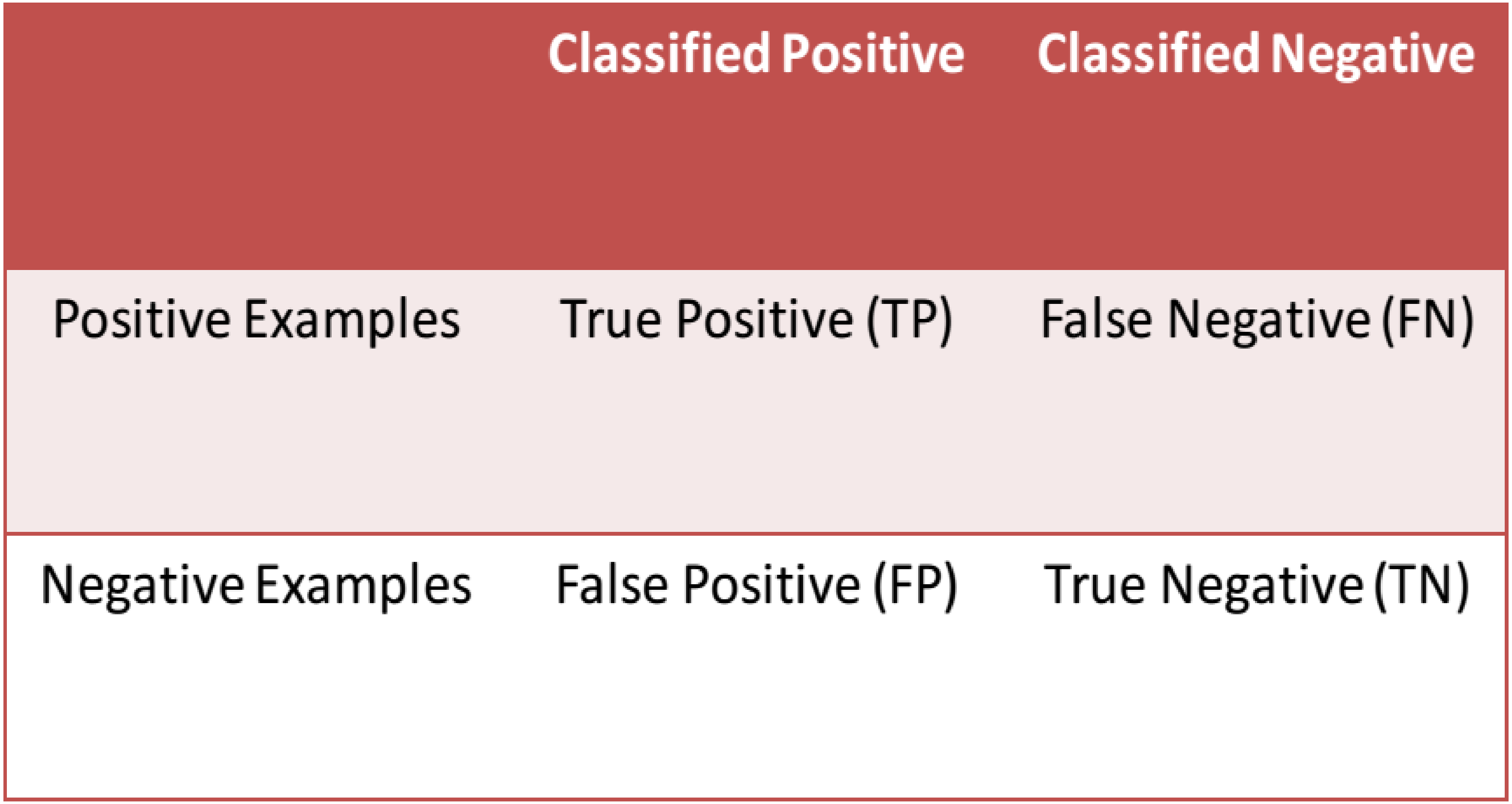
Evaluation Metrics

Classification Metrics

- Accuracy, Precision, Recall, Specificity/FPR
- PR curve & ROC curve
- Kappa, R^2

Confusion Matrix



	Classified Positive	Classified Negative
Positive Examples	True Positive (TP)	False Negative (FN)
Negative Examples	False Positive (FP)	True Negative (TN)

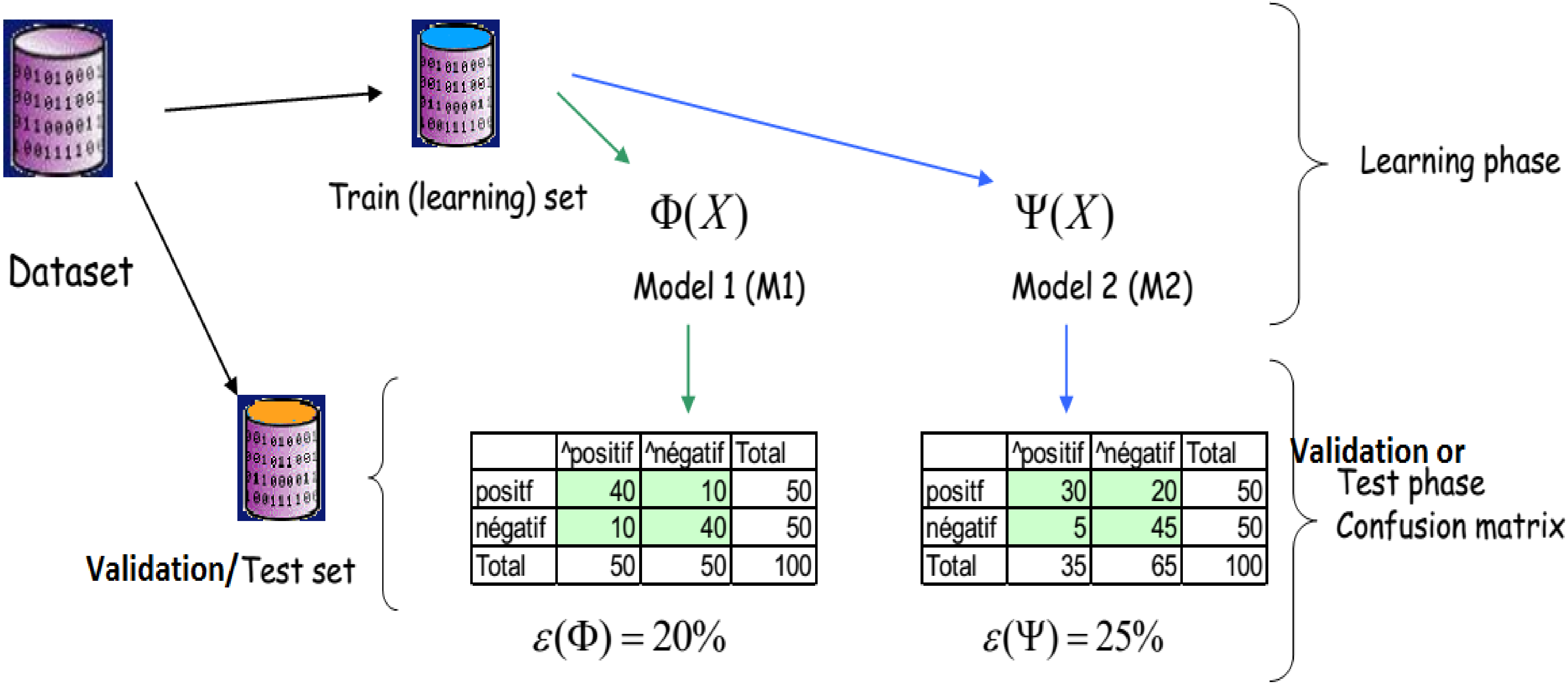
Error Rate/Accuracy metric

- The error rate (computed on a test/validation set) is the most popular summary measure because it is an estimator of the probability of misclassification and it is easy to calculate.

- Error rate = (FP+FN) / total

- Accuracy = 1 – (Error rate)

Error Rate Metric



	^positif	^négatif	Total
positf	40	10	50
négatif	10	40	50
Total	50	50	100

$$\varepsilon(\Phi) = 20\%$$

	^positif	^négatif	Total
positf	30	20	50
négatif	5	45	50
Total	35	65	100

$$\varepsilon(\Psi) = 25\%$$

- Which model is best?

Error Rate may be too simplistic!!

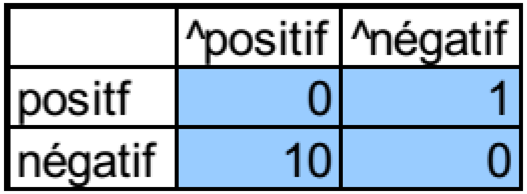
- When misclassification costs are not same
- When class imbalance datasets are used
- When test data may not be representative

Case-I: Non-symmetrical misclassification costs

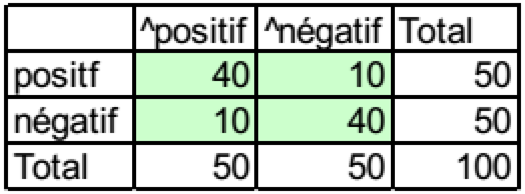
- Our conclusion makes the assumption that we have an unit misclassification costs matrix (the error costs are symmetric)

- Assumption of having unit misclassification costs is not true in most cases

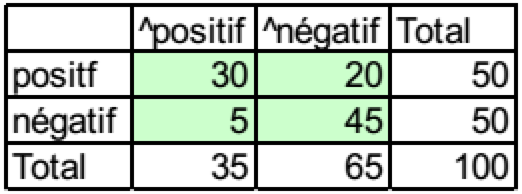
Case-I: Non-symmetrical misclassification costs



	^positif	^négatif
positf	0	1
négatif	10	0



	^positif	^négatif	Total
positf	40	10	50
négatif	10	40	50
Total	50	50	100



	^positif	^négatif	Total
positf	30	20	50
négatif	5	45	50
Total	35	65	100

Average cost of misclassification

$c(\Phi) = 1.1$

$c(\Psi) = 0.7$

- Which model is best now?

Case-I: Non-symmetrical misclassification costs

- Specifying the misclassification costs matrix is often difficult. The costs can vary according to the circumstances.

- Should we try a large number of matrices for comparing M1 and M2?

- Can we use a tool which allows us to compare the models regardless of the misclassification costs matrix?

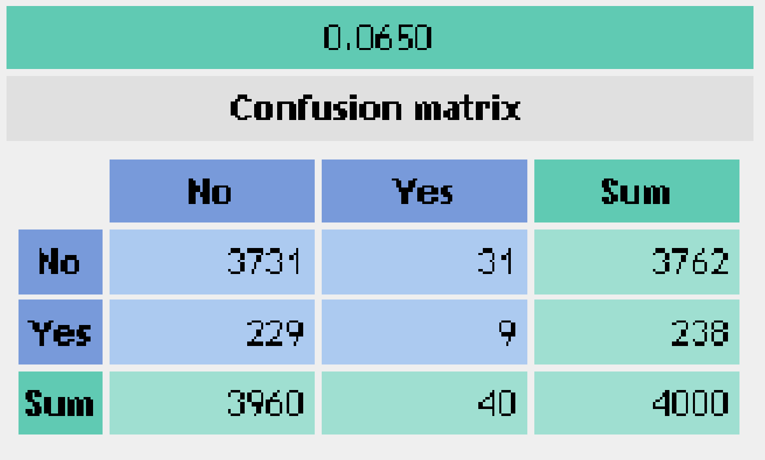
Case-II: Imbalanced Dataset

- When the learning process deals with class imbalance, the confusion matrix and the error rate do not provide a good idea about the classifier relevance.

- Dataset with 4000 observations
  - 3762 are of class No
  - 238 are of class Yes

Case-II: Imbalanced Dataset

- Confusion matrix of, say, model M1. The error rate of model M1 is 0.0650



0.0650

Confusion matrix

	No	Yes	Sum
No	3731	31	3762
Yes	229	9	238
Sum	3960	40	4000

- Default Model: It predicts the most frequent class as prediction. The test error rate of the default classifier is 238 / 4000 = 0.0595

Case-II: Imbalanced Dataset

- The default classifier seems to be the best in class imbalance situation

- This anomaly is due to the necessity to predict the class value, using a specific cutoff threshold.

- Yet, in numerous domains, the most interesting is to measure the propensity to be a positive class value (the class of interest - e.g. the propensity to purchase a product, the propensity to fail for a credit applicant, etc.)

Case-II: Imbalanced Dataset

- Specifying the cutoff threshold is often difficult since it can vary according to the circumstances.

- Should we try a large number of cutoffs to determine the performance of model?

- Can we use a tool which allows us to estimate the performance of model regardless of the cutoff threshold?

Alternative to Error rate/Accuracy

Alternative metrics

- ROC(Receiver Operating Characteristic) curve:
    - It plots a curve between FPR and Recall measures
- PR curve:
    - It plots a curve between Recall and Precision measures

Precision, Recall & FPR metrics

- Recall = of those that are known as positive, how many are predicted positive?

  Recall = TP / (All known positives) = TP / (TP + FN)

- Precision = of those that are predicted positive, how many are positive?

  Precision = TP / (All predicted positives) = TP / (TP + FP)

- FPR = of those that are known as negative, how many are predicted positive?

  FPR = FP / (All known negatives) = FP / (TN + FP)

ROC curve

- The ROC curve is a tool for the performance evaluation and the comparison of classifiers

- Its scope goes beyond to the interpretations provided by the analysis of the confusion matrix (which depends on the cutoff threshold used)

ROC curve: Benefits

- It does not depend on the misclassification costs matrix. It enables us to know if M1 (or M2) dominates M2 (or M1) for whatever the misclassification costs matrix

- It is valid even in the case of imbalanced classes. It allows us to determine the performance of model irrespective of cutoff threshold values using AUC measure

ROC curve: Benefits

- It provides a graphical tool which enables to compare classifiers. We know immediately which are the interesting classifiers

- The results are relevant when the test sample is not representative Even if the classes distribution of the test set do not provide a good estimation of the prior probability of classes

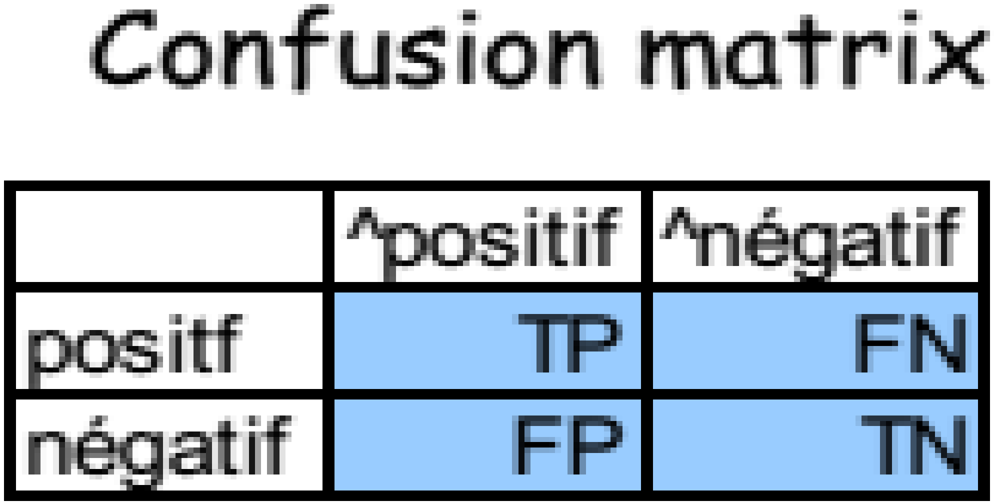
When to use ROC curve?

- It is applicable for only binary predictive classification problems: Y = {+, -}. The “+” value is the target class

- It is applicable for classifiers who can provide an estimate of P(Y=+/X) Or any SCORE that indicates the propensity to be "+" (which allows to sort the instances)

How to draw ROC curves?

How to draw ROC curve



Confusion matrix

	^positif	^négatif
positf	TP	FN
négatif	FP	TN

- TPR (True Positive Rate) = TP / Positives
- FPR (False Positive Rate) = FP / Negatives

TPR vs FPR

- TPR = of those that are known as positive, how many are predicted positive?
  - the higher TPR, the fewer positive data points we will miss.

- FPR = of those that are known as negative, how many are predicted positive?
  - the higher FPR, the more negative data points we will missclassified.

How to draw ROC curve

- P(Y=+/X) >= P(Y=-/X) is equivalent to the decision rule P(Y=+/X) >= 0.5 (threshold = 0.5). This decision rule provides a confusion matrix MC(1) with TPR(1) and FPR(1)

- If we use another threshold (e.g. 0.6), we obtain another confusion matrix MC(2) with TPR(2) and FPR(2)

- By varying the threshold, we have a succession of confusion matrices MC(i), for which we can calculate TPR(i) and FPR(i). The ROC curve is a scatter plot with FPR on the x-axis, and TPR on y-axis.

Sort the instances according to the score value (in descending order)

Individu	Score (+)	Classe
1	1	+
2	0.95	+
3	0.9	+
4	0.85	-
5	0.8	+
6	0.75	-
7	0.7	-
8	0.65	+
9	0.6	-
10	0.55	-
11	0.5	-
12	0.45	+
13	0.4	-
14	0.35	-
15	0.3	-
16	0.25	-
17	0.2	-
18	0.15	-
19	0.1	-
20	0.05	-

Positives = 6
Negatives = 14

Cut = 1

	^positif	^négatif	Total
positf	1	5	6
négatif	0	14	14
Total	1	19	20

TPR = 1/6 = 0.2 ; FPR = 0/14 = 0

Cut = 0.95

	^positif	^négatif	Total
positf	2	4	6
négatif	0	14	14
Total	2	18	20

TPR = 2/6 = 0.33 ; FPR = 0/14 = 0

Cut = 0.9

	^positif	^négatif	Total
positf	3	3	6
négatif	0	14	14
Total	3	17	20

TPR = 3/6 = 0.5 ; FPR = 0/14 = 0

Cut = 0.85

	^positif	^négatif	Total
positf	3	3	6
négatif	1	13	14
Total	4	16	20

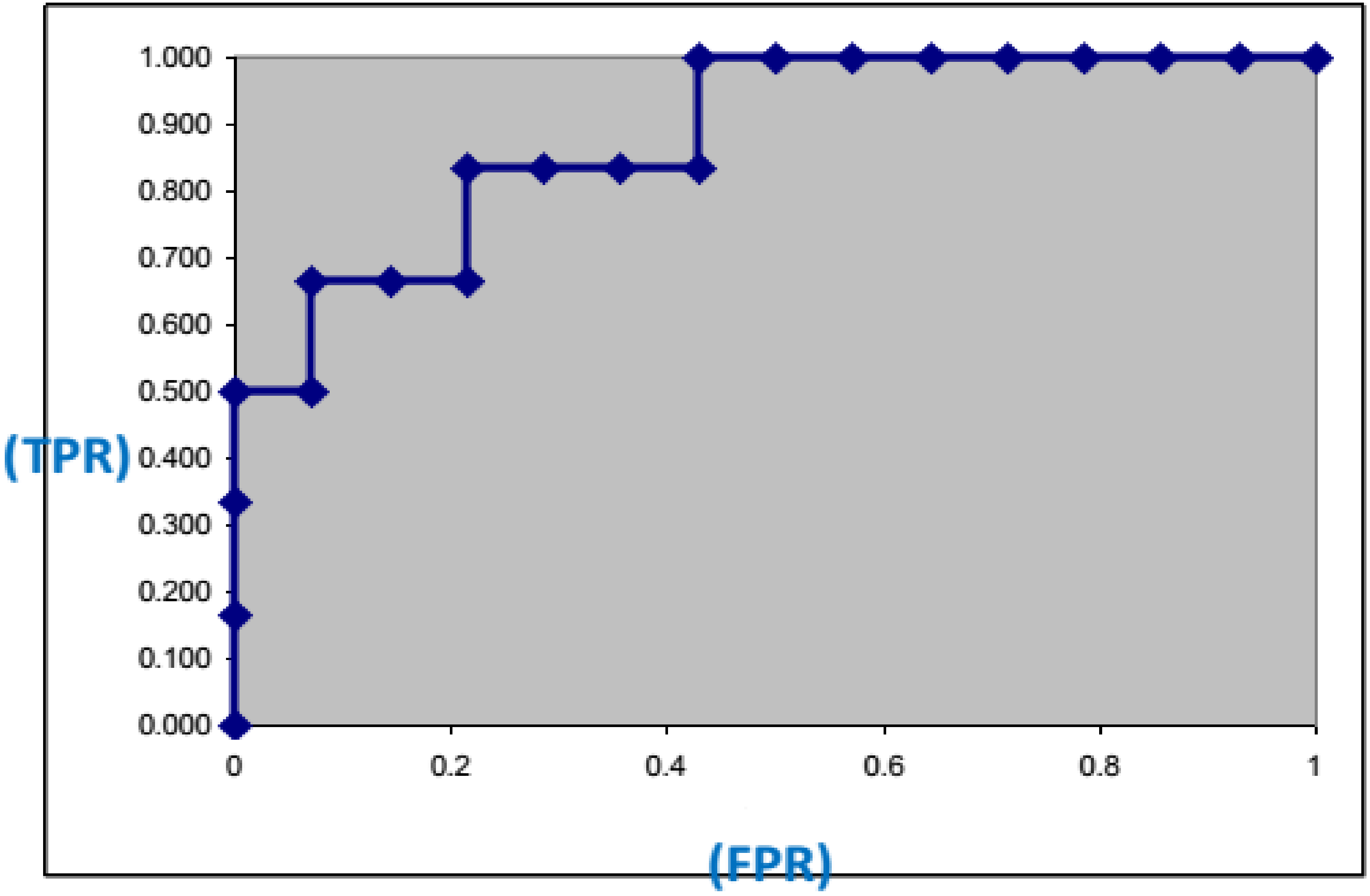
TPR = 3/6 = 0.5 ; FPR = 1/14 = 0.07

Cut = 0

	^positif	^négatif	Total
positf	6	0	6
négatif	14	0	14
Total	20	0	20

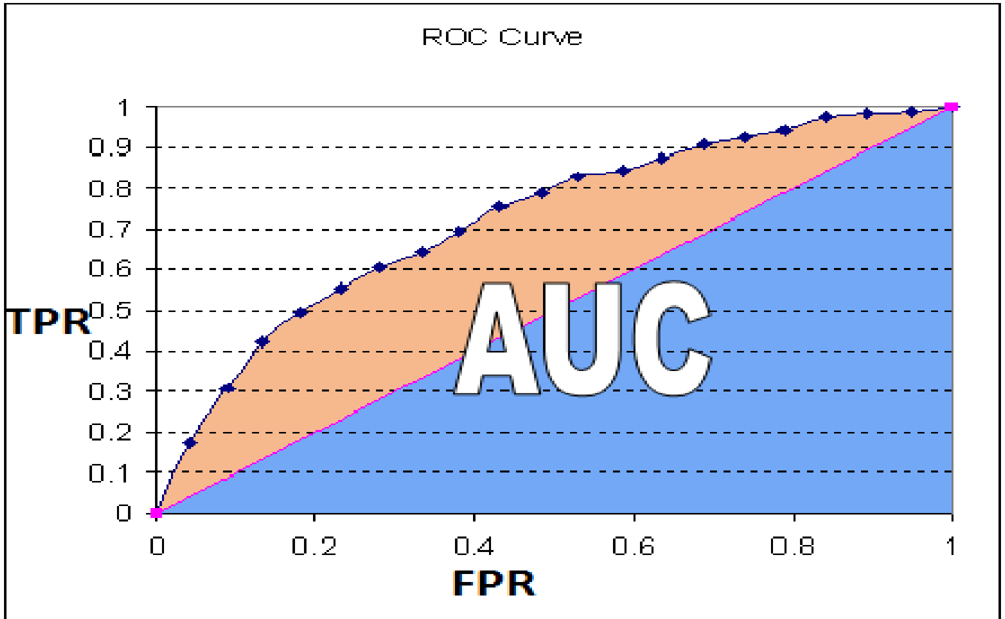
TPR = 6/6 = 1 ; FPR = 14/14 = 1

ROC curve


1.000
0.900
0.800
0.700
0.600
0.500
(TPR)
0.400
0.300
0.200
0.100
0.000
0
0.2
0.4
0.6
0.8
1
(FPR)

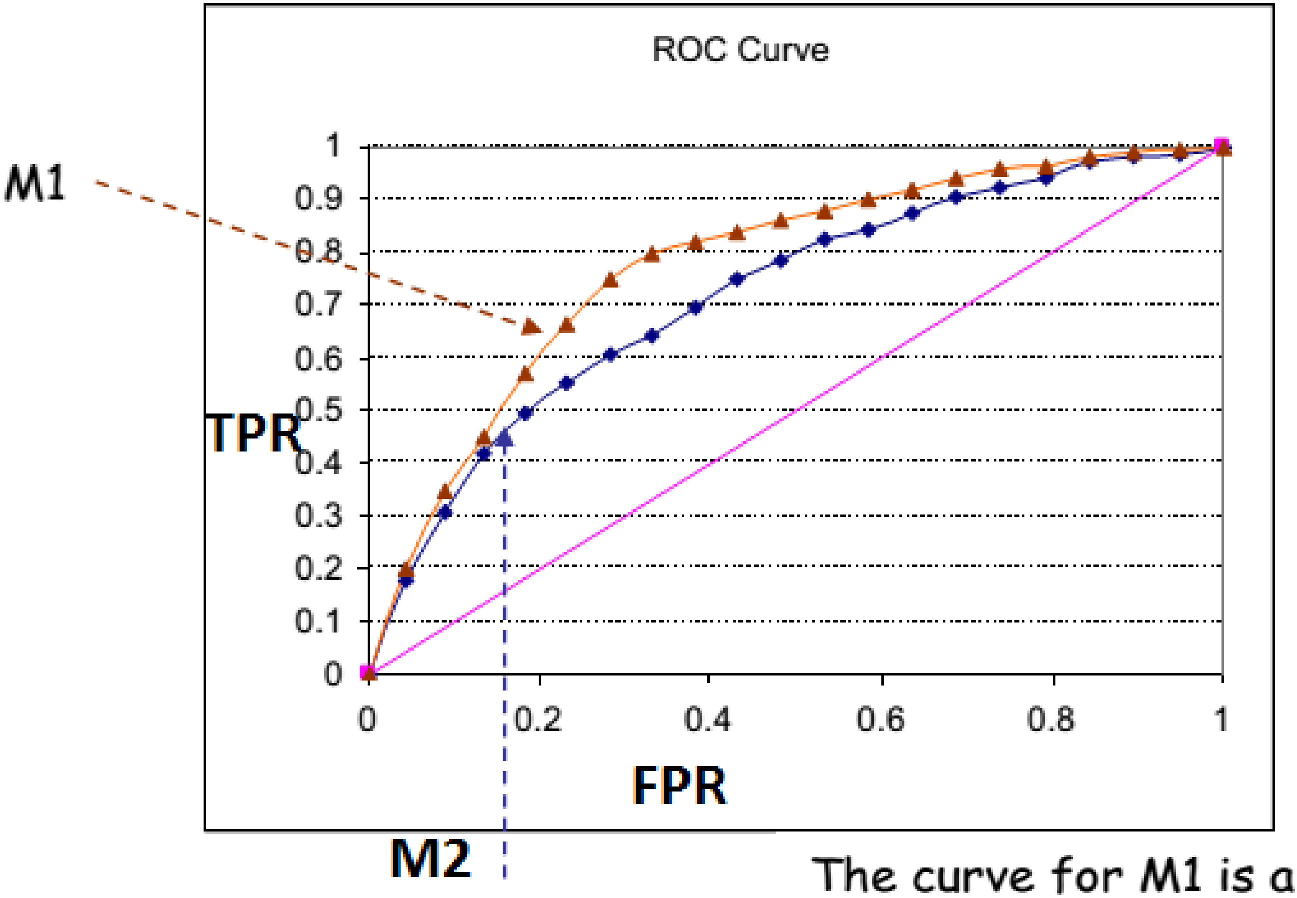
Area Under Curve(AUC)

- AUC corresponds to the probability of a positive instance to have a higher score than a negative instance (best situation AUC = 1)
- If the SCORE is assigned randomly to the individuals (the classifier is not better than random classifier), AUC = 0.5 This is the diagonal line in the graphical representation



Model Selection

How to show that the classifier M1 is always better than M2 whatever the misclassification costs matrix used?



The curve for M1 is always above to the one of M2: there is no situation (misclassification costs matrix) for which M2 would be better than M1.

Summary

- In many applications, the ROC curve provides more interesting information than the error rate. This is especially true when we deal with a non representative test sample; in the case of imbalanced classes; when the misclassification costs are not well defined.

- The ROC curve is effective only in the binary problems; the classifier must provide a score function for the target class P(Y = + /X) (or, at least, the propensity to be positive)

- Some extensions of the ROC principle to multiclass classification problems exist but they often have a lack of simplicity, reducing the interest of the tool.

Other Confusion matrix based metrics

- Recall = of those that known as positive, how many did you predicted positive?

  Recall = TP / (All known positives) = TP / (TP + FN)

- Precision = of those that predicted positive, how many are correct?

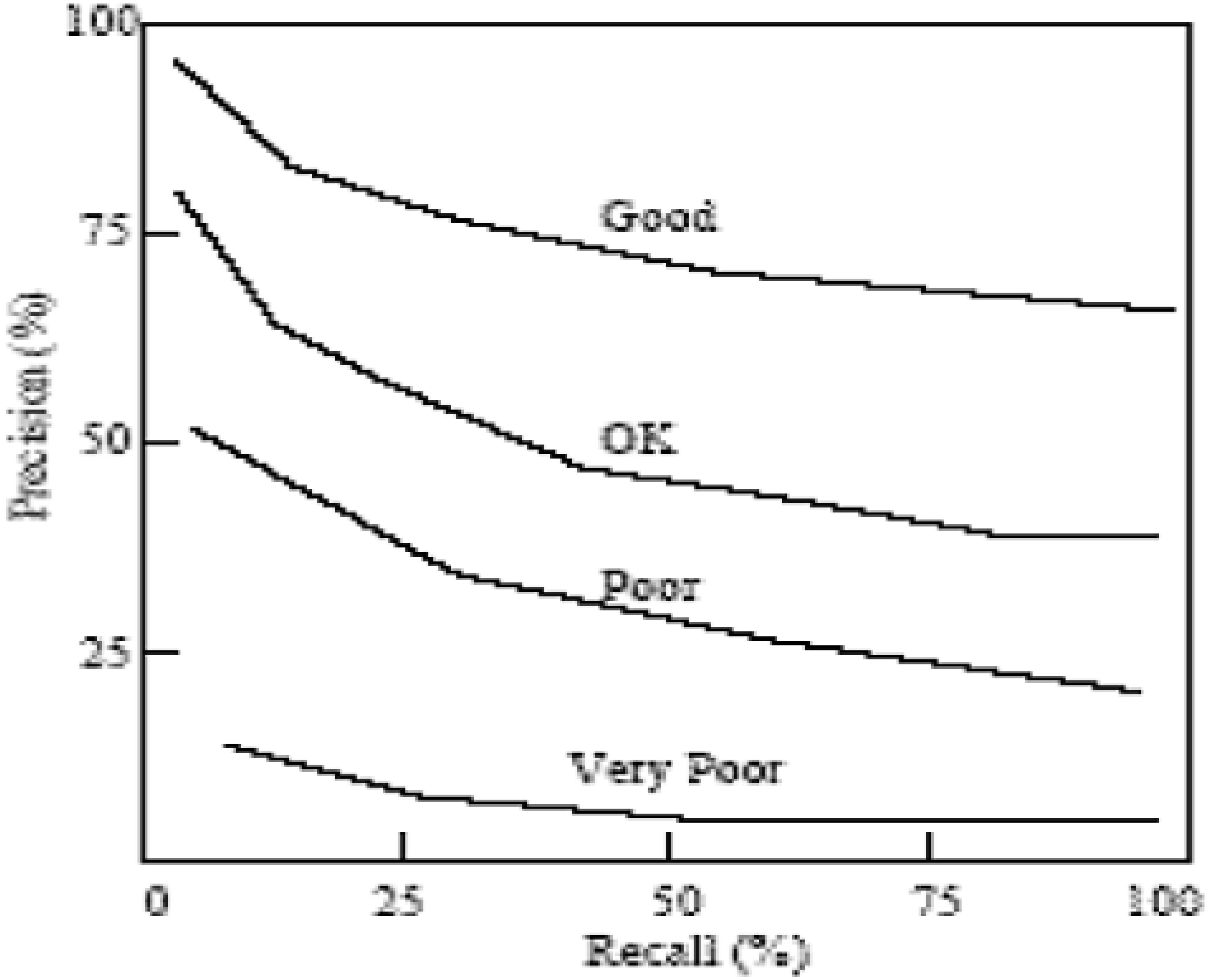
  Precision = TP / (All predicted positives) = TP / (TP + FP)

- Useful when notion of "negative" (and hence FPR) is not well defined

- F-score is harmonic mean:

$$F = \frac{2}{1/P + 1/R} = \frac{2PR}{R + P}$$

PR curve


100
75
50
25
Precision (%)
Good
OK
Poor
Very Poor
0
25
50
75
100
Recall (%)

Regression Metrics

- Root Mean Squared Error(RMSE)
- R^2

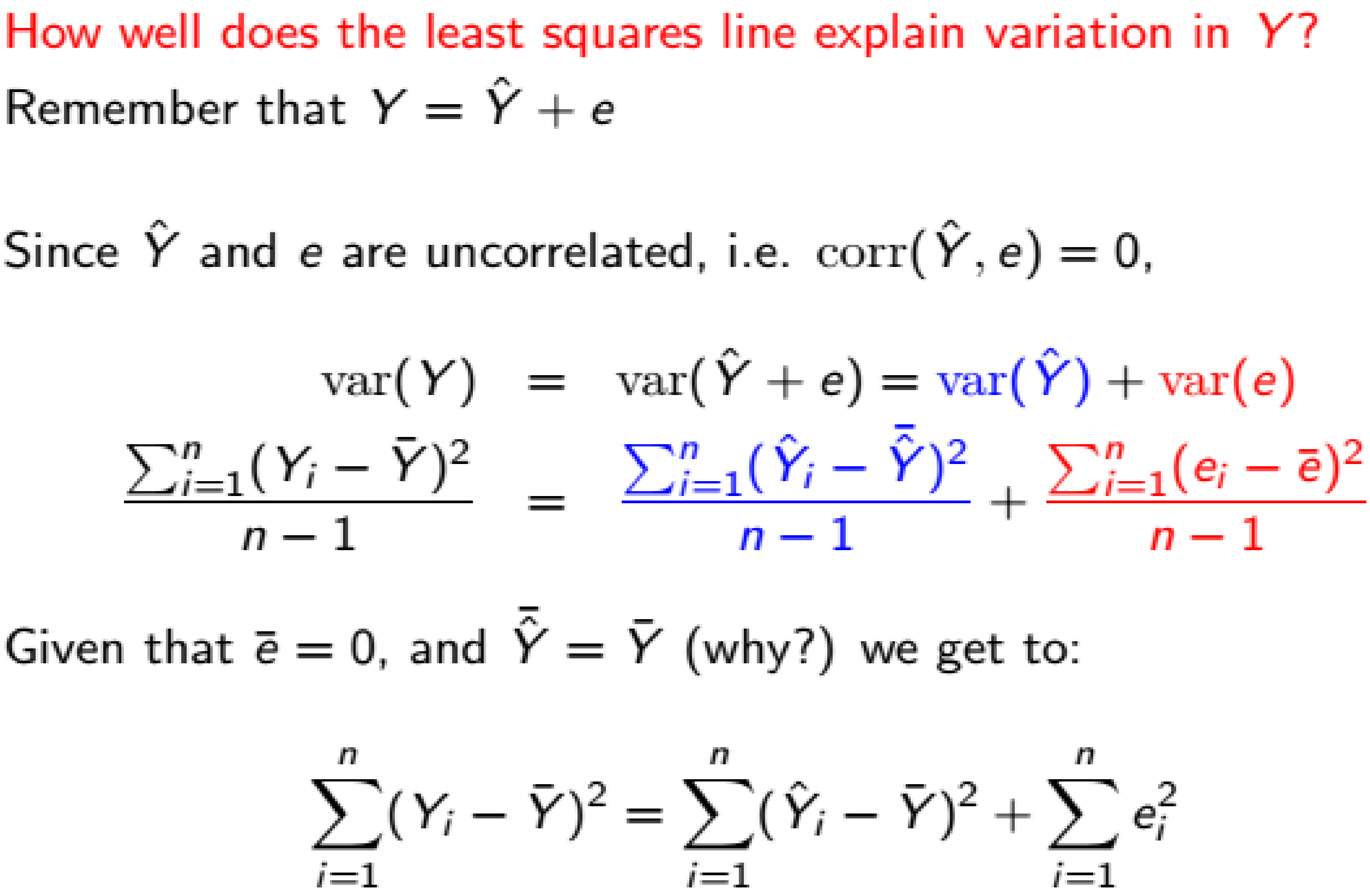
RMSE

- It is based on Residuals/Errors of each sample.
- RMSE = The square root of average squared residuals.
- It represents average distance between the observed and predicted values.

R^2 : Capturing Variance

If model captures the total variance of target variable then it is considered as best model. How much of the total variance captured/explained by model can be measured by R^2 metric.

Expression for R^2



How well does the least squares line explain variation in Y?

Remember that $Y = \hat{Y} + e$

Since $\hat{Y}$ and e are uncorrelated, i.e. $\mathrm{corr}(\hat{Y}, e) = 0$,

$$\begin{aligned} \mathrm{var}(Y) &= \mathrm{var}(\hat{Y} + e) = \mathrm{var}(\hat{Y}) + \mathrm{var}(e) \\ \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} &= \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2}{n-1} + \frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-1} \end{aligned}$$

Given that $\bar{e} = 0$, and $\bar{\hat{Y}} = \bar{Y}$ (why?) we get to:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} e_i^2$$

Expression for R^2

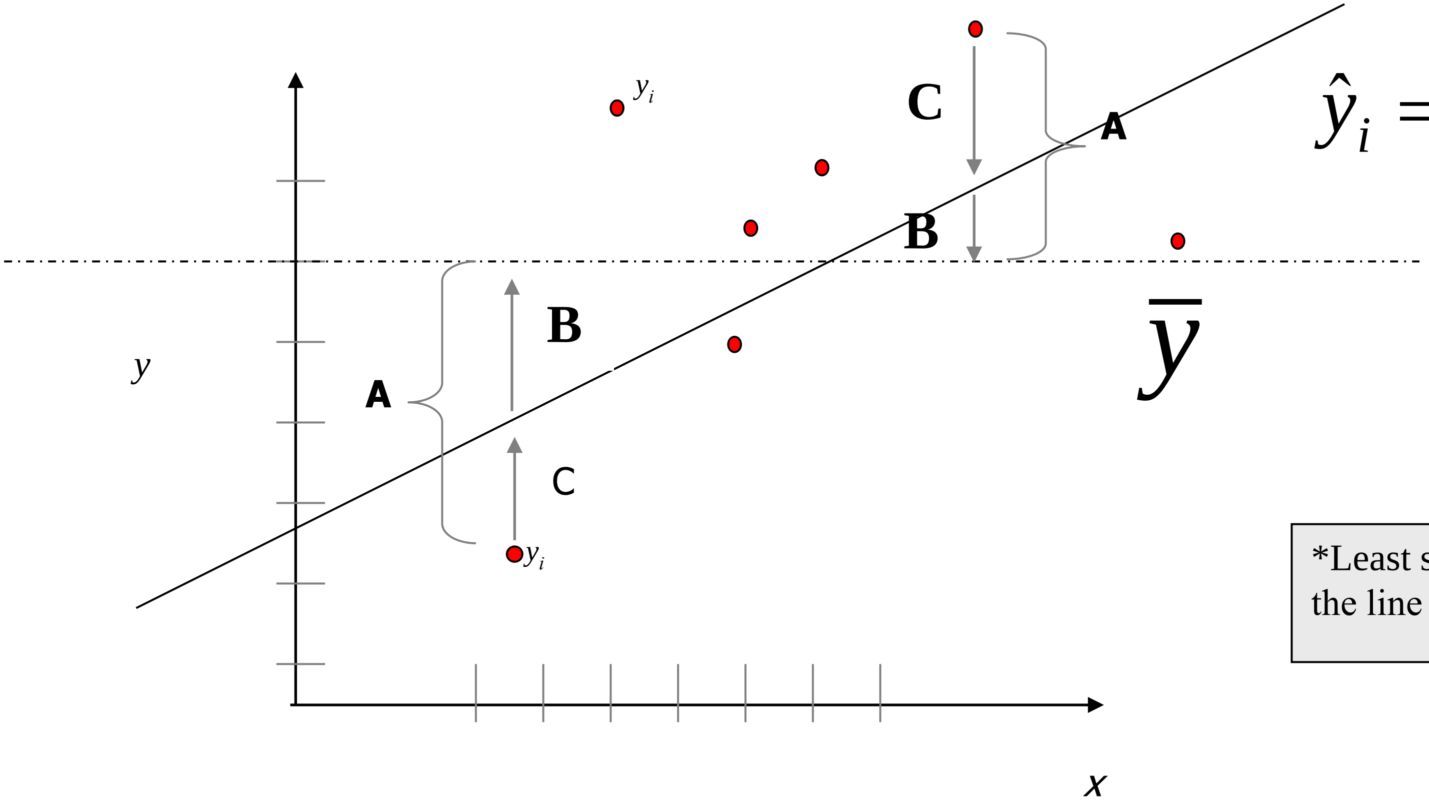
$$\underbrace{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}_{\boxed{\begin{array}{c}\textbf{Total Sum of}\\ \textbf{Squares}\\ \textbf{SST}\end{array}}}=\underbrace{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}_{\boxed{\begin{array}{c}\textbf{Regression SS}\\ \textbf{SSR}\end{array}}}+\underbrace{\sum_{i=1}^{n}e_i{}^2}_{\boxed{\begin{array}{c}\textbf{Error SS}\\ \textbf{SSE}\end{array}}}$$

SSR: Variation in Y explained by the regression line.

SSE: Variation in Y that is left unexplained.

$$\text{SSR} = \text{SST} \Rightarrow \text{perfect fit.}$$

Expression for R^2



$$\bar{y}$$

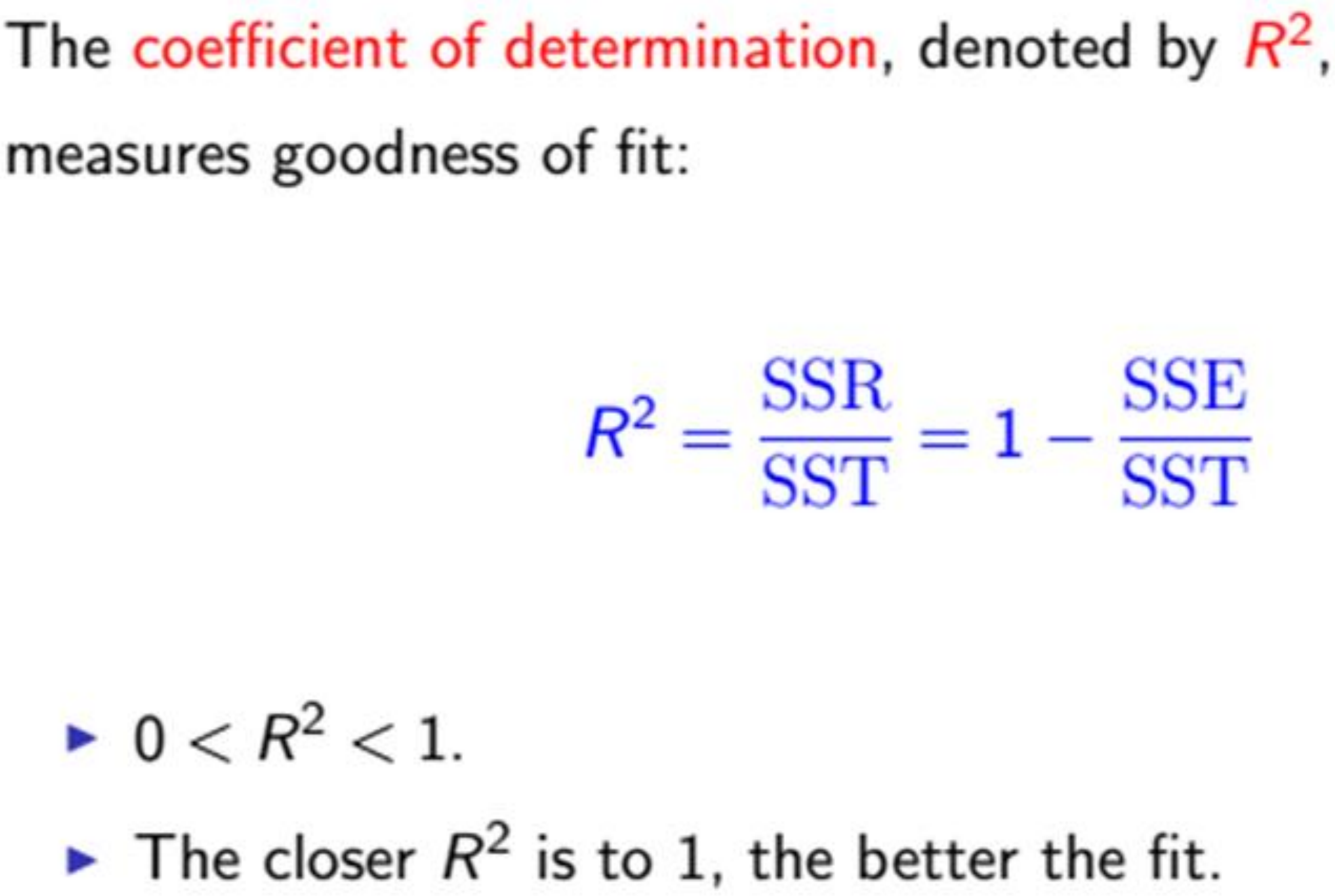
$$\hat{y}_i = \beta x_i + \alpha$$

*Least squares estimation gave us the line (β) that minimized C^2

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

A^2 = SST **B^2 = SSR** C^2 = SSE

Expression for R^2



The coefficient of determination, denoted by R^2, measures goodness of fit:

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$$

- $0 < R^2 < 1$.
- The closer R^2 is to 1, the better the fit.

RMSE vs R^2

Model1: RMSE=1 and variance = 4.2

R^2=76%

Model2: RMSE=1 and variance=2

R^2=67%