Machine Learning for Official Statistics and SDGs

Second Live Lecture (Webinar): Starts in **15** minutes

Christophe Bontemps, UN SIAP

UNITED NATIONS
SIAP
Statistical Institute for Asia and the Pacific

Machine Learning for Official Statistics and SDGs

Second Live Lecture (Webinar):

Starts in **10** minutes

Christophe Bontemps, UN SIAP

UNITED NATIONS siap Statistical Institute for Asia and the Pacific

Machine Learning for Official Statistics and SDGs

Second Live Lecture (Webinar):

Starts in **5** minutes

Christophe Bontemps, UN SIAP

UNITED NATIONS siap Statistical Institute for Asia and the Pacific

Machine Learning for Official Statistics and SDGs

Classification

UNITED NATIONS SIAP
Statistical Institute for Asia and the Pacific

[- Reminder -]

- Mute yourself always!

[- Reminder -]

- Mute yourself always!
- The lecture is recorded

[- Reminder -]

- Mute yourself always!
- The lecture is recorded
- Ask questions in the chat

[- Agenda -]

- Introduction

[- Agenda -]

- Introduction
- Classification in a (*Machine learning Framework*)

[- Agenda -]

- Introduction
- Classification in a (*Machine learning Framework*)
- Q&A

[- Agenda -]

- Introduction
- Classification in a (*Machine learning Framework*)
- Q&A
- Next week

[Classification]

What is a classification problem?

[Classification]

What is a classification problem?

- The goal is to understand why an observation belongs to a certain category

[Classification]

What is a classification problem?

- The goal is to understand why an observation belongs to a certain category
- y takes discrete values: 0/1, high school/primary school/no education; urban/rural

[Classification]

What is a classification problem?

- The goal is to understand why an observation belongs to a certain category
- y takes discrete values: 0/1, high school/primary school/no education; urban/rural
- Some variables xs may explain why y belongs to a particular category

[Classification]

What is a classification problem?

- The goal is to understand why an observation belongs to a certain category
- y takes discrete values: 0/1, high school/primary school/no education; urban/rural
- Some variables xs may explain why y belongs to a particular category

[Classification]

What is a classification problem?

- The goal is to understand why an observation belongs to a certain category
- y takes discrete values: 0/1, high school/primary school/no education; urban/rural
- Some variables xs may explain why y belongs to a particular category

A **classifier** is a tool that provides a classification for y using (*or not*) additional information from other variables

[Supervised *vs* unsupervised Classification]

[Supervised *vs* unsupervised Classification]

- In **supervised** classification, we **observe** the category for each observation

[SUPERVISED *vs* UNSUPERVISED CLASSIFICATION]

- In **supervised** classification, we **observe** the category for each observation

  One may learn and estimate the impact of other variables on that classification (e.g. logit regression)

[SUPERVISED *vs* UNSUPERVISED CLASSIFICATION]

- In **supervised** classification, we **observe** the category for each observation

  One may learn and estimate the impact of other variables on that classification (e.g. logit regression)
- In **unsupervised** classification, we **ignore** the category (if any) of each observation

[Supervised *vs* unsupervised Classification]

- In **supervised** classification, we **observe** the category for each observation

  One may learn and estimate the impact of other variables on that classification (e.g. logit regression)

- In **unsupervised** classification, we **ignore** the category (if any) of each observation

  The goal is to classify observations from those variables (clustering) without having any information of what a category means.

[Supervised *vs* unsupervised Classification]

- In **supervised** classification, we **observe** the category for each observation

  One may learn and estimate the impact of other variables on that classification (e.g. logit regression)

- In **unsupervised** classification, we **ignore** the category (if any) of each observation

  The goal is to classify observations from those variables (clustering) without having any information of what a category means.

[SUPERVISED *vs* UNSUPERVISED CLASSIFICATION]

- In **supervised** classification, we **observe** the category for each observation

  One may learn and estimate the impact of other variables on that classification (e.g. logit regression)

- In **unsupervised** classification, we **ignore** the category (if any) of each observation

  The goal is to classify observations from those variables (clustering) without having any information of what a category means.

- We'll focus on **supervised** classification

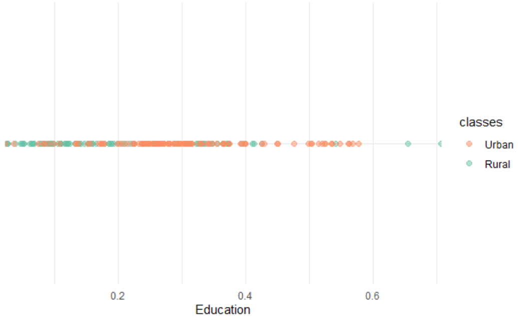
[Classification: An example]

[Classification: An example]

- You observe households in *Urban* or *Rural* and *Education*.

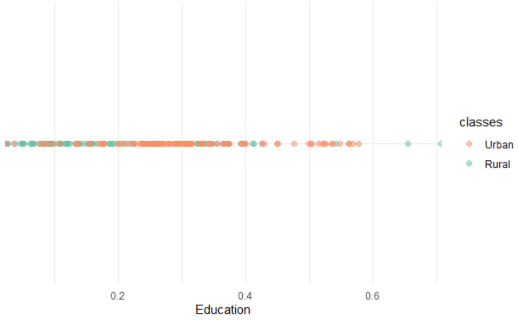
[Classification: An example]

- You observe households in *Urban* or *Rural* and *Education*.



[Classification: An example]

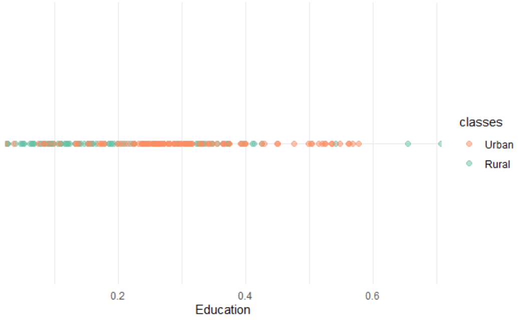
- You observe households in *Urban* or *Rural* and *Education*.



- A classifier "finds" the value of *Education* separating "*Rural*" from "*Urban*"

[Classification: An example]

- You observe households in *Urban* or *Rural* and *Education*.



- A classifier "finds" the value of *Education* separating "*Rural*" from "*Urban*"
  Typically with a threshold rule: "if $x \geq T_0$ then category is *Urban*"

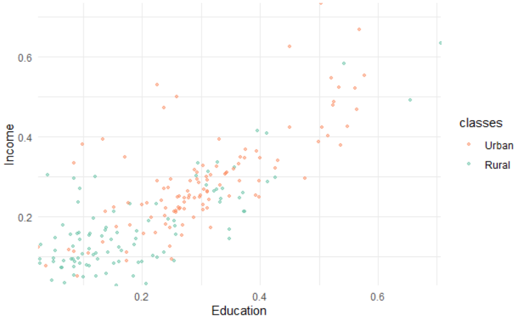
[Classification: A 2-D example]

[CLASSIFICATION: A 2-D EXAMPLE]

- You observe households in *Urban* or *Rural* areas and **two** variables (features): *Education* and *Income*

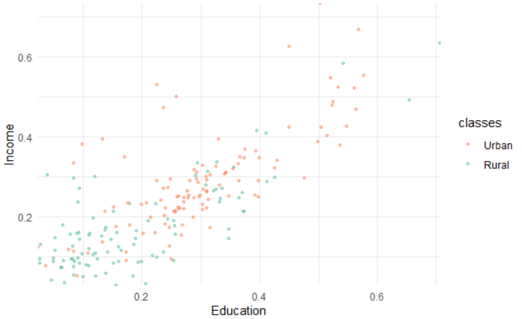
[Classification: A 2-D example]

- You observe households in *Urban* or *Rural* areas and **two** variables (features): *Education* and *Income*



[Classification: A 2-D example]

- You observe households in *Urban* or *Rural* areas and **two** variables (features): *Education* and *Income*



Where is the boundary? How to find it?

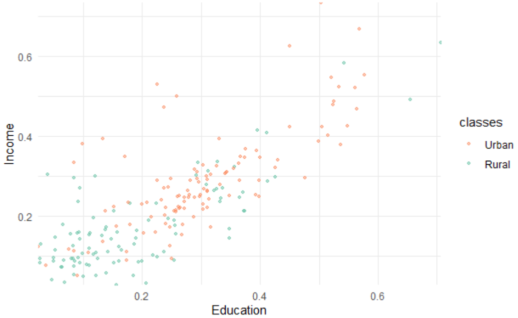
[Classification: A 2-D example]

[Classification: A 2-D example]

- A classifier will determine a **boundary** using both *Education* and *Income* to separate "Rural" from "Urban"

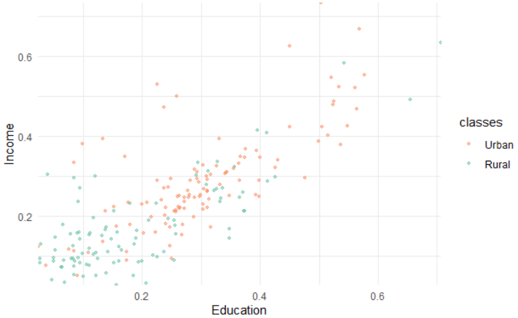
[Classification: A 2-D example]

- A classifier will determine a **boundary** using both *Education* and *Income* to separate "Rural" from "Urban"



[Classification: A 2-D example]

- A classifier will determine a **boundary** using both *Education* and *Income* to separate "Rural" from "Urban"



- The rule can be based on a linear relationship between *Education* and *Income* or can be non linear.

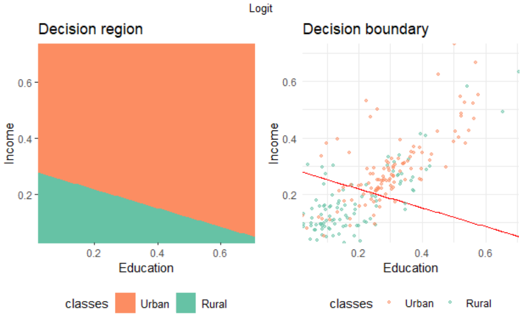
[Classification: A 2-D example]

[Classification: A 2-D example]

- Example of a linear classifier

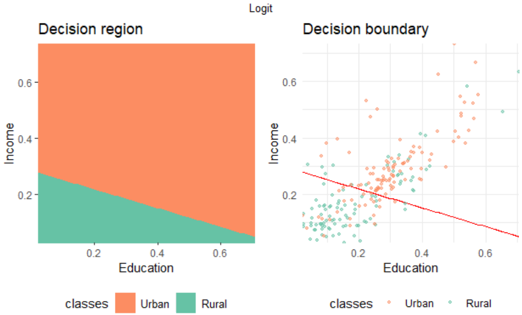
[Classification: A 2-D example]

- Example of a linear classifier



[Classification: A 2-D example]

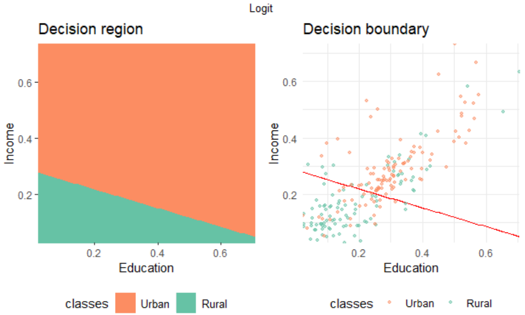
- Example of a linear classifier



- The separation rule is $x'\beta \geq T_0$ for a particular T_0: the ***threshold***

[Classification: A 2-D example]

- Example of a linear classifier



- The separation rule is $x'\beta \geq T_0$ for a particular T_0: the ***threshold***

  e.g.: $\beta_0 + \beta_1 Education + \beta_2 Income \geq T_0 \Leftrightarrow Urban$

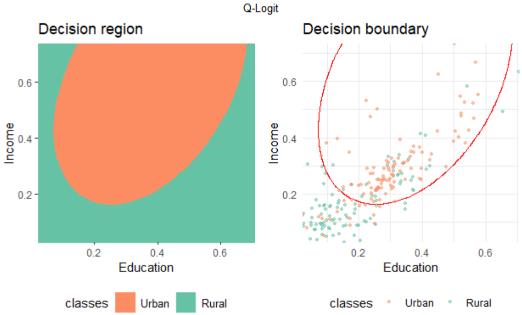
[Classifiers examples]

[Classifiers examples]

- Example of non-linear classifier

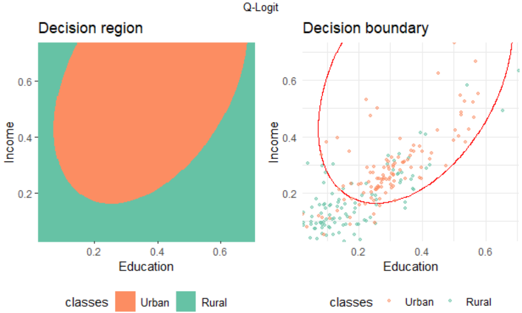
[Classifiers examples]

- Example of non-linear classifier



[Classifiers examples]

- Example of non-linear classifier



- The rule that separated the two classes is non linear in the variables *Education* and *Income*

[How to select the right model?]

[How to select the right model?]

- What is the goal?

[How to select the right model?]

- What is the goal?

  Have the “best‘” classification

[How to select the right model?]

- What is the goal?

  Have the “best‘” classification

↪ Need for a criterion to determine what is a good classifier

[How to select the right model?]

- What is the goal?

  Have the “best‘” classification

$\hookrightarrow$ Need for a criterion to determine what is a good classifier

- Measures of fit in classification are different and specific

[Measures of fit in classification]

There are several popular measures of fit, differing in their spirit and their goal

[Measures of fit in classification]

There are several popular measures of fit, differing in their spirit and their goal

- Accuracy

[Measures of fit in classification]

There are several popular measures of fit, differing in their spirit and their goal

- Accuracy
- Confusion matrix

[Measures of fit in classification]

There are several popular measures of fit, differing in their spirit and their goal

- Accuracy
- Confusion matrix
- Sensitivity & Specificity

[MEASURES OF FIT IN CLASSIFICATION]

There are several popular measures of fit, differing in their spirit and their goal

- Accuracy
- Confusion matrix
- Sensitivity & Specificity
- Kappa

[Measures of fit in classification]

There are several popular measures of fit, differing in their spirit and their goal

- Accuracy
- Confusion matrix
- Sensitivity & Specificity
- Kappa

...

[Measures of fit in classification]

There are several popular measures of fit, differing in their spirit and their goal

- Accuracy
- Confusion matrix
- Sensitivity & Specificity
- Kappa

...

Each criterion answers to a different question

[Accuracy and Confusion Matrix]

Accuracy corresponds to the probability of being "accurate"

$$\Pr\left[y_0 = \widehat{f}(x_0)\right]$$

[Accuracy and Confusion Matrix]

Accuracy corresponds to the probability of being "accurate"

$$\Pr\left[y_0 = \widehat{f}(x_0)\right]$$

- where $\widehat{f}(\cdot)$ is the classifier.

[Accuracy and Confusion Matrix]

Accuracy corresponds to the probability of being "accurate"

$$\Pr\left[y_0 = \widehat{f}(x_0)\right]$$

- where $\widehat{f}(\cdot)$ is the classifier.

$\hookrightarrow$ We want the **maximum** possible accuracy.

[Accuracy and Confusion Matrix]

Accuracy corresponds to the probability of being "accurate"

$$\Pr\left[y_0 = \widehat{f}(x_0)\right]$$

- where $\widehat{f}(\cdot)$ is the classifier.

↪ We want the **maximum** possible accuracy.

- Equivalently, we may want to **minimize** the *error rate* or *misclassification rate*

$$\Pr\left[y_0 \neq \widehat{f}(x_0)\right]$$

[Confusion Matrix & Accuracy]

A classifier predicts in which class each observation should be:

[Confusion Matrix & Accuracy]

A classifier predicts in which class each observation should be:

	Observed (True)	
Predicted	TP (True Positive)	FP (False Positive)
	FN (False Negative)	TN (True Negative)

Table: Confusion Matrix

[Confusion Matrix & Accuracy]

A classifier predicts in which class each observation should be:

	Observed (True)	
Predicted	TP (True Positive)	FP (False Positive)
	FN (False Negative)	TN (True Negative)

Table: Confusion Matrix

- Accuracy is then the ratio:

$$Acuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$= \frac{TruePositives + TrueNegatives}{N}$$

[Confusion Matrix & Accuracy]

A classifier predicts in which class each observation should be:

	Observed (True)	
Predicted	TP (True Positive)	FP (False Positive)
	FN (False Negative)	TN (True Negative)

Table: Confusion Matrix

- Accuracy is then the ratio:

$$Acuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
$$= \frac{TruePositives + TrueNegatives}{N}$$

- It is the proportion of accurate predictions

[Confusion Matrix & Accuracy]

In practice, with a classifier we have:

		Observed (True)	
		Urban	Rural
Predicted	Urban	87 (TP)	28 (FP)
	Rural	24 (FN)	69 (TN)

Table: Confusion Matrix

[Confusion Matrix & Accuracy]

In practice, with a classifier we have:

		Observed (True)	
		Urban	Rural
Predicted	Urban	87 (TP)	28 (FP)
	Rural	24 (FN)	69 (TN)

Table: Confusion Matrix

- Here *Urban* is the "positive" class

[Confusion Matrix & Accuracy]

In practice, with a classifier we have:

		Observed (True) Urban	Observed (True) Rural
Predicted	Urban	87 (TP)	28 (FP)
Predicted	Rural	24 (FN)	69 (TN)

Table: Confusion Matrix

- Here *Urban* is the "positive" class
- Accuracy is then the ratio:

$$Accuracy = \frac{87+69}{87+69+28+24}$$
$$= \frac{156}{208} = 0.75$$

[Confusion Matrix & Accuracy]

In practice, with a classifier we have:

		Observed (True)	
		Urban	Rural
Predicted	Urban	87 (TP)	28 (FP)
	Rural	24 (FN)	69 (TN)

Table: Confusion Matrix

- Here *Urban* is the "positive" class
- Accuracy is then the ratio:

$$Accuracy = \frac{87+69}{87+69+28+24} = \frac{156}{208} = 0.75$$

- We have an accurate prediction in 75% of the cases.

[Problem 1: Accuracy is one number]

Accuracy is not the *panacea* and may be misleading

- One may be more interested in **correctly** predicting a particular outcome!

[Problem 1: Accuracy is one number]

Accuracy is not the *panacea* and may be misleading

- One may be more interested in **correctly** predicting a particular outcome!

↪ This is often the case if the **cost** of being wrong differ

[Problem 1: Accuracy is one number]

Accuracy is not the *panacea* and may be misleading

- One may be more interested in **correctly** predicting a particular outcome!

$\hookrightarrow$ This is often the case if the **cost** of being wrong differ

- One may need other measures, focused on one particular outcome

[Problem 1: Accuracy is one number]

Accuracy is not the *panacea* and may be misleading

- One may be more interested in **correctly** predicting a particular outcome!

↪ This is often the case if the **cost** of being wrong differ

- One may need other measures, focused on one particular outcome
- Compute *Sensitivity* & *Specificity* from the confusion matrix

[Problem 1: Accuracy is one number]

Accuracy is not the *panacea* and may be misleading

- One may be more interested in **correctly** predicting a particular outcome!

↪ This is often the case if the **cost** of being wrong differ

- One may need other measures, focused on one particular outcome
- Compute *Sensitivity* & *Specificity* from the confusion matrix
- They may go in different directions

[Sensitivity or *True Positive Rate*]

		Observed (True) Urban	Observed (True) Rural
Predicted	Urban	87 (TP)	28 (FP)
Predicted	Rural	24 (FN)	69 (TN)

Table: Confusion Matrix

[Sensitivity or *True Positive Rate*]

		Observed (True)	
		Urban	Rural
Predicted	Urban	87 (TP)	28 (FP)
	Rural	24 (FN)	69 (TN)

Table: Confusion Matrix

- *Sensitivity*: focuses on predicted positives (here *Urban*) *vs* observed positives

$$Sensitivity = \frac{TP}{TP + FN}$$

$$= \frac{87}{87 + 24} = 0.78$$

[Sensitivity or *True Positive Rate*]

		Observed (True) Urban	Observed (True) Rural
Predicted	Urban	87 (TP)	28 (FP)
Predicted	Rural	24 (FN)	69 (TN)

Table: Confusion Matrix

- *Sensitivity*: focuses on predicted positives (here *Urban*) *vs* observed positives

$$Sensitivity = \frac{TP}{TP + FN}$$

$$= \frac{87}{87 + 24} = 0.78$$

- On *Urban*, we correctly predict in 78% of the cases

[Specificity or *True Negative Rate*]

		Observed (True) Urban	Observed (True) Rural
Predicted	Urban	87 (TP)	28 (FP)
	Rural	24 (FN)	69 (TN)

Table: Confusion Matrix

[SPECIFICITY OR *True Negative Rate*]

		Observed (True)	
		Urban	Rural
Predicted	Urban	87 (TP)	28 (FP)
	Rural	24 (FN)	69 (TN)

Table: Confusion Matrix

- *Sensitivity* focuses on predicted negatives (*Rural*) *vs* observed negatives

$$Specificity = \frac{TN}{TN + FP}$$

$$= \frac{69}{69 + 28} = 0.71$$

[Specificity or *True Negative Rate*]

		Observed (True)	
		Urban	Rural
Predicted	Urban	87 (TP)	28 (FP)
	Rural	24 (FN)	69 (TN)

Table: Confusion Matrix

- *Sensitivity* focuses on predicted negatives (*Rural*) *vs* observed negatives

$$Specificity = \frac{TN}{TN + FP}$$

$$= \frac{69}{69 + 28} = 0.71$$

- On *Rural*, we predict correctly in **only** 71% of the cases

[Problem 2: Imbalanced outcomes]

Imagine you observe much more Urban than Rural

Observed (True)	
Urban	Rural
95	5

[PROBLEM 2: IMBALANCED OUTCOMES]

Imagine you observe much more Urban than Rural

Observed (True)	
Urban	Rural
95	5

- A "*stupid*" classifier predicting only *Urban* $\cdots$

[Problem 2: Imbalanced outcomes]

Imagine you observe much more Urban than Rural

Observed (True)	
Urban	Rural
95	5

- A "*stupid*" classifier predicting only *Urban* $\cdots$

		Observed (True)	
		Urban	Rural
Predicted	Urban	95 (TP)	5 (FP)
	Rural	0 (FN)	0 (TN)

[Problem 2: Imbalanced outcomes]

Imagine you observe much more Urban than Rural

Observed (True)	
Urban	Rural
95	5

- A "*stupid*" classifier predicting only *Urban* · · ·

		Observed (True)	
		Urban	Rural
Predicted	Urban	95 (TP)	5 (FP)
	Rural	0 (FN)	0 (TN)

· · · would have a very good *Accuracy* and *Sensitivity*

[Problem 2: Imbalanced outcomes]

Imagine you observe much more Urban than Rural

Observed (True)	
Urban	Rural
95	5

- A "*stupid*" classifier predicting only *Urban* $\cdots$

		Observed (True)	
		Urban	Rural
Predicted	Urban	95 (TP)	5 (FP)
	Rural	0 (FN)	0 (TN)

$\cdots$ would have a very good *Accuracy* and *Sensitivity*

Accuracy = (TP + TN)/ 100 = 95 %

[PROBLEM 2: IMBALANCED OUTCOMES]

Imagine you observe much more Urban than Rural

Observed (True)	
Urban	Rural
95	5

- A "*stupid*" classifier predicting only *Urban* $\cdots$

		Observed (True)	
		Urban	Rural
Predicted	Urban	95 (TP)	5 (FP)
	Rural	0 (FN)	0 (TN)

$\cdots$ would have a very good *Accuracy* and *Sensitivity*

Accuracy = (TP + TN)/ 100 = 95 %

Sensitivity = TP/(TP + FN) = 95/95 = **100 %**

[The Kappa (κ) index]

Kappa (κ) is defined to measure the accuracy with imbalanced classes

- Its formal definition is given by

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

[The Kappa (κ) index]

Kappa (κ) is defined to measure the accuracy with imbalanced classes

Its formal definition is given by

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

P_o is the **current classifier accuracy** which is compared here with the accuracy of an uniformed classifier P_e

[The Kappa (κ) index]

Kappa (κ) is defined to measure the accuracy with imbalanced classes

Its formal definition is given by

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

P_o is the **current classifier accuracy** which is compared here with the accuracy of an uniformed classifier P_e

P_e is the **accuracy of an uniformed classifier** that would operate purely by chance, using no information.

[The Kappa (κ) index]

Kappa (κ) is defined to measure the accuracy with imbalanced classes

Its formal definition is given by

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

P_o is the **current classifier accuracy** which is compared here with the accuracy of an uniformed classifier P_e

P_e is the **accuracy of an uniformed classifier** that would operate purely by chance, using no information.

NB: P_o is simple accuracy while P_e is more complex to compute.

[The Kappa (κ) index]

Kappa (κ) is defined to measure the accuracy with imbalanced classes

Its formal definition is given by

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

P_o is the **current classifier accuracy** which is compared here with the accuracy of an uniformed classifier P_e

P_e is the **accuracy of an uniformed classifier** that would operate purely by chance, using no information.

NB: P_o is simple accuracy while P_e is more complex to compute.

- The larger κ is, the better the model for a given distribution of classes in a data set

[Logit as You know it]

[Logit as You know it]

y is discrete ($y \in 0, 1$), so no *direct* linear relationship between y and the explanatory variables x (*i.e. Education, Income*)

[Logit as You know it]

y is discrete ($y \in 0, 1$), so no *direct* linear relationship between y and the explanatory variables x (*i.e. Education, Income*)

- Logit estimates the probabilities π ($\in [0, 1]$)

$$\pi = Probablity[y = 1]$$

[Logit as You know it]

y is discrete ($y \in 0, 1$), so no *direct* linear relationship between y and the explanatory variables x (*i.e. Education, Income*)

- Logit estimates the probabilities π ($\in [0, 1]$)

$$\pi = Probablity[y = 1]$$

- The definition of the logit model is:

$$\pi = Pr\,(y = 1) = F(x'\beta) = \frac{1}{1 + \exp(-x'\beta)}$$

[Logit as You know it]

y is discrete ($y \in 0, 1$), so no *direct* linear relationship between y and the explanatory variables x (*i.e. Education, Income*)

- Logit estimates the probabilities π ($\in [0, 1]$)

$$\pi = Probablity[y = 1]$$

- The definition of the logit model is:

$$\pi = Pr\,(y = 1) = F(x'\beta) = \frac{1}{1 + \exp(-x'\beta)}$$

$\hookrightarrow$ So basically: $log(\frac{\pi}{1-\pi}) = x'\beta$ where $\frac{\pi}{1-\pi}$ is the odd ratio $\in [0, \infty]$ with values indicating high or low probability that $y = 1$

[Logit as You know it]

y is discrete ($y \in 0, 1$), so no *direct* linear relationship between y and the explanatory variables x (*i.e. Education, Income*)

- Logit estimates the probabilities π ($\in [0, 1]$)

$$\pi = Probablity[y = 1]$$

- The definition of the logit model is:

$$\pi = Pr\,(y = 1) = F(x'\beta) = \frac{1}{1 + \exp(-x'\beta)}$$

$\hookrightarrow$ So basically: $log(\frac{\pi}{1-\pi}) = x'\beta$ where $\frac{\pi}{1-\pi}$ is the odd ratio $\in [0, \infty]$ with values indicating high or low probability that $y = 1$

$\hookrightarrow$ *"The logit models log of odd ratios as linear in x"*

[Logit as a classifier]

[Logit as a classifier]

- Once estimated, $\widehat{\pi}_i$ provide a simple rule for classification

[Logit as a classifier]

- Once estimated, $\widehat{\pi}_i$ provide a simple rule for classification

$$\widehat{\pi}_i > t_0 \Leftrightarrow \widehat{y}_i = 1$$

[Logit as a classifier]

- Once estimated, $\widehat{\pi}_i$ provide a simple rule for classification

$$\widehat{\pi}_i > t_0 \Leftrightarrow \widehat{y}_i = 1$$

Where t_0 is a **threshold** probability

[Logit as a classifier]

- Once estimated, $\widehat{\pi}_i$ provide a simple rule for classification

$$\widehat{\pi}_i > t_0 \Leftrightarrow \widehat{y}_i = 1$$

Where t_0 is a **threshold** probability

- If $t_0 = 1/2$ (default), then the rule is equivalent to:

$$x_i'\widehat{\beta} > 0 \Leftrightarrow \widehat{y}_i = 1$$

[Logit as a classifier]

- Once estimated, $\widehat{\pi}_i$ provide a simple rule for classification

$$\widehat{\pi}_i > t_0 \Leftrightarrow \widehat{y}_i = 1$$

Where t_0 is a **threshold** probability
- If $t_0 = 1/2$ (default), then the rule is equivalent to:

$$x_i'\widehat{\beta} > 0 \Leftrightarrow \widehat{y}_i = 1$$

- If $t_0 \neq 1/2$, then there exist a threshold T_0 such that :

$$x_i'\widehat{\beta} > T_0 \Leftrightarrow \widehat{y}_i = 1$$

[Logit as a classifier]

- Once estimated, $\widehat{\pi}_i$ provide a simple rule for classification

$$\widehat{\pi}_i > t_0 \Leftrightarrow \widehat{y}_i = 1$$

Where t_0 is a **threshold** probability

- If $t_0 = 1/2$ (default), then the rule is equivalent to:

$$x_i'\widehat{\beta} > 0 \Leftrightarrow \widehat{y}_i = 1$$

- If $t_0 \neq 1/2$, then there exist a threshold T_0 such that :

$$x_i'\widehat{\beta} > T_0 \Leftrightarrow \widehat{y}_i = 1$$

$\hookrightarrow$ The logit classifier depends on the linear combination of the x's

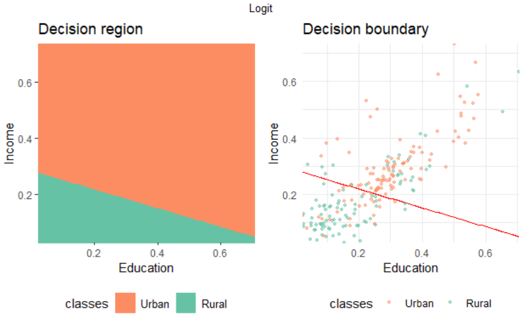
[Importance of the threshold]

[Importance of the threshold]

- The rule $x'\beta = T_0$ defines the partition of the space

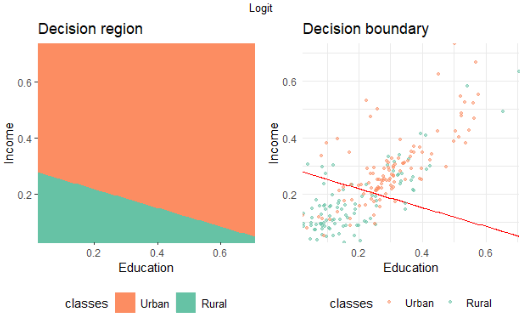
[Importance of the threshold]

- The rule $x'\beta = T_0$ defines the partition of the space



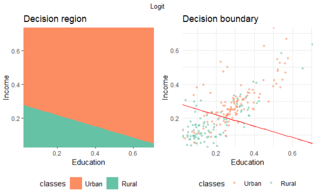
[IMPORTANCE OF THE THRESHOLD]

- The rule $x'\beta = T_0$ defines the partition of the space

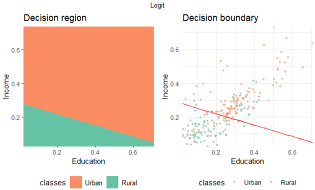


- This partition is sensitive to the choice of the threshold T_0 (and hence t_0)

[Importance of the threshold]



[Importance of the threshold]



- Changing t_0 will change the predictions & the classification
  A **higher** t_0 will allocate **less** observations to the $y = 1$ category (Urban)
  A **lower** t_0 will allocate **more** observations to the $y = 1$ category
- The choice of t_o should be done according to the data and observed classes repartition
- *Specificity* and *Sensitivity* are affected by t_0

[The ROC Curve]

[THE ROC CURVE]

- We want the *Specificity* and *Sensitivity* to be both **maximized** (ideally both would be 1)

[The ROC Curve]

- We want the *Specificity* and *Sensitivity* to be both **maximized** (ideally both would be 1)
- The ROC curve help visualize the best choice

[The ROC Curve]

- We want the *Specificity* and *Sensitivity* to be both **maximized** (ideally both would be 1)
- The ROC curve help visualize the best choice
- The ROC plots both Sensitivity and Specificity values for different thresholds

[The ROC Curve]

- We want the *Specificity* and *Sensitivity* to be both **maximized** (ideally both would be 1)
- The ROC curve help visualize the best choice
- The ROC plots both Sensitivity and Specificity values for different thresholds

↪ *Be careful of the axes*

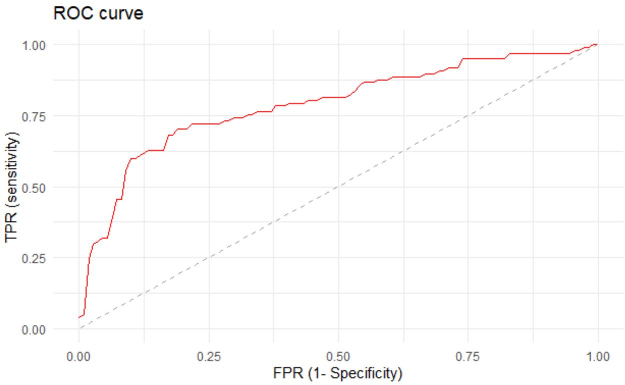
[The ROC Curve]

[The ROC Curve]

The ROC represents values of 1- Specificity = FPR *vs* Sensitivity = TPR for many values of the threshold t_0

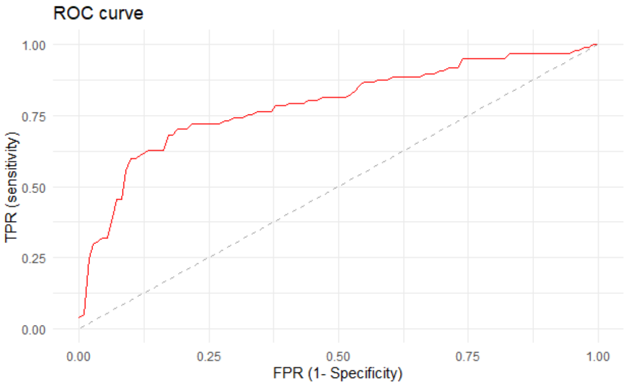
[The ROC Curve]

The ROC represents values of 1- Specificity = FPR *vs* Sensitivity = TPR for many values of the threshold t_0



[The ROC Curve]

The ROC represents values of 1- Specificity = FPR *vs* Sensitivity = TPR for many values of the threshold t_0



↪ *sometimes on a ROC curve, x is sensitivity with inverted x-axis*

[The ROC Curve: How to read?]

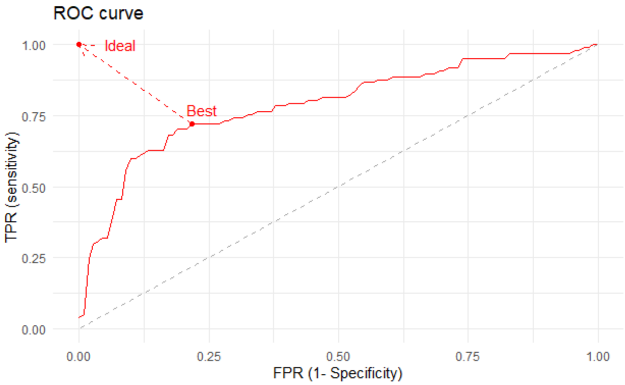
[The ROC Curve: How to read?]

[The ROC Curve: How to read?]

- Optimally, the curve should touch top-left corner

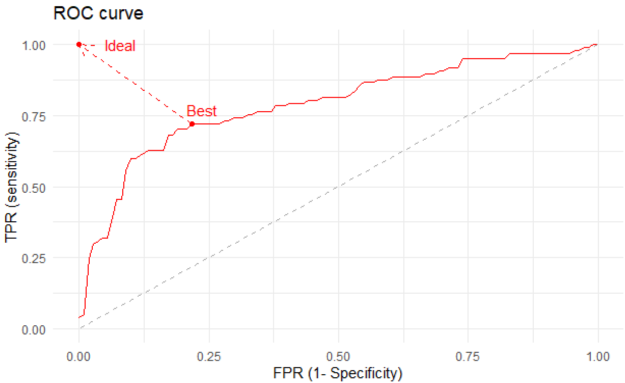
[The ROC Curve: How to read?]

- Optimally, the curve should touch top-left corner



[THE ROC CURVE: HOW TO READ?]

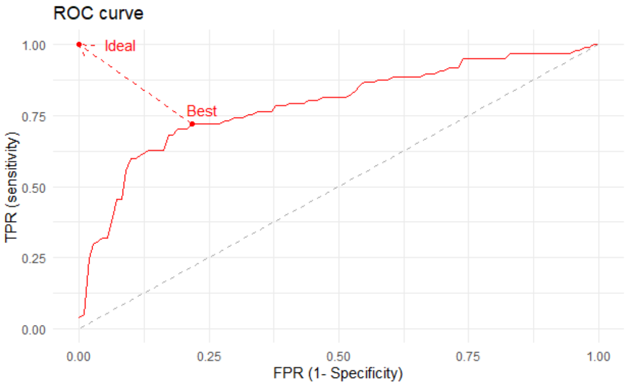
- Optimally, the curve should touch top-left corner



- If $t_0 \nearrow$, more cases classified as *Negatives*, less *Positives*

[The ROC Curve: How to read?]

- Optimally, the curve should touch top-left corner



- If $t_0 \nearrow$, more cases classified as *Negatives*, less *Positives*
- If $t_0 \nearrow$, specificity $\nearrow$ and sensitivity $\searrow$

[AUC as a measure of fit]

[AUC as a measure of fit]

A model that works well, whatever the threshold is certainly desirable

[AUC as a measure of fit]

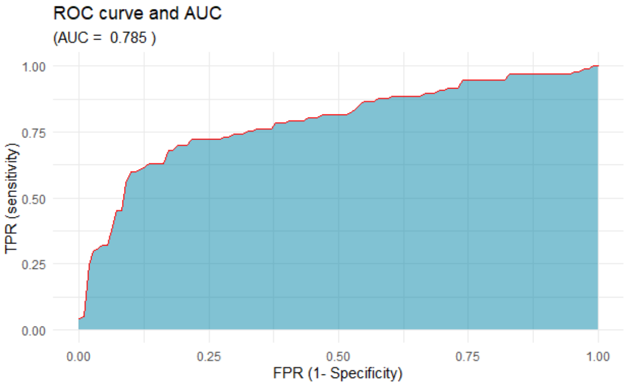
A model that works well, whatever the threshold is certainly desirable

- Using the AUC is also a measure of fit of a model

[AUC AS A MEASURE OF FIT]

A model that works well, whatever the threshold is certainly desirable

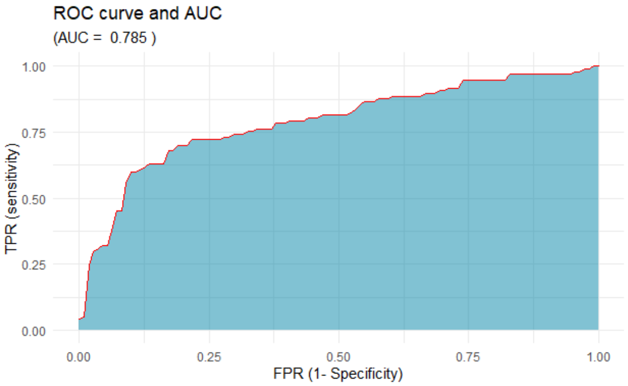
- Using the AUC is also a measure of fit of a model



[AUC as a measure of fit]

A model that works well, whatever the threshold is certainly desirable

- Using the AUC is also a measure of fit of a model



- The greater the area, the better the model

[Comparing the measures]

[Comparing the measures]

- We have several measures at hand

[Comparing the measures]

- We have several measures at hand
- We should evaluate those models on their predictive performance on a new "*unseen*" data set

[Comparing the measures]

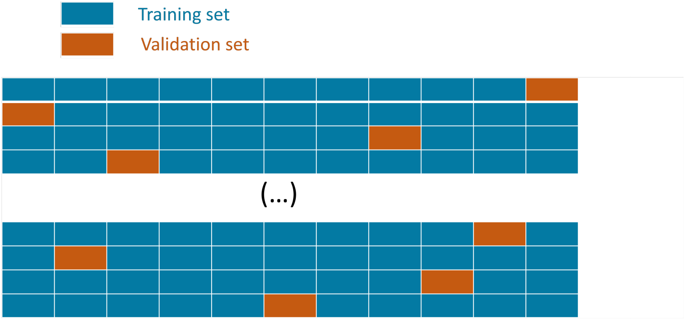
- We have several measures at hand
- We should evaluate those models on their predictive performance on a new "*unseen*" data set

↪ This is what Cross-Validation can do

[COMPARING THE MEASURES]

- We have several measures at hand
- We should evaluate those models on their predictive performance on a new "*unseen*" data set

↪ This is what Cross-Validation can do



[Comparing the measures]

[COMPARING THE MEASURES]

- For any model, CV gives several classifications

[Comparing the measures]

- For any model, CV gives several classifications
- All the criteria derive from the confusion matrix

[Comparing the measures]

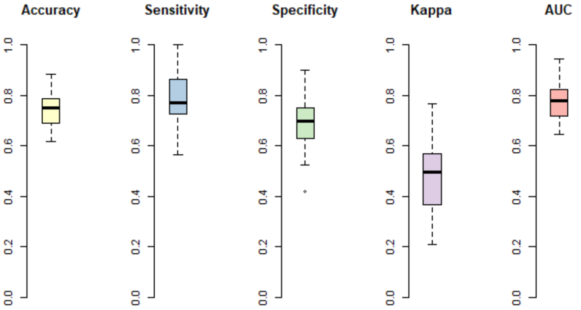
- For any model, CV gives several classifications
- All the criteria derive from the confusion matrix

$\hookrightarrow$ Examine them all!

[Comparing the measures]

- For any model, CV gives several classifications
- All the criteria derive from the confusion matrix

↪ Examine them all!



[How to chose the best model?]

[How to chose the best model?]

- We have several criteria for one model

[How to chose the best model?]

- We have several criteria for one model
- We should again evaluate the classifier based on "*unseen*" data set

[How to chose the best model?]

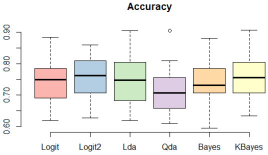
- We have several criteria for one model
- We should again evaluate the classifier based on "*unseen*" data set

↪ Run Cross-Validation an all!

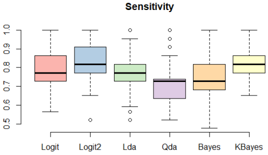
[How to chose the best model?]

- We have several criteria for one model
- We should again evaluate the classifier based on "*unseen*" data set

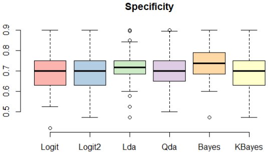
↪ Run Cross-Validation an all!



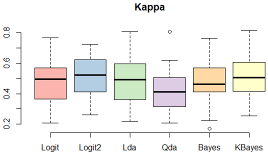
Accuracy of all CV validation sets,on all models



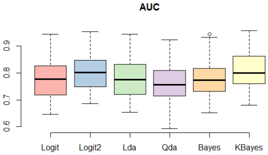
Sensitivity of all CV validation sest, on all models



Specificity of all CV validation sest, on all models



Kappa of all CV validation sets, on all models



AUC of all CV validation sets, on all models

[Takeaways]

- In classification, the **Confusion matrix** is important

[TAKEAWAYS]

- In classification, the **Confusion matrix** is important
- Many adjustment measures: **accuracy**, **sensitivity** and **specificity**.

[TAKEAWAYS]

- In classification, the **Confusion matrix** is important
- Many adjustment measures: **accuracy**, **sensitivity** and **specificity**.
  - *Sensitivity* is accuracy restricted to the positives.

[Takeaways]

- In classification, the **Confusion matrix** is important
- Many adjustment measures: **accuracy**, **sensitivity** and **specificity**.
  - *Sensitivity* is accuracy restricted to the positives.
  - *Specificity* is accuracy restricted to the negatives.

[Takeaways]

- In classification, the **Confusion matrix** is important
- Many adjustment measures: **accuracy**, **sensitivity** and **specificity**.
  - *Sensitivity* is accuracy restricted to the positives.
  - *Specificity* is accuracy restricted to the negatives.
- When outcome is *imbalanced*, one may use **kappa** has a better measure for accuracy.

[Takeaways]

- In classification, the **Confusion matrix** is important
- Many adjustment measures: **accuracy**, **sensitivity** and **specificity**.
  - *Sensitivity* is accuracy restricted to the positives.
  - *Specificity* is accuracy restricted to the negatives.
- When outcome is *imbalanced*, one may use **kappa** has a better measure for accuracy.

  Which measure you should consider depends on the context and your goal.

[Takeaways]

- In classification, the **Confusion matrix** is important
- Many adjustment measures: **accuracy**, **sensitivity** and **specificity**.
  - *Sensitivity* is accuracy restricted to the positives.
  - *Specificity* is accuracy restricted to the negatives.
- When outcome is *imbalanced*, one may use **kappa** has a better measure for accuracy.

  Which measure you should consider depends on the context and your goal.
- **Logit** is a benchmark parametric model for classification

[Takeaways]

- In classification, the **Confusion matrix** is important
- Many adjustment measures: **accuracy**, **sensitivity** and **specificity**.
  - *Sensitivity* is accuracy restricted to the positives.
  - *Specificity* is accuracy restricted to the negatives.
- When outcome is *imbalanced*, one may use **kappa** has a better measure for accuracy.

  Which measure you should consider depends on the context and your goal.
- **Logit** is a benchmark parametric model for classification

  One may use the **ROC** to change the threshold parameter

[Takeaways]

- Use (*Training-Validation*) sets to **select** parameters within a model

[Takeaways]

- Use (*Training-Validation*) sets to **select** parameters within a model
- Use (*Training-Validation*) sets to **compare** models on the same criteria

[Takeaways]

- Use (*Training-Validation*) sets to **select** parameters within a model
- Use (*Training-Validation*) sets to **compare** models on the same criteria
- Several criteria / measures of fit / cost functions are available

[Takeaways]

- Use (*Training-Validation*) sets to **select** parameters within a model
- Use (*Training-Validation*) sets to **compare** models on the same criteria
- Several criteria / measures of fit / cost functions are available
- Time is the limit...

[Q&A]

Write your questions in the chat

[Next week]

[NEXT WEEK]

- Module 3: "Regression" (Multiple dimension, penalization methods, ...)

[Next week]

- Module 3: "Regression" (Multiple dimension, penalization methods, ...)
- Webinar on "Regression" Thursday, same time

[Next week]

- Module 3: "Regression" (Multiple dimension, penalization methods, ...)
- Webinar on "Regression" Thursday, same time
- Complete the activities before the webinar!

[Next week]

- Module 3: "Regression" (Multiple dimension, penalization methods, ...)
- Webinar on "Regression" Thursday, same time
- Complete the activities before the webinar!
- **Continue to post your thoughts on the forums**

[Next week]

- Module 3: "Regression" (Multiple dimension, penalization methods, ...)
- Webinar on "Regression" Thursday, same time
- Complete the activities before the webinar!
- **Continue to post your thoughts on the forums**

Have a nice week!