# CS 335-337: Assignment 3

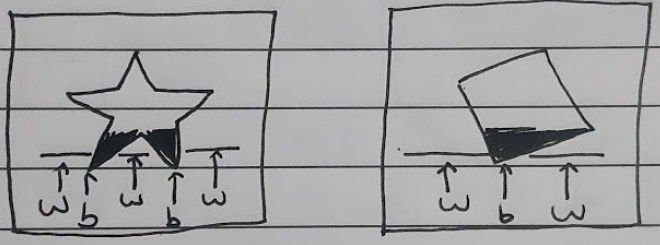
**Anish Deshpande - 180100013**

Date :

## Assignment 3

**1.1** b) I have used 5 features to classify the shapes.

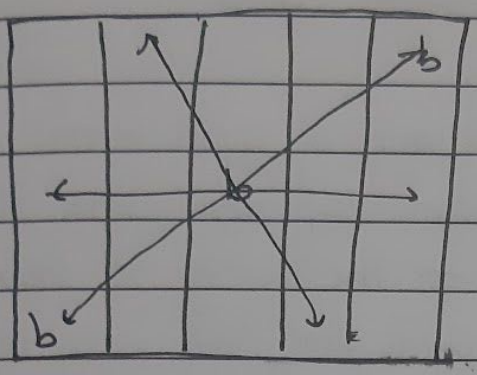
i) The first is the number of transitions from white to black or black to white in any given row. This number will be atmost 2 for any shape except a star, and for a star, there will be a few rows where there are 4 transitions.




ii) The next 4 features are all extracted from information from 10×10 patches surrounding each point.

* We take patches whose centre is on the shape.
* We want it to be a corner / edge, so only let 30/100 pixels be black, this excludes the interior of a shape.

* We distinguish between a corner and edge by taking a smaller 5×5 sub-patch, and seeing whether all points on some straight line are black:




If all points in a straight line are black, on any one line, then increment the 'on_edge' counter.

∴* The number of such 'on_edge' pixels is a feature.

After determining that it is not an edge point, we calculate the number of black pixels in the 10×10 patch.

The trend: num(black): star < triangle < square. circles are set to 0 because of the patch condition.

* The min, max, and average of these counts are our features.

corner  ← count num. of black pixels.

2.1

a) Categorical cross-entropy loss:

$$E(w) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_k^{(i)} \times \log\left(P(Y=k \mid W_k, \phi(x^{(i)}))\right)$$

$$\left(P(Y=k \mid W_k, \phi(x^i))\right)^X = \sigma_{W_k}(x^i) = \left(\frac{1}{1+e^{W_k^T \phi(x^i)}}\right)$$

In binary logistic regression, $k = \{1,2\}$ (only 2 classes). equivalent to 0,1

$$P(Y=k \mid W_k, \phi(x^i)) = \frac{e^{W_k^T \phi(x^i)}}{\sum_{k=1}^{K} e^{W_k^T \phi(x^i)}} \qquad \forall k \in 1 \ldots K$$

In the special case, we only have 2 categories $P(Y=1)$ and $P(Y=0)$.

Hence, we may write $P(Y=0)$ as $1 - P(Y=1)$.

also, say the class labels are 0 and 1

$\therefore y_i \in \{0,1\}$

$\therefore y^{(i)}$ is 1 if class label is 1

$(1-y^{(i)})$ is 1 if class label is 0

and vice versa. $(y_0^i = 1 - y_i^i \quad \forall i)$

$\therefore$ ~~$E(w) = -\frac{1}{N}\sum$~~ however, taking $\therefore$

$$P(Y=k \mid W_k, \phi(x^i)) = \frac{e^{W_k^T \phi(x^i)}}{1 + \sum_{k=1}^{K-1} e^{W_k^T \phi(x_i)}} \qquad \forall k \in 1 \ldots K-1$$

And $P(Y=K \mid W_k, \phi(x^i)) = \frac{1}{1+\sum_{k=1}^{K-1} e^{W_k^T \phi(x^i)}} \quad (k=K)$

This reduces the number of weight vectors / variables by 1, it removes redundancy.

$\therefore$ We have $P(Y=0 \mid W_k, \phi(x^i)) = \frac{e^{W_0^T \phi(x^i)}}{1+\sum_{k=0}^{0} e^{W_k^T \phi(x^i)}}$

( k runs from 0 to 0, as K=1 ( ~~not~~ class labels 0, 1))

$\therefore P(Y=0) = \frac{e^{W_0^T \phi(x^i)}}{1+e^{W_0^T \phi(x^i)}}$ and $P(Y=1) = \frac{1}{1+e^{W_0^T \phi(x^i)}}$

$\therefore E(W) = \frac{-1}{N} \sum_{i=1}^{N} \sum_{k=0}^{1} y_k^{(i)} \log\left(P(Y=k \mid W_k, \phi(x^i))\right)$

$$E(W) = \frac{-1}{N} \sum_{i=1}^{N} \left[ y_0^{(i)} \log\left(\frac{e^{W_0^T \phi(x^i)}}{1+e^{W_0^T \phi(x^i)}}\right) + y_1^{(i)} \log\left(\frac{1}{1+e^{W_0^T \phi(x^i)}}\right) \right]$$

$$E(W) = \frac{-1}{N} \sum_{i=1}^{N} \left[ (1-y_1^{(i)}) \log\left(\frac{1+e^{W_0^T \phi(x^i)}-1}{1+e^{W_0^T \phi(x^i)}}\right) + y_1^{(i)} \log\left(\frac{1}{1+e^{W_0^T \phi(x^i)}}\right) \right]$$

$\therefore \sigma_{W_0}(x^{(i)}) = \frac{1}{1+e^{W_0^T \phi(x^i)}}$

$$\therefore E(w) = \frac{-1}{N}\sum_{i=1}^{N} y_1^{(i)} \log \sigma_{w_0}(x^i) + (1-y_1^{(i)})\left(\log(1-\sigma_{w_0}(x^i)\right)$$

Generalising $y_1$ to $y$ and $w_0$ to $w$

$$E(w) = \frac{-1}{N}\sum_{i=1}^{N} y^{(i)} \log \sigma_w(x^i) + (1-y^{(i)}) \log(1-\sigma_w(x^i))$$

This is the required expression.
Hence proved.

b) $$E(w) = \frac{-1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_k^{(i)} \times \log(P(y=k \mid W_k, \phi(x^{(i)})))$$

to find $\frac{\partial E(w)}{\partial w} = \frac{\partial E(w)}{\partial z} \times \frac{\partial z}{\partial w}$ (chain rule)

Let $z = \phi(x).W \quad (N \times K)$

$\phi(x)$ is constant w.r.t $W$

$$\therefore \frac{\partial z}{\partial w} = \phi(x)^T$$

$$E(w) = \frac{-1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_k^{(i)} \log\left(\frac{e^{z_{ik}}}{\sum_{p=1}^{K} e^{z_{ip}}}\right)$$

$$E(w) = \frac{-1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_k^{(i)}\left[z_{ik} - \log\left(\sum_{p=1}^{K} e^{z_{ip}}\right)\right]$$

Date : _____

$$\therefore \frac{\partial E(w)}{\partial z_{ik}} = \frac{-1}{N}\left( y_k^{(i)} \left[ 1 - \frac{1}{\sum_{p=1}^{K} e^{z_{ik}}} \times e^{z_{ik}} \right] \right)$$

$$\frac{\partial E(w)}{\partial z_{ik}} = \frac{-1}{N}\left( y_k^{(i)} - \frac{y_k^{(i)} e^{z_{ik}}}{\sum_{p=1}^{K} e^{z_{ik}}} \right)$$

Say $Y'$ = N×K matrix of $y_k^{(i)}$s.

$$\frac{\partial E(w)}{\partial z} = \frac{-1}{N}\left( Y' - \text{softmax}(\text{weights}^{(w)}, \phi(x)) \right)$$

$$\frac{\partial E(w)}{\partial w} = \left(\frac{1}{N}\right)^{*} \phi(x)^{T} \left( \sigma_w(\phi(x)) - Y \right)$$

Where $\sigma_w(\phi(x)) = \text{softmax} = \dfrac{e^{\phi(x)W}}{\sum_{\text{columns}} e^{\phi(x)W}}$.

2.2

b) For binary logistic regression,
test accuracy obtained = ~~0.8525~~ [0.863]
Model 'M', which predicts '0' all the time,
test accuracy = ~~0.86~~ [0.84]

Clearly, accuracy is not a good metric. This is because of the nature of the data. It is extremely unlikely for the song to have made it to the Billboard Hot 100, so any model can cheat and just predict that it won't make it, for high accuracy.

c) For our model, F1 score
= 0.301 (excluding artist name)
= 0.522 (including artist name, preferred).
Consider the model M, which predicts 0 all the time.

$$F1\ score = \frac{2TP}{2TP + FN + FP}$$

$\because TP = 0$ and $FN > 0$

F1 score = 0

Clearly, F1 score is a great indicator of the model's performance, as it takes into consideration the extent of misclassification also: (FN, FP) terms, comparing those with the TPs.

Date :

An imbalanced class-distribution exists here, so false negatives and false positives are very important. So, F1 score is a good metric for model evaluation.

e) Perceptron accuracy on D1 ≈ 0.78
Logistic regression on D1, accuracy ≈ 0.84

Logistic regression outperforms the perceptron. They are both hyperplanes, but because of the non-linear nature of the logistic curve, it can fit better to real-world situations.
A linear model is more often than not unable to separate and classify actual data, like the fact that the relationship between the properties of a song and whether it makes it to the billboard top 100 is complex.