MIDTERM EXAM

NAME: Shihab Muhtasim

ID : 21301610

FALL 23

section 2 Set

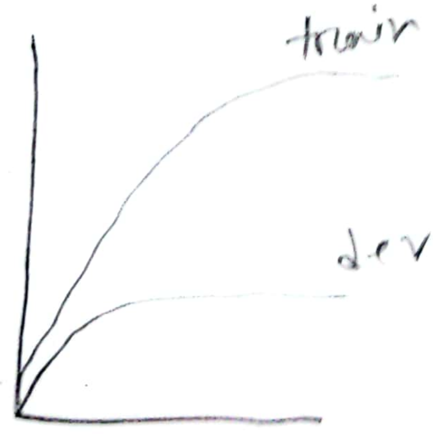
Ans to or 1(A)

Overfitting of a model is when a model is too complex for given data that it memorizes it's data and fails to detect patterns. Again, using higher models for simple data can cause overfitting.

We can solve it by:

① Using simpler models or making the complex model simpler by using less polynomials.

② We can feed the model less data so that it tries to learn



Overfitting performs good in training and bad during valid testing.

Ans to Q 1 (B)

TF : shallow appears 2 times

But appears 7 times.

$$TF(\text{shallow}; d) = \log_{10}(c(\text{shallow}, d) + 1)$$

$$= \log_{10}(2+1) = 0{\cdot}47$$

$$TF(\text{but}, d) = \log_{10}(7+1) = 0{\cdot}9$$

$$IDF(\text{shallow}) = \log_{10}\left(\frac{N}{df_{\text{shallow}}}\right)$$

$$= \log_{10}\left(\frac{10\,000}{1000}\right)$$

$$= 1$$

$$IDF(\text{but}) = \log_{10}\left(\frac{10000}{9000}\right)$$

$$= 0{\cdot}045$$

Weight of shallow = $0{\cdot}47 \times 1 = 0{\cdot}47$

weight of but = $0{\cdot}045 \times 0{\cdot}9 = 0{\cdot}0405$

∴ shallow is more important than but in this case.

Ans to 1 (c)

Accuracy is bad performance metrix because it often works bad for imbalanced datasets. for example if a huge data is of positive class and bad classifier predicts even negative classes to positives classes still it can give good accuracy. Hence, its missleading.

Ans to qr 2

(A)

Issues in english language tokenizations:

Tokenization is a process of splitting into words.

① In case of punctuations it can be hard to tokenize a word

② If certain words are connected and have space between them. Ex: New York

③ Hyphen between words can cause problems.

Ans to Q2B

A, Paris = [3, 4, 0, 1] Oslo = [1, 2, 1, 1]

B, France = [3, 2, 1, 0]

A - B = [0, 2, -1, 1]

A - B + sweden = [3, 3, 0, 2]

$$\cos(\text{oslo}, A-B+\text{sweden}) = \frac{3+6+0+2}{\sqrt{1+4+1+1}\,\sqrt{9+9+0+4}}$$

$$= 0.886$$

A - B + Norway = [3, 4, 2, 3]

$$\cos(\text{oslo}, A-B+\text{Norway}) = \frac{3+8+2+3}{\sqrt{7} \times \sqrt{9+16+4+9}}$$

$$= 0.981$$

$$oslo = [1, 2, 1, 1]$$

$$A - B + Findland = [1, 3, 0, 1]$$

$$\cos(oslo, A - B + Findland) = \frac{1+6+0+1}{\sqrt{7}\ \sqrt{1+9+1}}$$

$$= 0.911$$

$$A - B + Denmark = [0, 2, 0, 3]$$

$$\cos(oslo, A - B + Denmark) = \frac{0+4+0+3}{\sqrt{7}\ \sqrt{4+9}}$$

$$= 0.7337$$

∴ Ans is oslo is to norway.

$P = .$

Ans to qr no 3(A)

(A)

classifier A:

$$W_A = [1\ 1\ 3]^T = \begin{bmatrix}1\\1\\3\end{bmatrix}$$

$$X = \begin{bmatrix}1\\1\\0\end{bmatrix}$$

$$W_A^T = [1\ 1\ 3]$$

$$W_A^T X + b_A = [1\ 1\ 3] \times \begin{bmatrix}1\\1\\0\end{bmatrix} + 0$$

$$= 2 + 0 = 2$$

$$\sigma(2) = \frac{1}{1+e^{-2}} = 0.88 = y_A$$

classifier B:

$$W_B^T X + b_B = [1\ 2\ 0] \times \begin{bmatrix}1\\1\\0\end{bmatrix} + (-1)$$

$$= 3 - 1 = 2$$

$$\sigma(2) = 0.88 = y_B$$

x's original lable = 1

$$L_{CE} = -\log\left(\hat{y}^{y}(1-\hat{y})^{1-y}\right)$$

$$= -\log\left(0.88^{1}(1-0.88)^{1-1}\right)$$

$$= 0.055$$

∴ Both classifier will have 0.055 ~~lo~~cross entropy loss.

Ans to or 3(B)

Although it may be hard but it is possible. We can use ~~BOW~~ - term document frequency and ~~inver~~ inverse doc freq to analyze the sentiments of the reviews from dataset and then collect symantic information for each review. The words that ~~colle~~ have most important ~~in~~ words in each review we can set them to positive or neg class either by human annotation or by using lexicon list of word and then run a classifier on that dataset.