NOM BOUKHELOUA

Prénom Rémi

Promo 2020

Date

20150257: BOUKHELOUA Rémi
M1:
ST2ML2-DE (19/04/2019)
Amphi jaune

13

MATIÈRE ML and TM (1/2)

6. Naive Bayes algorithm is based on probabilities and the ~~formula~~ following formula: $P(c|x) = \frac{P(x|c) \times P(c)}{P(x)}$.

It ~~will~~ is a classification method that ~~will~~ performs well with multi classes dataset and better than others in general.
For a given input, it will compute and output a probability that the input is in a class or an other.
It's ~~problem~~ main problem is that ~~it~~ assumes independance of features what is rarely the case in real life problem. To counter that we can remove correlated features. We can't boost or bagg it and tuning is very restricted.

7. A minimal edit distance is the minimum number of operation we have to make to transform one word to another one.
It exists three operations: add a letter, delete a letter and change a letter. The two first ones cost 1 while the second, as a combination of the two others, costs 2.
For example :

A ~~N~~ N D
↓ ↓ ↓
~~N~~ N O D E

~~Find~~ d("AND", "NODE") = 1 delete + 2 additions = 3

11. From a corpora, we will first tokenize each text or speech. Tokenize first begin with splitting the text into words, then you remove stop words such as "or", "and" which have no impact. but also non english words.
On the rest of words, we can decide of lowering or keeping caps. The advantage of it is that it will reduce number of column ("YEAH" and "yeah" become "yeah") but we will loose some sens as caps on showing written sentiments.
After that, we decide to lematize which is change a word into a closer one ("has", "have", "had" become "have") or to stem which is keeping only root of word
We can now construct the matrix counting how much time a word appears or only if it appears (for Binary Naive Bayes). But having word we may prefer to work with frequencies to avoid having too big numbers. But words appearing a lot of time, may have less impact than less frequent ones. That's why we introduce log functio. We will obtain a TF IDF matrix. As it's now only composed of number, we can use this matrix on a machine learning method.

13. A general bagging framework ~~perfor~~ generates ~~differ~~ a lot of classifier or ~~subsets of~~ subsets of the train dataset. It will (same algorithm) then do the mean of them.

For random forest, we will create a lot of decision trees based on diffrent subsets of the dataset and then, instead of doing a mean, we will combine them by addition to have a bigger tree. As it is based on nodes we just can't make a mean...

14. Boosting is based on training several time increasing the the importance of ~~failed~~ failed set and reducing ~~the~~ on the succeed one. To do so, we introduce a weight at each row and identical at start. Higher it will be, higher will be the importance. The classifier will then ~~learn~~ focus on failed ones.

Ada boost is based on creating stump (1 node decision tree). with boosting method. It will generate ~~a lot of stumps~~ a stump at each boosting iteration and combine them, giving each one a weight, in order to have a good model.

15. As bagging is based on generating lot of model on diffrent part of train datset so each little model has a diffrent train and then meaning them, it will reduce the impact of training set on the model and so ~~the variance~~ will reduce the variance.

Boosting is based on focusing on failed rows, so it will ~~the~~ adapt the model each ~~time~~ iterations so it will reduce variance.

BIAS ?

6. Latent Dirichlet Allocation (LDA) is a method to make text with a certain repartition of words from different topics. For example, if you ~~have~~ have 3 topics A, B and C, you can use LDA to select words ~~that~~ with a certain repartition (20% of A, 80% of B).
It is based on Dirichlet coefficient which is a statistical variable representing the dispersion.
It could also be used to group text with same topic. Meaning Dirichlet you would be able to group ~~topic A~~ text from topic A together, topic B together and so on (based on their word repartition)

8. A bag of word of a text is words contained in the text ~~and that we can easily represent it as a matrix with bag~~
We could easily represent a text based on it ~~letter~~ as a matrix.
If the word is ~~there~~ in the bag, we put a one, if not a zero. or nb of time it appears
It is easy to do and to use in a machine learning method.
The problem is that we do not keep the order of the words what means that "I like math and hate physic" and "I like physic and hate math" would have the same representation. Moreover we do it based on a dictionnary so it ~~has to be~~ will not adapt on ~~the~~ unknown words. Finally, if we do it on long text, as we put a new and ~~feature for each~~ column for each word, it will ~~be~~ create a high number of dimension which ~~can~~ longer computational time.

NOM BOUKHELOVA

Prénom Rémi

Promo 2020

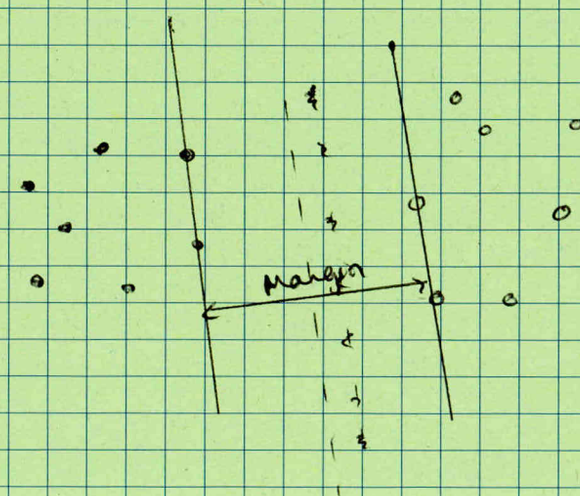
Date

MATIÈRE ML and TM (2/2)

10. Embedding can be creating n grams which are group of following n words. Unlik bag of words, it will keep ~~a~~ ~~of~~ order of words. ~~Ex~~ Creating bi bigrams (2 gram) out of "I like math and hate physic" would be ("I","like"), ("like","math"), ("math","and"), ("and","hate"), ("hate","physic") without ~~shortcut~~ lematization / steming. Based on this, we could create a Markov chain such as from one ngram we ~~h~~ would have probabilities of following one. And with a lot of dataset have a complex network from what we would generate text. or recognize author from a text (ex Shakespear).

12. Word2vec is a way of creating ~~numerical~~ vectors from a word that will allow ~~all~~ common operation on them. As a result we could have : King - man + woman = queen.

3.

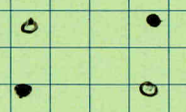


Support Vector Machine is model that create an hyperplane between classes. Without kernel trick, the hyperplan is linear. It will compute the hplane ~~with the highest~~ maximizing the margin. As we guess the ~~following~~ graphic 2Dimensional, I drew ~~on~~ the hyperplane maximizing it.

But with more dimensions and ~~for~~ ~~et~~ hidden ones not scaled, hyperplane could be very difficult!

9. F measure is frequency measures that we use when we transform a corpora into matrix.
We could use $\frac{\text{nb of apparition}}{\text{nb of words}}$ but also a ~~most~~ more interesting one $\frac{\text{nb of apparition}}{\text{nb of apparition of most common word}}$. In this case in a log, the result of most common word would be 0.

10. A straight line in a two dimensional feature space ~~could only separate~~ can't even separate two classes as following.



So I would say that it is 1. NB OF POINTS

4. In a neural network, we calculate weight on training set. It will be highly dependent of it. We could use some bagging method to generate ~~ma~~ several neural network and at the end mean them.
~~Also~~ We also have to care about the number of neurons in the hidden layer that should not be over twice number of input ~~layer~~ layer neurons to reduce overfitting.
~~I'm not sure but~~ Generally, cross validation process also reduces ~~overf~~ train set dependence, ~~but I am not sure~~ that could also be a possibility even though regarding computational time, I'm not sure it would be acceptable.

5. Instead of using the output of a the node, we could use a sigmoid function with threshold to output only 0 or 1 depending on the threshold.

efrei PARIS
Ecole d'ingénieurs du numérique

Ne rien inscrire dans ce cadre

Prénom Rémi

Nom BOUKHECOUA

Promotion 2020

Groupe

Promotion M1

Advanced Machine Learning and Text Mining

ST2ML2

DE - 1h45 min

Date Horaire

Sujet proposé par :

Calculatrice autorisée : ☐ **OUI** ☒ **NON**

Documents autorisés : ☐ **OUI** ☒ **NON** **Type de documents :**

Ordinateur portable autorisé : ☐ **OUI** ☒ **NON**

Internet : ☐ **OUI** ☒ **NON**

Traducteur électronique, dictionnaire : ☒ **OUI** ☐ **NON**

Consigne :

Merci de restituer uniquement : **« choisir un élément de la liste »**

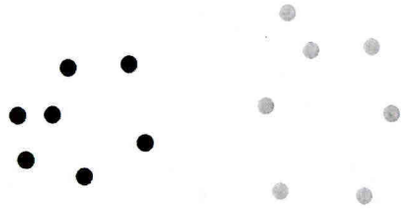
Rappel :

- Tous les appareils électroniques (téléphones portables, ordinateurs, tablettes, montres connectées ...) doivent être éteints et rangés.
- Il est interdit de communiquer.
- Toute fraude ou tentative de fraude fera l'objet d'un rapport de la part du surveillant et sera sanctionnée par la note zéro, assortie d'une convocation devant le conseil de discipline. Aucune contestation ne sera possible. Tous les documents et supports utilisés frauduleusement devront être remis au surveillant.
- Aucune sortie de la salle d'examen ne sera autorisée avant la moitié de la durée de l'épreuve.

« Code cours » - « Intitulé du cours » - « type de l'épreuve » **Date de l'épreuve**

Ne rien inscrire dans ce cadre

1. What is "minimizing the empirical risk"? Why and when are we using it?
2. What is VC dimension for a straight line in two dimensional feature space? Explain your answer.
3. Show SVM separating hyperplane for the dataset plotted bellow and explain why (without the kernel trick)



4. How to fight overfitting in a Neural Network learning?
5. What difficulties with convergence of Neural Network fitting process?
6. Explain Naïve Bayes algorithm
7. What is minimal edit distance?
8. Explain what is a bag of words model, its advantages and disadvantages
9. What is F-measure?
10. What is embedding? Example(s)?
11. What are the ways to create a matrix representation from corpora to use as an input for machine learning models? (feature creation)
12. Describe a schematic algorithm of skip-gram word2vec
13. What is particular for the random forest algorithm comparing to a general bagging framework?
14. What is a general concept of Adaboost algorithm
15. Explain boosting/bagging in bias-variance set up
16. What is the generative model behind Latent Dirichlet Allocation (LDA) ?