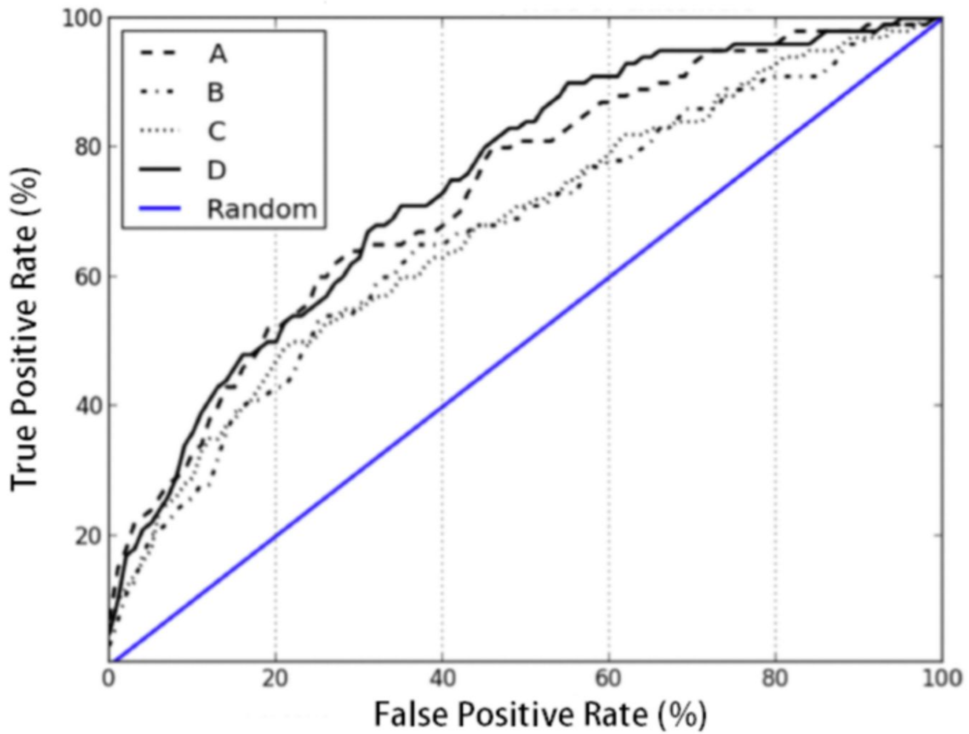
ROC curve (receiver operating characteristic)



For LogReg, think of it as sliding the purple/red line along the sigmoid function

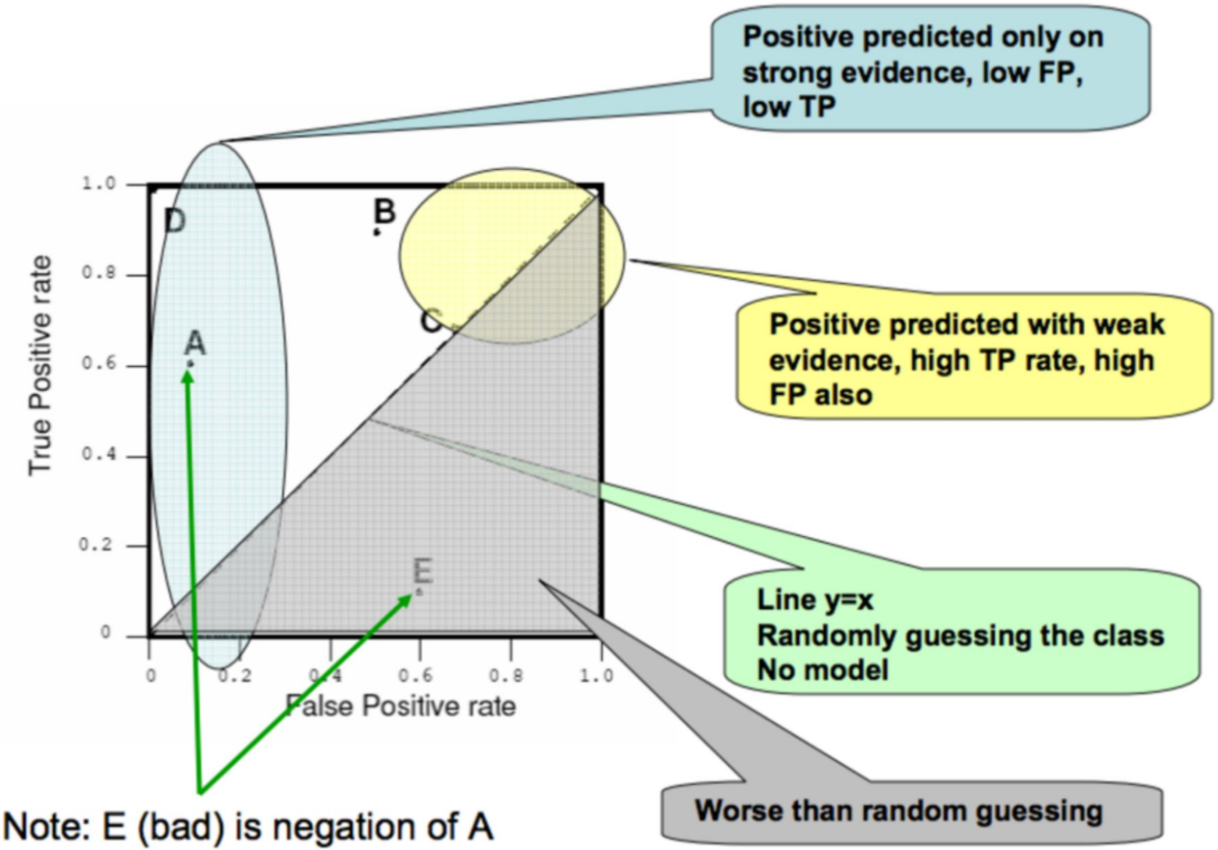


Comparing classifiers based on their ROC curve



Possible metric : AUC
Area-under-curve

What is the “ideal” / “worst” classifier ?



Note: E (bad) is negation of A