INF 264 - Exercice 3

Instructions

- Deadline: 13/09/2019, 23.59
- Submission place: `https://mitt.uib.no/courses/19532/assignments`
- Format: Your answers are to be returned in a single pdf report. You can also return scanned pages for your calculations.

The present document is the third of a series of eight weekly exercises. The deadline for weekly exercises is on Friday evening. You will be awarded points for solving weekly exercises. Each week you can earn at most 1.5 raw points (depending on how many tasks you solve). Your solutions do not need to be perfect; we award points for reasonable effort. Weekly exercises give at most 10 points towards the final grade; if you get more than 10 raw points, it will count as 10 points. To pass the course, you will need to get at least 5/10 points from the weekly exercises. Note that you do not have to get points every week, just at least 5 points in total.

In this exercise we will introduce entropy, construct a simple decision tree, learn how to apply cross-validation for model selection in the context of regression and finally learn how to perform model selection for classification on unbalanced data.

1 Entropy and decision trees

In this first section, you will briefly refresh your memories on the concept of entropy from information theory, then you will construct a toy-example decision tree on a small dataset. **This section is to be done entirely by hand, not with any programming !**

1.1 Entropy

Let X a random discrete variable taking values in $\{x_0, \dots, x_{N-1}\}$. Denote by p_X the density of X:

$$X \sim p_X = \big\{P(x_0), \dots, P(x_{N-1})\big\} = \{p_0, \dots, p_{N-1}\}. \tag{1}$$

Recall the definition of the entropy of X:

$$H(X) = \sum_{n=0}^{N-1} p_n \log_2\left(\frac{1}{p_n}\right). \tag{2}$$

1. Let us consider the special case where $N = 2$, i.e X follows a Bernoulli distribution. We can write $p_0 = p$ and $p_1 = 1 - p_0 = 1 - p$, thus the entropy of X is simply:

   $$H(X) = p \log_2\left(\frac{1}{p}\right) + (1-p)\log_2\left(\frac{1}{1-p}\right). \tag{3}$$

   Plot $H(X)$ as a function of the parameter p. From the observation of the obtained graph, explain why entropy is said to be a measure of disorder.

2. More generally, it is possible to show that the entropy of X is maximal if and only if X follows a uniform distribution, that is if and only if $p_0 = \dots = p_{N-1} = \frac{1}{N}$. Prove the previous statement. You can for instance first show that when X follows a uniform distribution then its entropy is equal to $\log_2(N)$, and secondly use the Jensen's inequality theorem recalled below to show that $H(X) \leq \log_2(N)$ always holds.
   Hint: look at the general definition of $H(X)$ in (2) and find what are f, λ_n and x_n from the Jensen's inequality.

Jensen's inequality. *Let $f : I \to \mathbb{R}$ a concave function and $(\lambda_0, \ldots, \lambda_{N-1}) \in [0,1]^N$ such that $\sum_{n=0}^{N-1} \lambda_n = 1$. Then:*

$$\forall (x_0, \ldots, x_{N-1}) \in I^N : \quad \sum_{n=0}^{N-1} \lambda_n f(x_n) \leq f\left(\sum_{n=0}^{N-1} \lambda_n x_n \right). \tag{4}$$

1.2 Relation between entropy and information

1.2.1 Arresting a criminal

Police officers A, B and C are looking for a criminal. They know for sure that he is hiding in one of the 5 rooms on the 1st floor of a building. The police officers received the information that each room has 80 people in it.

1. Officer A did not receive anymore information with respect to the position of the criminal. What is the entropy of the criminal with respect to officer A ?

2. Officer B received in his personal receiver the additional information that the criminal is in the 2nd room. What is the entropy of the criminal with respect to officer B ?

3. Officer C received even more information: he learned via his personal receiver that the criminal is in the 2nd room, that this room consists of 8 rows each with exactly 10 people sitting and that the criminal is sat somewhere among rows 2, 3, 4. What is the entropy of the criminal with respect to officer C ?

4. What can you say about the relationship between information and entropy ?

1.2.2 Guessing cards

Consider a traditional 52-card deck, with 4 colors (hearts, diamonds, clubs and spades) and 13 ranks (from highest to lowest: ace, king, queen, jack and numbered ranks from 10 to 2), where each rank comes in every color. Suppose the deck was meticulously shuffled and is placed face-down. We are interested in the entropy of the random variable X that models the identity (rank and color) of the top card.

1. What is the density of X ? Compute its entropy $H(X)$.

Consider now the random variable Y that models the event E_Y: "Top card is an ace". We want to measure the impact of the information carried by Y on the entropy of X.

2. What is $P(E_Y)$? Compute the density of Y.

3. What are the densities of $X|Y{=}true$ and $X|Y{=}false$? Compute their respective entropies $H(X|Y{=}true)$ and $H(X|Y{=}false)$.

4. The entropy of X given information Y is defined as the mean over Y of $H(X|Y{=}y)$, that is in our case:

$$H(X|Y) = P(Y{=}true) \cdot H(X|Y{=}true) + P(Y{=}false) \cdot H(X|Y{=}false) \tag{5}$$

   Imagine a friend of yours draws the top card, looks at it without revealing anything and tells you that you can ask any question about its identity and he will answer (truthfully) to it. Information Y in this context can be understood as you asking the question "Is the top card an ace ?". Explain in your own words the physical interpretation of the formula above in the case of our variables X and Y.

5. Compare the entropies $H(X)$ and $H(X|Y)$. How did the addition of the information carried by Y affect the entropy of X ?

6. The quantity $IG(X|Y) = H(X) - H(X|Y)$ is called the information gain of X given information Y. For Y to be a valuable source of information about X, should $IG(X|Y)$ be small or large ? Explain the intuition behind that.

7. Is it more informative to ask the question "Is the top card an ace ?" or to ask the question "Is the top card a spades ?" when trying to identify the top card ?

1.3 Decision trees

Consider the following train dataset:

X	Y	Z	Class
1	0	0	A
0	0	0	B
0	1	1	A
1	0	1	B
0	0	1	A
1	1	1	B
0	1	1	A
0	1	0	A
0	0	0	B
1	1	1	B

1. Construct (on paper) a decision tree from the train dataset based on the entropy metric. Make your reasoning and computations apparent.

Now consider the following pruning dataset:

X	Y	Z	Class
0	1	1	B
0	0	0	A
0	1	0	B
0	0	1	B
0	0	0	B

2. Use this pruning dataset to perform reduced error post-pruning of your decision tree.

2 Model selection for regression

Consider the Boston Housing dataset: `https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html`. In this dataset, each sample corresponds to a house, whose target price is to be infered from 13 features contained in the dataset. In this second section you will complete a template code that answers the following questions:

1. Load the Boston dataset from sklearn. Store the 13 features in a matrix X and the target prices in a vector Y.

2. For each of the 13 features extracted, scatterplot the target prices as a function of this feature.

We want to build a polynomial regression model that learns the target prices from the 13 features. We are not sure about which polynomial order we should use to obtain the best model. Moreover, we were told that adding a regularization term can be useful to prevent overfitting, but once again we do not know how to choose this hyper-parameter. A solution is to perform model selection with cross-validation on the hyper-parameters.

3. Finish the implementation of a KFold cross-validation procedure in order to perform model selection of a Ridge model with respect to its hyper-parameters (polynomial order and regularization value), using the MSE metric. More specifically:

i Split the whole dataset into a train and a test set, then set aside the test set.

ii Split the train set into 5 (train,validation) folds using the KFold class from sklearn.

iii Loop over each hyper-parameters instance you cross-validate on (outer loop).

iv For current instance of hyper-parameters, loop over each (train,validation) fold (inner loop).

v In the inner loop, assess a "surrogate" model with current instance of hyper-parameters on the current (train,validation) fold, i.e fit your surrogate model on the train fold and evaluate it using the MSE metric on the validation fold.

vi In the outer loop (after the inner loop finishes), compute the mean validation MSE over each (train,validation) fold.

vii Select the model with the smallest mean validation MSE.

viii Train the selected model on the whole train set, then evaluate it on the test set which was set aside at the end of step i.

Cross-validation is to be done on the polynomial order ranging in $\{1, 2, 3\}$ and on the regularization value ranging in $\{0, 0.001, 0.01, 0.1\}$ for a total of 12 hyper-parameters combinations. Indicate which hyper-parameters combination obtained the best results with respect to the MSE metric during the KFold cross-validation procedure.

4. When fitting a Ridge model of 3rd degree, you should have encountered the warning "Singular matrix in solving dual problem. Using least-squares solution instead.". Can you explain why this warning occured ? How much do you need to increase the regularization hyper-parameter in order to get rid of this warning ?

3 Model selection for classification

In this last section, we will first illustrate how misleading the accuracy metric can be when assessing a classifier on unbalanced data. Finally, we will cross-validate different classifiers on an unbalanced dataset with a relevant metric. Complete the template code that answers the following questions:

1. Create randomly generated binary datasets, with a 1st class ratio ranging in $\{0.6, 0.75, 0.9, 0.95, 0.98, 0.99\}$. For each dataset generated this way, train a K-NN classifier with $K = 10$, then evaluate it on its corresponding test set with respect to the accuracy metric, the $F1$-score metric and the confusion matrix metric. Plot all of those results in a single figure (using subplots). Does the accuracy metric appear to assess the quality of your model in an appropriate way ?

2. Download a custom randomly generated binary dataset by clicking on this link: `https://mitt.uib.no/files/1893088/download?download_frd=1`. Visualize this dataset. How unbalanced is it ? Inspiring yourself from section 2, perform model selection on three different classification models: a K-NN classifier with $K = 20$, a logistic regression classifier and a decision tree classifier. In order to do this, use Kfold cross-validation with the number of folds set to 10. Indicate and justify which metric you decided to use in order to cross-validate the different models. Finally, train and evaluate the best model on the whole dataset (evaluate on the test set with the $F1$-score and the confusion matrix).