Su Mo Tu We Th Fr Sa

Date: / /

~~Appendix~~

Table of Contents

Su Mo Tu We Th Fr Sa

Date: 02/07/2021

Finding a suitable title for research proposal.

Title:

i) Assessing the pe' performances of various algorithms for sentiment analysis.

ii) Comparing different sentiment analysis algorithms for optimal performance

iii) Implementing different sentiment analysis algorithms to find the best one.

~~iv) Assessing product reviews~~

~~iv) Assessing customer reviews to~~ better

iv) Assessing sentiment analysis algorithms to harness online product reviews.

Su Mo Tu We Th Fr Sa

Date: 03/07/2021

Research question:

With respect to title (iv),

i) Which methodology is the most effective to analyze the sentiments from product reviews?

ii) Which algorithms is the most efficient when ~~terms of~~ analyzing sentiments from online product reviews?

iii) Which algorithms provide the best accuracies when it comes to analyzing sentiment from online reviews?

iv) Which algorithm is the most accurate to better analyze sentiment from online product reviews.

Su Mo Tu We Th Fr Sa

Date: 05 / 07 / 2021

Linear regression → Linear regression is for handling regression, continuous output

Logistic regression → ~~Linear~~ Logistic regression is for handling classification, discreet output

$$z = wx + b$$

$$\hat{y} = \sigma(wx+b)$$

$$= \sigma(z)$$

$$= \frac{1}{1+e^{-z}}$$

$$\frac{\partial L_{CE}(\hat{y}, y)}{\partial w_j} = [\sigma(wx+b) - y] x_j$$

$$\theta_{t+1} = \theta_t - \eta x$$

Su Mo Tu We Th Fr Sa

Date: 06/07/2021

Bag of words:

Step i) Other preprocessing

Step ii) Create vocabulary from unique words

Step iii) Create a matrix, also known as text vectorization

*Drawback is that the order of occurences is lost

* This is why create bigrams.

TF-IDF (Term Frequency - Inverse Document Frequency)

TF → Specifies how frequently a term appears.
— Calculates how many times it appears, with respect to total number of words.

IDF →
- whether a term is rare or frequent.
- Rare words have high IDF score.
- It's a log normalized value

Su Mo Tu We Th Fr Sa

Date: 08 /07 /2021

Naive Bayes:

Let document d =

classes $C = \{c_1, c_2, c_3 \dots\}$

m labeled documents. $(d_1 c_1, d_2 c_3, d_3 c_1 \dots)$

Output $\hat{h} : d \to c$

~~Bow: $\hat{h}(?) = \hat{c}$~~

Bow: $h(d) = \hat{c}$

$d = \{$ sweet : 3
happy : 6
sad : 3
sweet : 2 $\}$

For doc d, class c,

$$P(c|d) = \frac{P(d|c)\,P(c)}{P(d)}$$

$$c_{MAP} = \arg\max_c P(c|d)$$

$$= \arg\max_c \frac{P(d|c)\,P(c)}{P(d)}$$

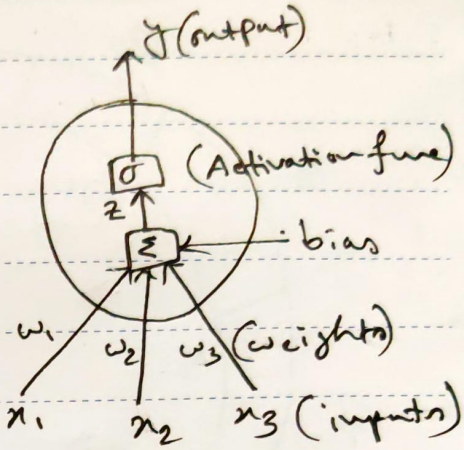
$$= \arg\max_c P(d|c)\,P(c)$$

$$= \arg\max_c P(x_1, x_2, x_3 \dots x_n|c)\,P(c)$$

Su Mo Tu We Th Fr Sa

Date: 09 / 07 / 2021

Neural net:



$$z = b + \sum w_i x_i$$

$$= w \cdot x + b$$

$$y = a = f(z)$$

$$y = \sigma(z) = \frac{1}{1+e^{-z}}$$

Other activation functions:

- tanh
- Softmax
- ReLU

* Sigmoid makes the value range from 0 to 1.

* $\tanh(y) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$

tanh makes the value range from −1 to 1

Su Mo Tu We Th Fr Sa

Date: 09/07/2021

* ReLU → z when z is positive, 0 otherwise.
Almost linear..

$$ReLU(y) = \max(z, 0)$$

Perceptron:

Binary output (0 or 1)
— no non linear activation function

$$- y = \begin{cases} 0, & \text{if } w.x+b \leq 0 \\ 1, & \text{if } w.x+b > 0 \end{cases}$$

AND, OR with Perceptron

AND

x_1	x_2	y
0	0	0
0	1	0
1	0	0
1	1	1

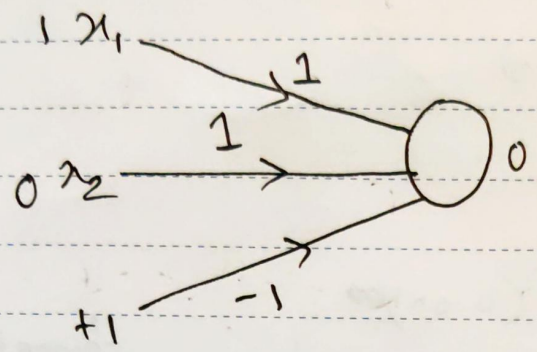
OR

x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	1

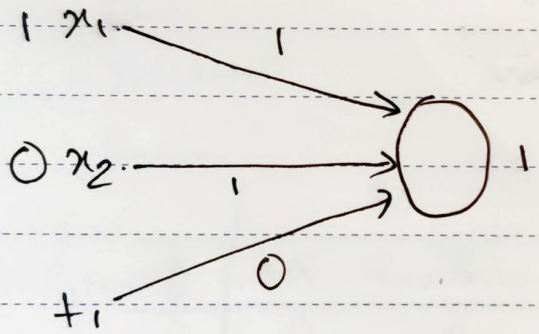
Su Mo Tu We Th Fr Sa

Date: 11 /07 / 2021

AND



OR



* XOR cannot be built with perceptron.

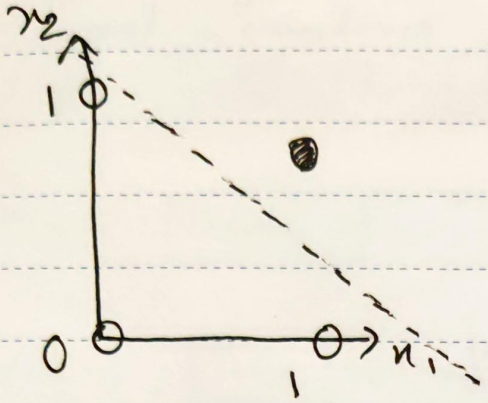
** Given input x_1, x_2, $w_1x_1 + w_2x_2 + b = 0$

* In standard linear format $x_2 = (-\frac{w_1}{w_2})x_1 + (\frac{-b}{w_2})$

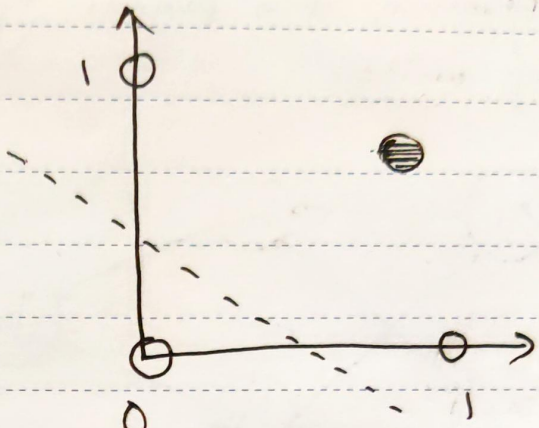
Su Mo Tu We Th Fr Sa

Date: 12 /07 /2021

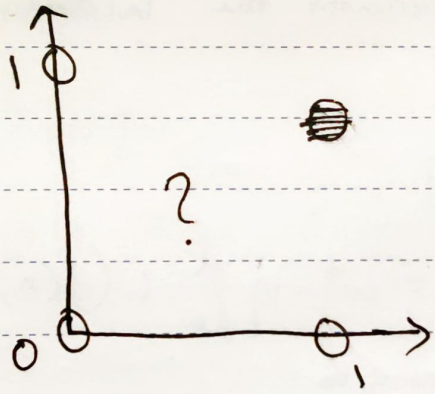
Decision boundaries:



a) x_1 AND x_2



b.) x_1 OR x_2



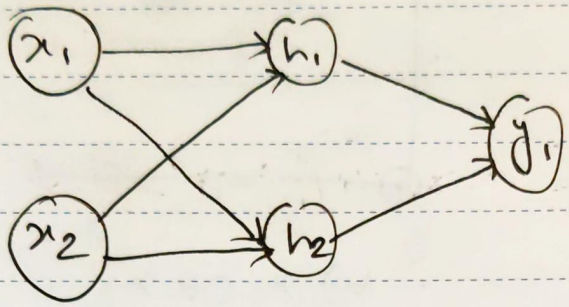
c.) x_1 XOR x_2

* We can see that XOR is not linearly separable.

Su Mo Tu We Th Fr Sa

Date: 13 / 07 / 2021

* XOR can't be calculated with single perceptron. However, it is possible by combining a layer of units.



* h_1 and h_2 represents the hidden layers.

Gradient descent:

$$w^{t+1} = w^t - \eta \frac{d}{dw} L(f(x; w), y)$$

For logistic regression

$$\frac{\partial L_{CE}(\hat{y}, y)}{\partial w_j} = [\sigma(wx+b) - y] x_j$$

Su Mo Tu We Th Fr Sa

Date: 14 / 02 / 2021

Bag of words (B.W) with N-grams:

N = 1 : This | is | a | sentence → unigrams

N = 2 : This is | is a | a sentence → bigrams

N = 3 : This is a | is a sentence → trigrams

* Last word in a N gram sentence given previous words.

$$P(w_n / w_{n-1} w_{n-2} \dots w_{n-N-1})$$

e.g. N = 2, $P(w_n | w_{n-1})$

N-gram sentence

$$P(w_n, w_{n-1}, w_{n-2} \dots w_{n-N-1})$$

e.g N = 2 $P(w_n, w_{n-1})$

Su Mo Tu We Th Fr Sa

Date: 16 /07/ 2021

Maximum likelyhood estimate,

$$P(w_i | w_{i-1}) = \frac{count(w_{i-1}, w_i)}{count(w_{i-1})}$$

E.g.:

<s> I am Aneftin </s>

<s> Aneftin I am </s>

<s> I do not like eggs and ham </s>

$P(I | <s>) = 0.67$, $P(Aneftin | <s>) = 0.33$

$P(am | I) = 0.67$, $P(Aneftin | am) = 0.5$

$P(</s> | Aneftin) = 0.5$, $P(do | I) = 0.33$

Random forest:

Let,

training set $X = x_1 \dots x_n$

" labels $Y = y_1 \dots y_n$

For, $b = 1 \dots B$ (bagging repeatedly B times)

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x')$$

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B} (f_b(x') - \hat{f})^2}{B-1}}$$

Su Mo Tu We Th Fr Sa

Date: 17/07/2021

Relationship to nearby neighbors:

$$\hat{y} = \sum_{i=1}^{n} w(x_i, x') y_i$$

Since, a forest averages the predictions of a set of m trees, with individual weight functions w_j, its predictions are:

$$\hat{y} = \frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{n} w_j(x_i, x') y_i = \sum_{i=1}^{n} \left(\frac{1}{m} \sum_{j=1}^{m} w_j(x_i, x') \right) y_i$$

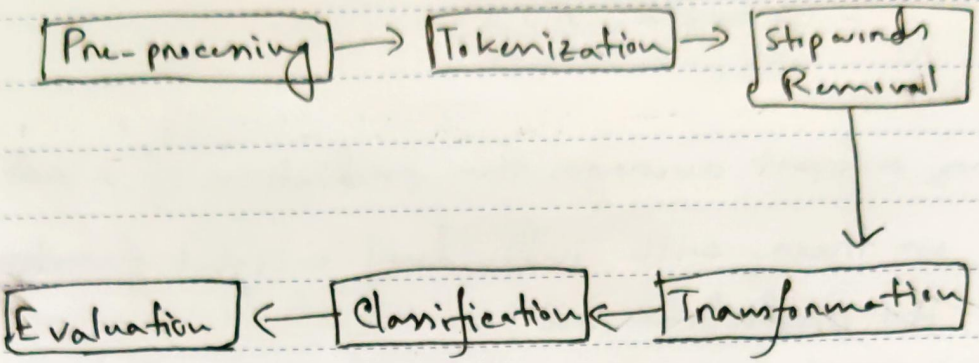
SGD: (Stochastic Gradient classifier):

* It's just a linear classifier that uses ~~SE~~ gradient descent on a loss function unlike logistic regression.

Su Mo Tu We Th Fr Sa

Date: 19/07/2021

Model:



(i) Pre-processing:

- Removing special characters, numbers, punctuations
- Tokenization
- Stemming
- Removing stop words

(*) Tokenization makes every word token

(*) Stemming converts the verb to its original form. E.g. - connected, connection, becomes connect.

Su Mo Tu We Th Fr Sa

Date: 20/02/2021

* Stopwords : are common English words, like, "a", "the", "are", "is"

By removing these the traing data becomes less ambiguous.

~~ttttt~~

Dataset: Amazon reviews (Labeled data)

Column names:

- ID
- ProductID
- User ID
- Profile Name
- Helpfulness Numerator
- Helpfulness Denominator
- Score
- Time
- Summary
- Text

Su Mo Tu We Th Fr Sa

Date: 21/07/2021

*Marking the dataset "Text" column based on Score.

- Score ≥ 4 ⇒ Positive
- Score == 3 ⇒ Neutral
- Score ≤ 2 ⇒ Negative

Visualizing label count:

Positive → 443777
Negative → 82037
Neutral → 42640

*Spliting the dataset into training and testing.

X_train.shape : (454763, 72127)
X_test. shape : (113691, 72127)
y_train.shape : (454763,)
y_test. shape : (113691,)

Su Mo Tu We Th Fr Sa

Date: 22/07/2021

i) Naive Bayes :-

Training accuracy : 0.834 [4th]

Testing accuracy : 0.828

ii) Stochastic Gradient Classifier (SGD) :-

Training accuracy : 0.867 [3rd]

Testing accuracy : 0.863

iii) Random Forest :-

Training accuracy : 0.999 [1st]

Testing accuracy : 0.886

iv) Logistic Regression :-

Training accuracy : 0.877 [2nd]

Testing accuracy : 0.867