Rules :

Don't hesitate adding answers / developing the ones already tgere

You can put the page number and lecture name to every question, so we can all find it and verify if needed (in case question is slightly different)

Good answer

Not sure

bad answer

QUIZZ 1

1. **You run gradient descent for 15 iterations with a=0.3 and compute J(theta) after each iteration. You find that the value of J(Theta) decreases quickly and then levels off (stabilize). Based on this, which of the following conclusions seems most plausible?** Veuillez choisir une réponse:

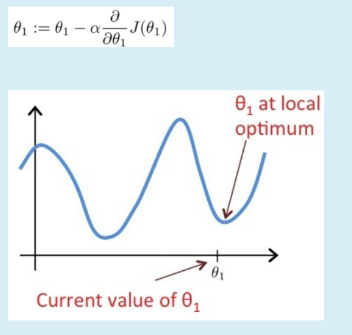
a. None of the above

b. a=0.3 is an effective choice for a learning rate

c. Rather than using the current value of a. use a lower value of a (a=0.1)

d. Rather than using the current value of a. use a larger value of a (a-1.0)

1. **Suppose θ1 is at a local optimum of J(θ1), such as shown in the figure. What will one step of gradient descent do?** Veuillez choisir au moins une réponse:



a. Move θ1 in the direction of the global minimum of J(01)

b. Leave 01 unchanged

c. Change 01 in a random direction

d. Decrease 01

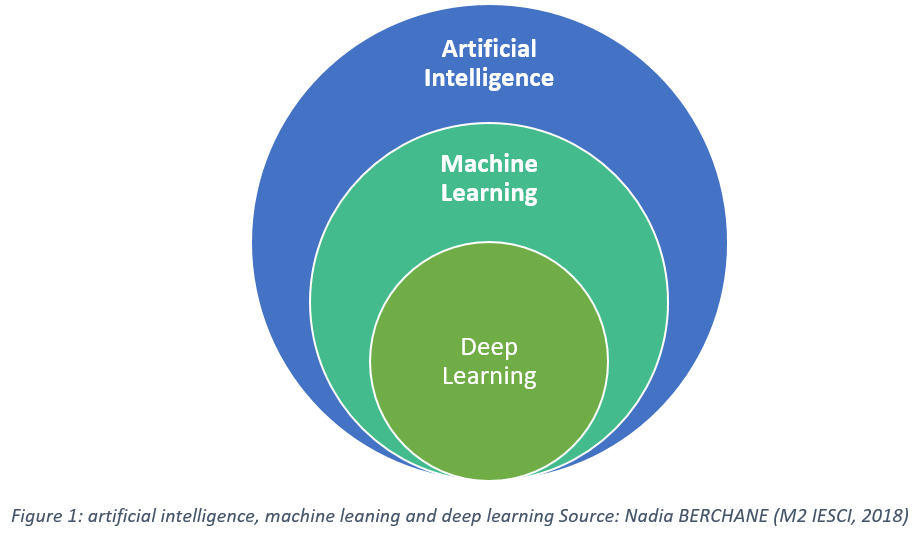
θ1 is at a local optimum. Hence, its derivative on this particular point is 0.

Hence, the formula θ1 = θ1 - α d(J(θ1))/dθ1 becomes simply θ1 = θ1 - α\*0 so finally θ1 = θ1

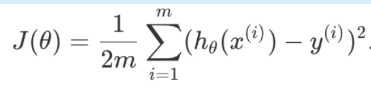
So θ1 is not going to change. Answer B

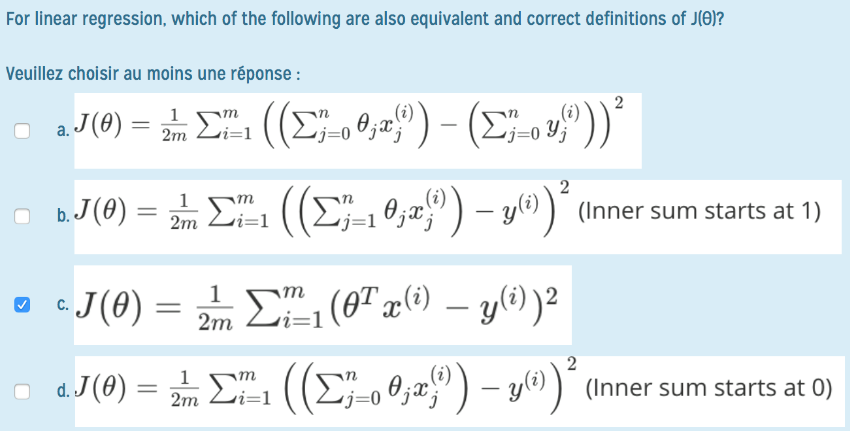
**4) Deep learning is a subset of machine learning which is a subnet of AI.** Veuillez choisir une réponse:

* True
* False



**5) When there are n features, we define the cost function as**





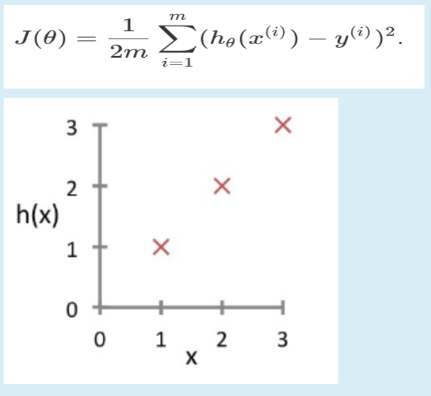
d) it's not the

b) because we add the intercept (so we have n+1 input variables)

c) it's the same as

d) with the matricial product that is not developed

**6) Suppose we have a training set with m=3 examples, plotted below. Our hypothesis representation is: hθ(x) = θ1x, with parameter θ1. The cost function J(θ1) is :**



**What is J(-1) ? (1 answer)**

1. 56/6
2. 26
3. 0
4. 14/6

We have J(-1) = 1/(2\*m) [sum i=1 to m] (h[-1](x\_i) - y\_i)²

because :

h[θ](x) = θx, so

h[-1](x) = -1x

so we have J(-1) = 1/(2x3) ( (1 - (-1) )² + (2 - (-2) )² + (3 - (-3) )² )

J(-1) = 1/6 \* ( 2² + 4² + 6²)

J(-1) = 1/6 \* (4 + 16 + 36)

J(-1) = 1/6 \* (56)

J(-1) = 56/6

answer a)

**7) How do you handle missing/corrupted data in a dataset?** Veuillez choisir une réponse:

a. Assign a unique category to string value

b. Drop missing rows or columns

c. All of the above

d. Replace missing values with mean/median/most frequent

**8) A computer program is said to learn from experience E with respect to some task T and some performance measure R if its performance on T. as measured by R improves with experience E.**

**Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to better filter spam. What is the task T in this setting?** Veuillez choisir au moins une réponse:

a. The number (or fraction) of emails correctly classified as spam/not spam.

b. Classify emails as spam or not spam.

c. Watching you label emails as spam or not spam.

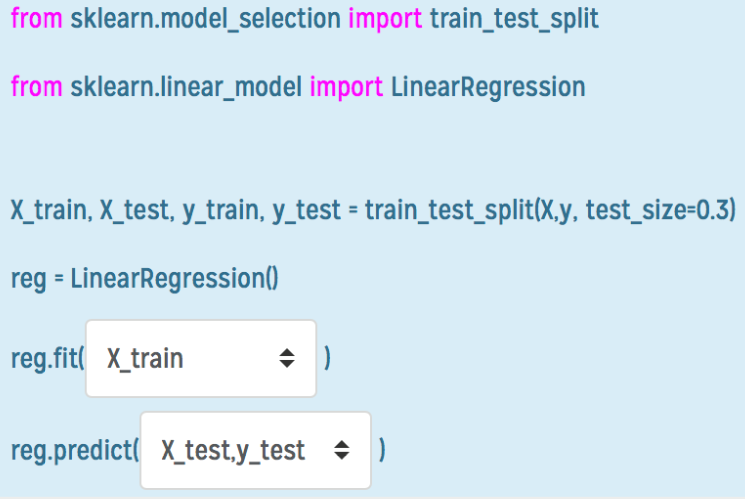
d. None of the above, this is not a machine learning algorithm.

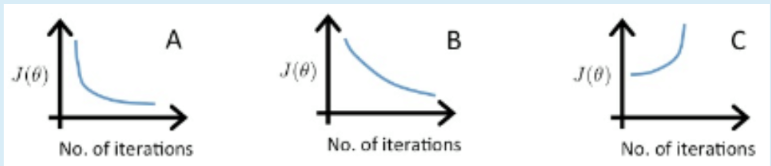
// ça sort d’ou en bas?

c) all of the above ?

at least d) and b) are true, not sure for a)

**9) Say we have a dataframe df in which we have all the features and the value to predict. After separating features and labels into X and Y, we want to split the model into training and test and train it. Fill the blanks in the following piece of code**

****

**10) Suppose a friend ran gradient descent three times, with alpha=0.01, alpha=0.1 and alpha=1 and got the following three plots (labeled A,B and C):** Veuillez choisir une réponse:

a. A is a=0.1, B is a=0.01, C is a=1.

b. A is a=0.01, B is a=0.1, C is a=l.

c. A is a=l, B is a=0.l, C is a=0.01.

d. A is a=l, B is a=0.1, C is a=0.l.

We see that C does not converge, so it has to be the largest α.

We see that B goes down much more slowly than A, so it has to have an α that is lower than the B.

From this, we have that αB < αA < αC

--> answer a)

QUIZZ 2

1. **You run gradient descent for 15 iterations with a=0.3 and compute J(theta) after each iteration. You find that the value of J(theta) decreases quickly and then levels off (stabilize). Base on this which of the following conclusions seems most plausible?** (1 answer)

a.none of the above

b) It is a good learning rate because it decreases quickly, then stabilizes. It would be bad if it was too slow, and/or if it didn't stabilize.

c.Rather than using the current value of a, use a lower value (ex a =0.1)

d.Rather than using the current value of a, use a larger value (ex a =1.0)

1. **Which of the following statements are true? Select all that apply**

a.Gradient descent is guaranteed to find the global minimum for any function J(teta0,teta1)

b.To make gradient descent converge, we must slowly decrease over time

c.Gradient can converged even if alpha kept fixed

d.For specific choice of cost function used in linear regression, there are no local optima (other than global minimum)

Justification:

a) false, because you may find a LOCAL minimum instead of a global one. (that is because you take your starting point (=the initial value of θ) randomly)

b) false, α can be kept fixed and it would still work (cf what we did with KIM)

c) true : we did it in KIM's lab, and if α is too large, we never stabilize.

d) true : the cost function of a linear regression is like a parabole (expression like y = ax² + bx + c), which has only 1 minimum

1. **Which of the following are example of classification**

a.All of the above

b.Predicting if a football player is good, medium or bad (because discrete value -> output,category)

c.Predicting the age of a passenger who survived the titanic ( continuous value-> number in output)

d.Spotting anomalies in an unlabelled set

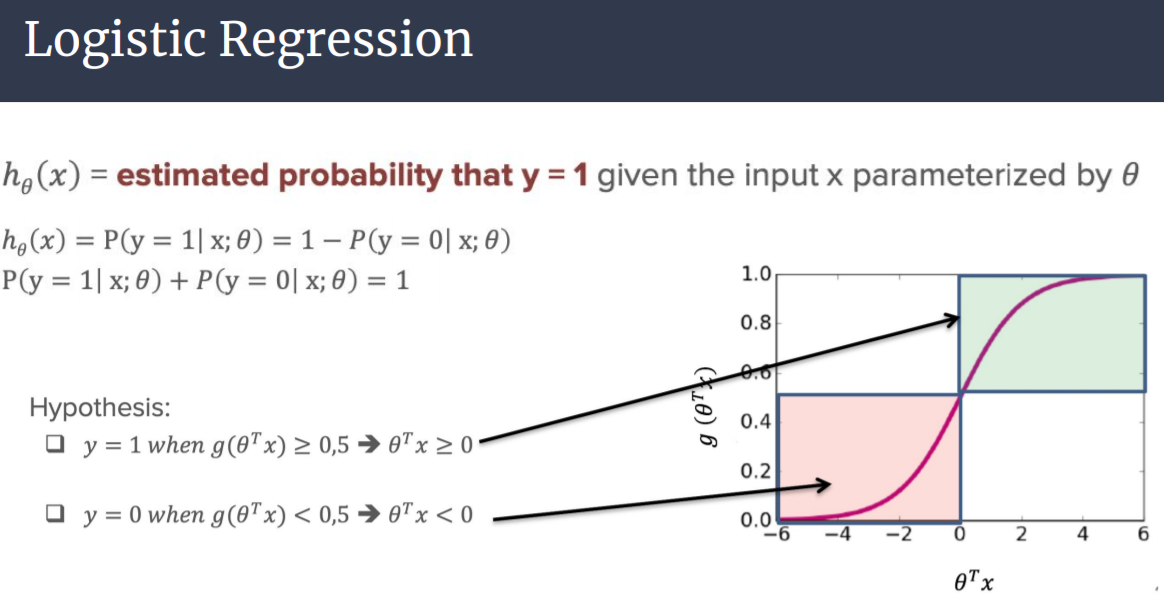
1. **Suppose we have trained a logistic regression classifier and it outputs on a new example hteta(x) =0.49. This means:**

a.Our estimate for p(y=1|x;teta) is 0.51

b.Our estimate for p(y=0|x;teta) is 0.51

c.The output will be positive

c.The output will be negative

****

1. **What is the purpose of the sigmoid function?**

a.Change the output of our linear regression model from -infinity <h(teta)<1

b.None of above

c.Create a new cost that will be convex and where we can run gradient descent

d.Optimizing to find the correct teta parameters

1. **Consider the medical diagnosis problem of classifying tumors as malignant or benign. If a hypothesis hteta(x) has overfit the training set. it means that:**

a.It does not make accurate predictions for examples in the training set and does not generalize well to make accurate predictions on new, previously unseen examples.

b.It **does not** make accurate predictions for examples in the training set but **does** generalize well to make accurate predictions on new, previously unseen examples.

c.It makes accurate predictions for examples in the training set and does not generalize well to make accurate predictions on new, previously unseen examples.

d.It makes accurate predictions for examples in the training set and generalize well to make accurate predictions on new, previously unseen examples.

1. **How many logistic regression do we perform when training a classifier to choose between k classes.**

a.None of the above

b.k

c.k+1

d.k-1

1. **Select the true statement about accuracy,precision and recall**

a.Recall is our ability to find relevant examples in a dataset

b.Accuracy is not helpful for an imbalanced dataset because of the imbalanced classification problem

c. Accuracy is the number of correctly classified points divided by the total number of points

d.We want to use F1, precision and recall for an imbalanced dataset

1. **K-Fold cross validation helps the model train on more data and get a better number for generalization error**

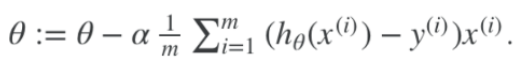
a.Vrai

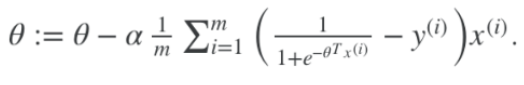
b.Faux

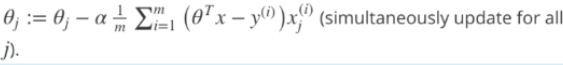
1. **For logistic regression, the gradient is given by:**

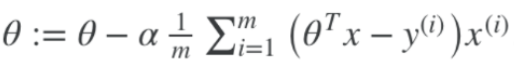


**Which of these is a correct gradient descent update for logistic regression with a learning rate of alpha? Check all that apply.**

a.

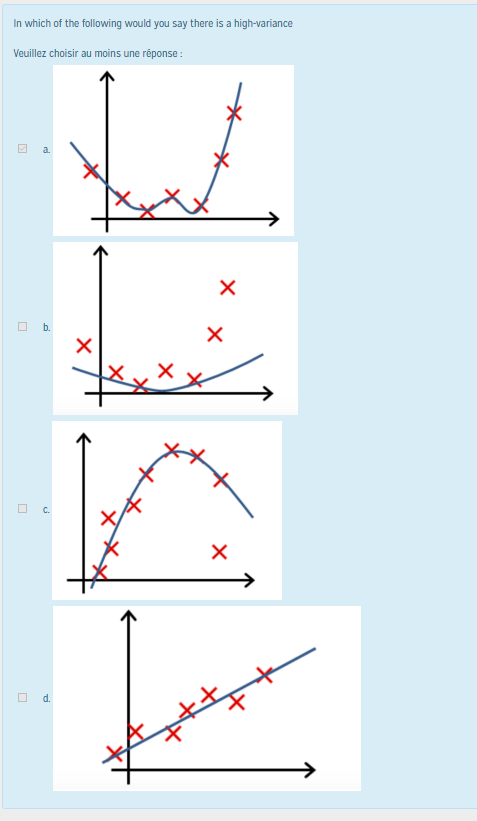
b.

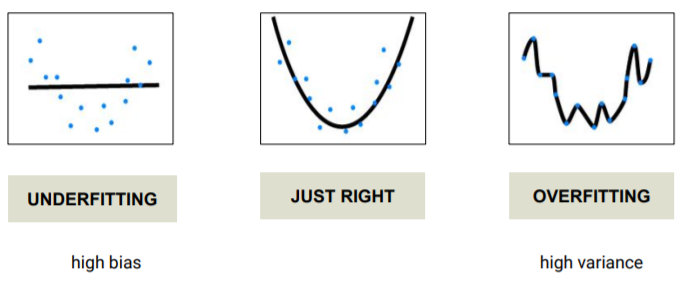
c.

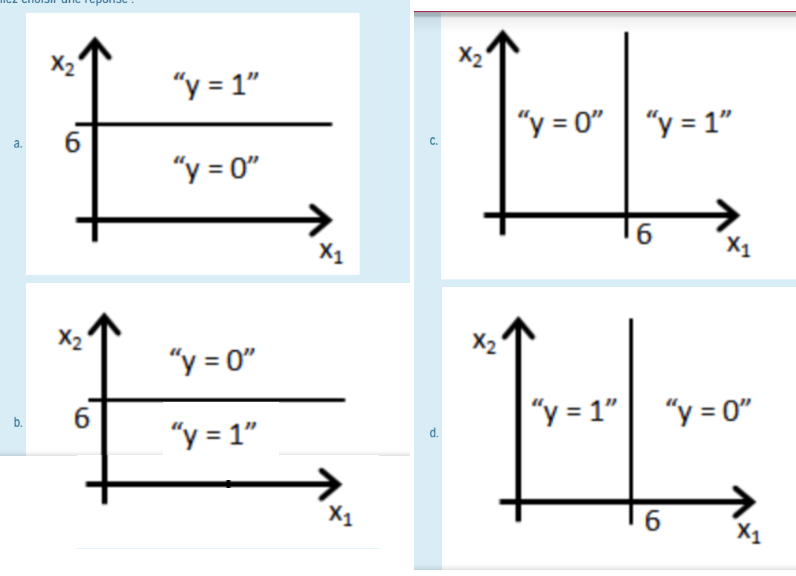
d.

1. **In which of the following would you say there is a high variance (**Veuillez choisir au moins une réponse)

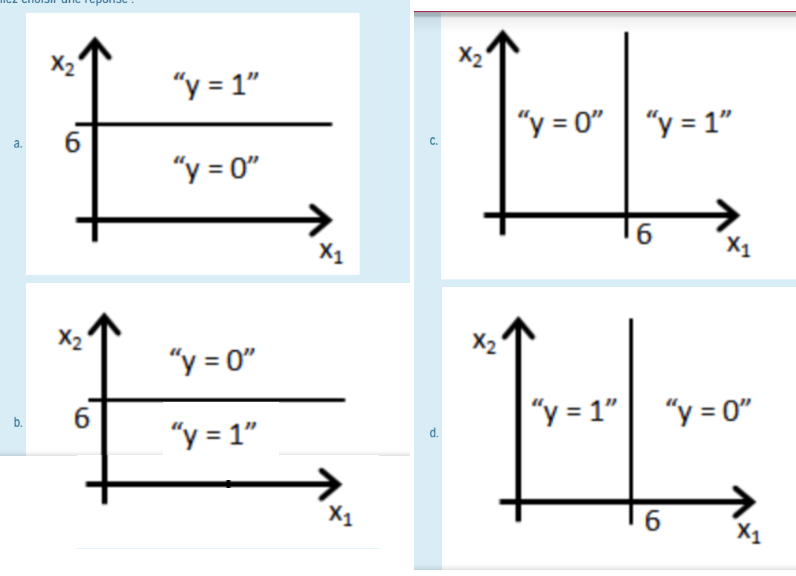
a.because it’s overfitting = high variance





1. **What is the correct decision boundary for logistic regression hθ(x) = g(θ0 + θ1\*x1 + θ2\*x2) with θ=[6;0;-1]** (1 answer) : 

* a
* b
* c
* d

1. **What is the correct decision boundary for logistic regression hθ(x) = g(θ0 + θ1\*x1 + θ2\*x2) with θ=[6;-1;0]** (1 answer) : 

* a
* b
* c
* d

Quizz 3

1. **The deeper a decision tree, the more it is prone to overfitting (only 1 answer)**

* True
* False

1. **How do we know which variable to use when splitting data in a decision tree? (only one answer)**

* I have no idea, leave me alone
* We use information Gain and Entropy (or similar) to detect the purest features
* we do it at random
* we tell the algorithm which features we want to use in a specific order

1. **Which of the following is a widely used and effective machine learning algorithm based on the idea of ensemble? (only one answer)**

* Decision Tree Regression
* Adaboost
* Decision Tree Classifier
* none of the above

1. **Select the disadvantages of decision trees? (at least 1 answer)**

* Decision trees are robust to outliers
* Decision trees are prone to be overfit
* Decision trees have too many parameters
* Decision trees are robust to outliers

1. **Which of the following is a disadvantage of decision trees ? (only 1 answer)**

* Decision trees are too intuitive
* Decision trees are prone to be overfit
* None of the above
* Decision trees are robust to outliers

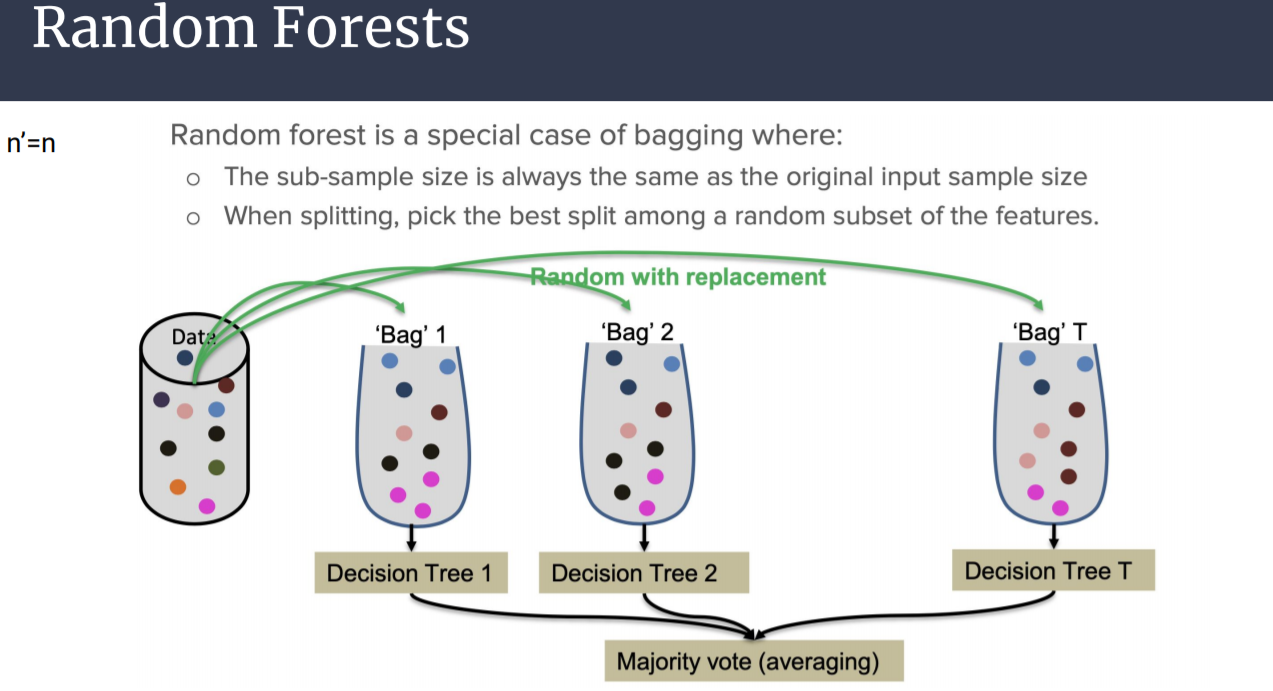
1. **Select the true statement(s) about Adaboost listed below: (at least 1 answer)**

* In Adaboost, each tree has a decision weight
* In Adaboost, some trees can have no incidence on the final decision
* In Adaboost, each tree is independent (->build on each other)
* Adaboost is an utilization of Bagging (-> utilization of boosting)

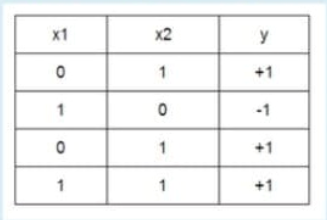
1. **Select the true statement(s) about Random Forest listed below (au moins 1 réponse):**

* In Random forests, each tree is independent it is random with replacement
* In Random forests, the more trees we have, the better it is for the score because we average the results got. So the more results we have, the more accurate the mean is.
* Random Forests are an utilization of boosting no special case of bagging
* In random forests, we can end up with twice the same data in a bag because of the idea of replacement we could take twice the same data, as it is put back in the bag after we took it

Remainder of what is a random forest :



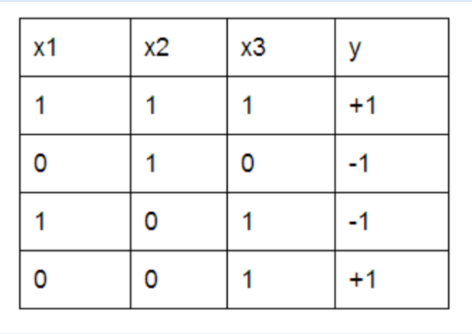
1. **Imagine we are training a decision tree and we are at a node (before a split). Suppose we have 2 features x1 and x2 and a label y that can take the value of (+1) or (-1). Which feature results in the best split? (only 1 answer)**



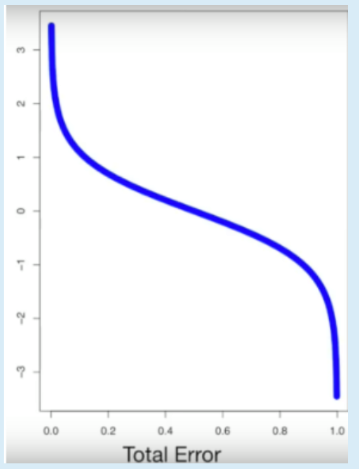
x2 because we see a direct correlation between the value of x2 and the final output y.

* x1

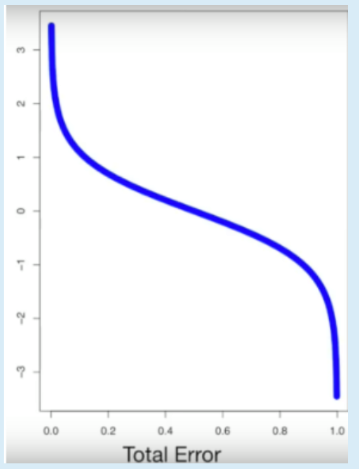
1. **Imagine we are training a decision tree and we are at a node (before a split). Suppose we have 3 features x1, x2 and x3 and a label y that can take the value of (+1) or (-1). Which feature results in the best split?** (only 1 answer)



* x2
* x1
* x3 100% sure !!!
* Both x1 and x2 work fine

**10) Say we are using Adaboost and need to decide the amount of say of a weak classifier that has 50% correct predictions. The amount of say of this stump will be**  (only 1 answer) **:**

* 0
* A positive number
* None of the listed
* A negative number

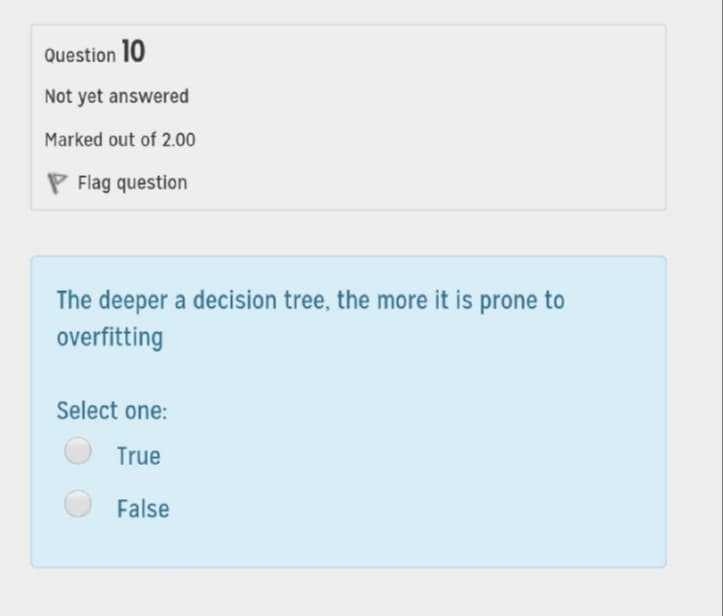


**11) Say we are using Adaboost and need to decide the amount of say of a weak classifier that has 75% correct predictions. The amount of say of this stump will be**  (only 1 answer)**:**

* 0
* A positive number
* None of the listed
* A negative number

**12) What will a decision tree do if we are at node of 43 elements, there are still features to split, and the min\_samples\_split value is set to 2? (only one answer)**

* None of the listed
* Split this branch, but only once more
* Keep splitting this branch
* Stop splitting this branch

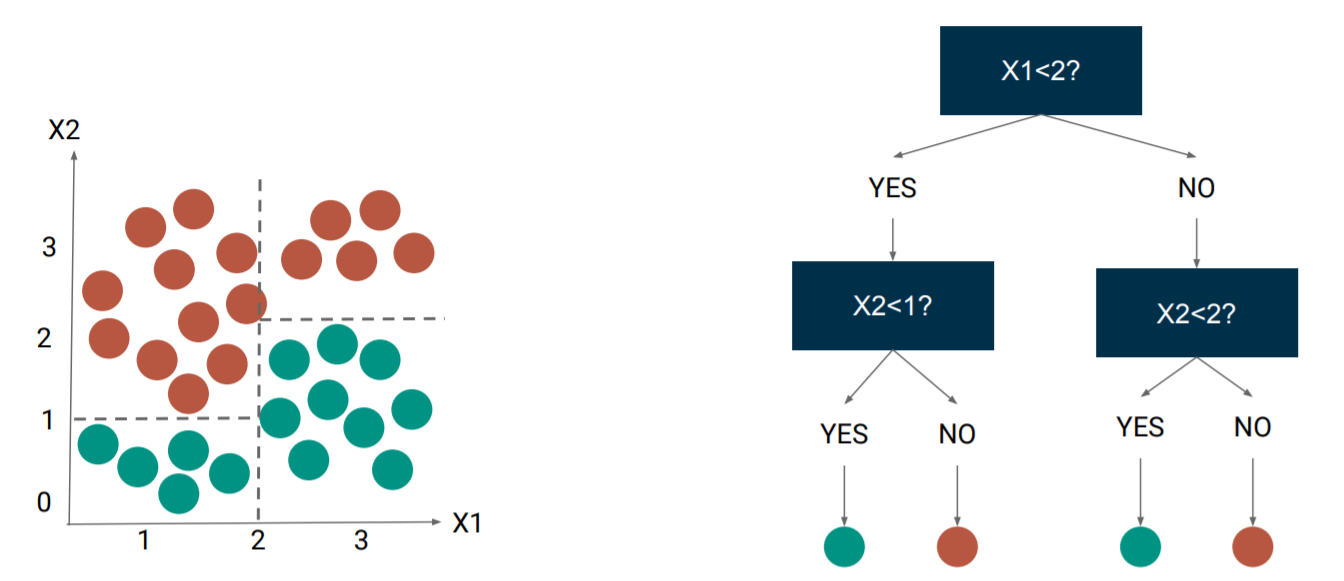
**13) What was the purpose of Grid Search CV (only one answer)**

* To search for the best variable to split in a decision tree
* To search for optimal parameters in a logistic or linear regression
* All of the above
* To test between multiple parameters and output the best one given a specific algorithm (SVM, Decision Tree, ...)

**Grid search** is an approach to parameter tuning that will methodically build and evaluate a model for each combination of algorithm parameters specified in a **grid**. The recipe below evaluates different alpha values for the Ridge Regression algorithm on the standard diabetes dataset. This is a one-dimensional **grid search**.

**14) Decision Tree (with depth>1) is always a linear classifier (only one answer)**

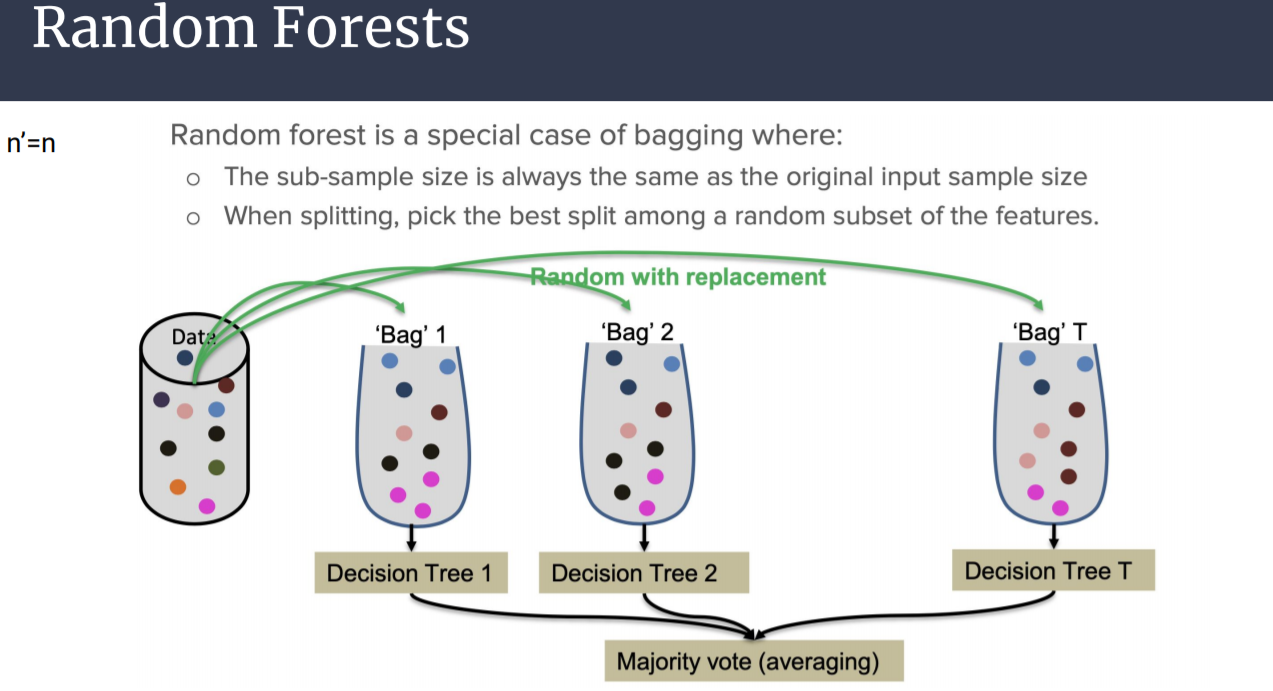
* True
* False , we can have a tree that acts like this for instance : data is clearly not linearly separated



**15) Which of the following is widely used and effective machine learning algorithm based on the idea of bagging? (only one answer)**

* Régression
* Random forest -> i would vote for that
* Classification
* Decision Tree : The other answers are not even machine learning algorithms

Regression : Reduce the number of input variables (or dimensions)

Random Forest : 

Classification : Saying that the output is in a "class" (true/false, dead/survived, etc)