Machine Learning Part I (B. Wilbertz)
M2 Probabilités & Finance, UPMC-Ecole Polytechnique
M2 P.M.A.
6th January, 2023

1.5h, books and mobile phones not allowed

A Kaggle in-class competition

1. (*48 points*) State your team name for the competition (probably your student id or full name)
2. (*1 point*) Which algorithms did you use for the final submission?
3. (*1 point*) What was your biggest insight when participating in the competition?

B Evaluation of ML trainings

(*10 points*) In a binary classification problem (classes `A` and `B`) a machine learning classifier is producing the following results:

probs	ground truth
0.1	A
0.2	A
0.3	A
0.4	B
0.5	A
0.6	B
0.7	A
0.8	B
0.9	B
1.0	B

TP = 4.

TN = 4.

FP = 1

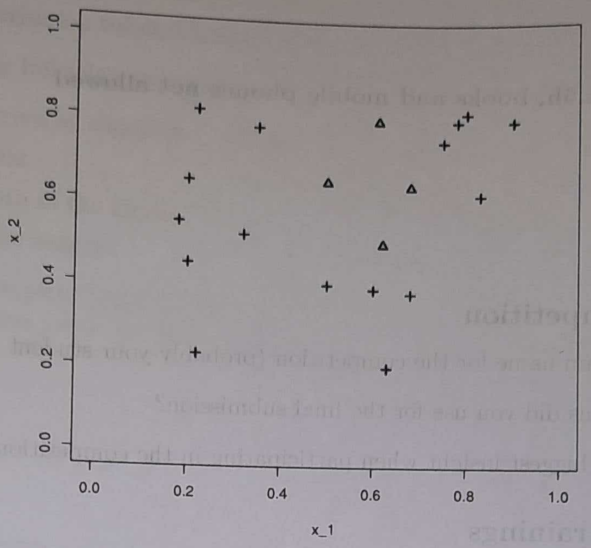
FN = 1

(probs are confidences for class `B`).

1. Draw the ROC curve for this model
2. Compute the AUC (area-under-the-curve) value for this model

C Decision Tree

(*10 points*) Consider the following binary classification problem with inputs (x_1, x_2) and outcome $y \in \{\Delta, +\}$.

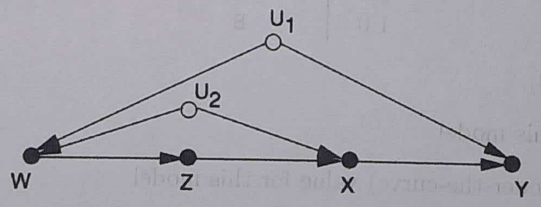


$\Sigma p_i (1 - p_i)$.

1. Compute the Gini index for this dataset
2. Perform the splitting operation of the decision tree algorithm on this dataset using the Gini index as loss criterium (spliting values should be estimated from the plot). Report the Gini loss reduction for every split and continue until no further reduction of the loss is possible.
3. Draw the resulting classifier in tree form

D Causal Inference

(*15 points*) In the following we want to estimate the effect of X on Y from below causal diagram, where U_1 and U_2 are unobserved variables.



1. Explain why we cannot use the back-door or front-door criterium to estimate this effect.
2. Transform $\mathbb{P}(Y|do(X))$ into a *do*-free expression using the 3 rules of the *do*-calculus.

E Multiple choice questions 15 pts → 3pt/20

For each question only one possible answer is correct.

1. (*1 point*) What are typical numbers for K in K-fold cross-validation?
   - (a) $K = 3$ and $K = 9$
   - (b) $K = 4$ and $K = 8$
   - (c) $K = 4$ and $K = 10$
   - [✓ (d)] $K = 5$ and $K = 10$ ←

2. (*1 point*) What is the usual relation between K in in K-fold cross validation and the number of bootstrap iterations n?
   - [✓ (a)] $n > K$ ✓
   - (b) $n \approx K$
   - (c) $n < K$
   - (d) n can be larger or smaller than K

3. (*1 point*) Which metric should be used for class imbalanced data sets?
   - (a) Accuracy
   - (b) Specificity
   - [✓ (c)] AUC ✓
   - (d) RMSE

4. (*1 point*) Which of the following metrics does not make sense for regression problems?
   - (a) AIC
   - (b) R^2
   - (c) RMSE
   - [✓ (d)] ROC ✓

5. (*2 points*) Which statement is correct?
   - [✓ (a)] The lasso method yields sparse models ←
   - (b) The ridge method yields sparse models
   - (c) The ridge method is designed to lower model bias
   - (d) The drawback of the lasso method is an increased model variance

6. (*2 points*) Which statement is correct?

   (a) SVM methods only works for linearly separable data

   (b) SVM method needs no cross-validation for hyperparameter tuning

   (c) The kernel trick for SVMs allows separation in a lower dimensional space

   [✓ (d)] The soft-margin criterion tolerates some violation of the data separation assumption

7. (*2 points*) Which of the following principles makes bagging work for random forests?

   [✓ (a)] Decorrelation of trees by choosing random predictor subsets

   (b) Pruning of the trees

   (c) Increasing the depth of the trees

   (d) Penalty term on the weights

8. (*1 point*) What is the typical number m of predictors to be taken at each split in the random forest algorithm for classification (total number of predictors p, size of training set n)?

   (a) $p/2$

   (b) $p/3$

   [✓ (c)] $\sqrt{p}$

   (d) $\log(p)$

9. (*2 points*) What can we say about the predictors when a gradient boosting algorithm yields only depth-1 trees?

   (a) We have to train with less data

   (b) We have to train with more data

   [✓ (c)] There is no interaction between the predictors

   (d) All predictors are linearly dependent

10. (*2 points*) Which statement is correct?

   [✓ (a)] Gradient boosting can overfit if the number of trees becomes too large

   (b) Random Forest can overfit if the number of trees becomes too large

   (c) A small learning rate η (also called shrinkage parameter) for gradient boosting reduces the number of trees needed

   (d) The boosting principle favors large trees to prevent overfitting

B - Evaluat° of ML Trainings

1)

P	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Classe	A	A	A	B	A	B	A	B	B	B

Résultat

si la prédict° est B (P ≥ seuil) et l'observat° est $\begin{cases} B \to \text{Vrai Positif (TP)} \\ A \to \text{Faux Positif (FP)} \end{cases}$

si la prédict° est A (P < seuil) et l'observat° est $\begin{cases} B \to \text{Faux Négatif (FN)} \\ A \to \text{Vrai Négatif (TN)} \end{cases}$

On convertit les classes en représentat° binaire (0 pour A, 1 par B)

Donc ici on a $\begin{cases} TP = 4 \\ FP = 2 \\ FN = 1 \\ TN = 3 \end{cases}$ => Total Positifs réel = TP + FN = 5

Pour chaque seuil t, on recalcule les métriques suivantes

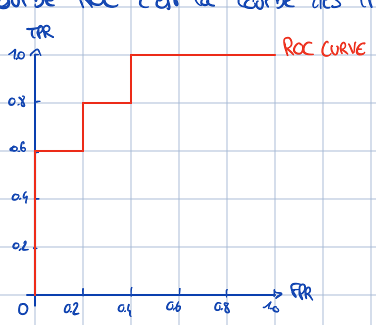
TPR (True Positive Rate) tq $TPR = \dfrac{\text{Vrais positifs}}{\text{Total Positifs réels}}$ → Du tableau des seuils →

= TP+FN de l'observat°

FPR (False Positive Rate) tq $FPR = \dfrac{\text{Faux positifs}}{\text{Total Positifs réels}}$

Seuil	Prédict° (B si P ≥ seuil)	TP	FP	TN	FN	TPR	FPR
1.0	[A,A,A,A,A,A,A,A,A,B]	1	0	5	4	1/5 = 0.2	0
0.9	[A,A,A,A,A,A,A,A,B,B]	2	0	5	3	0.4	0
0.8	[A,A,A,A,A,A,A,B,B,B]	3	0	5	2	0.6	0
0.7	[A,A,A,A,A,A,B,B,B,B]	3	1	4	2	0.6	0.2
0.6	[A,A,A,A,A,B,B,B,B,B]	4	1	4	1	0.8	0.2
0.5	[A,A,A,A,B,B,B,B,B,B]	4	2	3	1	0.8	0.4
0.4	[A,A,A,B,B,B,B,B,B,B]	5	2	3	0	1.0	0.4
0.3	[A,A,B,B,B,B,B,B,B,B]	5	3	2	0	1.0	0.6
0.2	[A,B,B,B,B,B,B,B,B,B]	5	4	1	0	1.0	0.8
0.1	[B,B,B,B,B,B,B,B,B,B]	5	5	0	0	1.0	1.0

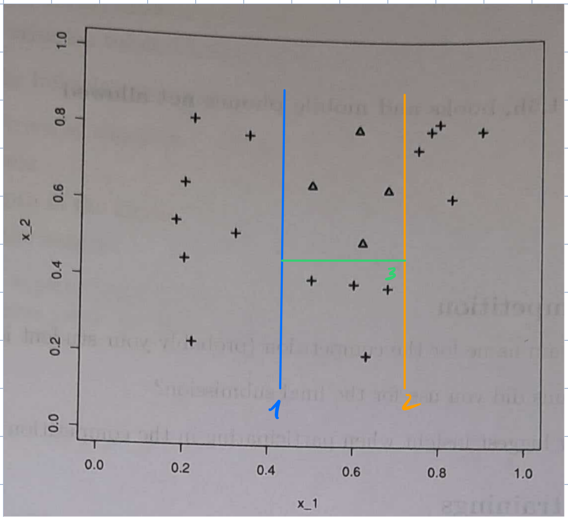
Courbe ROC c'est la courbe des (FPR, TPR)
X Y



2) $AUC = 0.6 \times (0.2 - 0.0) + 0.8 \times (0.4 - 0.2) + 1.0 \times (1.0 - 0.4)$

$= 0.88$

C - Decision Tree



1) $Gini = 1 - \sum_{i=1}^{m} p_i^2$ où p_i est la proport° de la classe i dans le dataset

Ici il y a 2 classes, Δ et +
4 éléments 16 éléments

$\Rightarrow p_\Delta = \frac{4}{20} = 0.2$

$p_+ = \frac{16}{20} = 0.8$

$\Rightarrow Gini = 1 - 0.2^2 - 0.8^2$

$= 0.32$

2) On découpe pour réduire l'indice de Gini qui est signe d'"impureté"

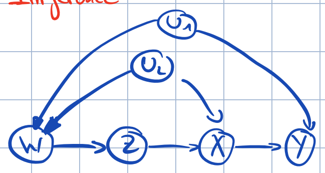
Découpage 1 $\Rightarrow Gini_{split} = \frac{7}{20} \times 0 + \frac{13}{20} \times \left(1 - \left(\frac{4}{13}\right)^2 - \left(\frac{9}{13}\right)^2\right) = 0.129$

Decoupage 2 $\Rightarrow Gini_{split} = \frac{5}{20} \times 0 + \frac{8}{20} \left(0 + \underbrace{\left(1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2\right)}_{1/2}\right) = \frac{4}{20} = 0.2$

Decoupage 3 $\Rightarrow Gini_{split} = 0$

D- Causal Inference

1)



No front door because there are no intermediates between X and Y

No back door because one should block $U_2 \rightarrow W$ but W is a collider → creates a path $U_1 \cdots U_2$ when conditioning on W

2)