1800500023
BHASKAR GHARU

1.2 (b) Features used for classification are perimeter of the shape, max distance of the point on perimeter from center and min distance of the points on perimeter from the center

This Triplet of features are unique for different polygons and therefore very effective for this classification problem.

2.2 (a) Given Cross Entropy function

$$E(W) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_k^{(i)} \times \log\left(P(Y=k|w_k\phi(x))\right)$$

for binary classification we can set $K=2$

$$E(W) = \frac{-1}{N}\sum_{i=1}^{N}\left[y_1^{(i)} \log\left(P(Y=1|w_1^T\phi(x^{(i)}))\right) + y_2^{(i)} \log\left(P(Y=2|w_2^T\phi(x^{(i)}))\right)\right)$$

Note that $y_1^{(i)} + y_2^{(i)} = 1$

$$P(Y=1|w_1^T\phi(x^{(i)})) = \frac{e^{w_1^T\phi(x^{(i)})}}{e^{w_1^T\phi(x^{(i)})} + e^{w_2^T\phi(x^{(i)})}}$$

$$P(Y=2 \mid w_2, \phi(x^{(i)})) = 1 - P(Y=1 \mid w_1, \phi(x^{(i)}))$$

$$= \frac{e^{w_2^T \phi(x^{(i)})}}{e^{w_1^T \phi(x^{(i)})} + e^{w_2^T \phi(x^{(i)})}}$$

let $P(Y=1 \mid w_1, \phi(x^{(i)})) = \sigma_w(x^{(i)})$

hence $P(Y=2 \mid w_2, \phi(x^{(i)})) = 1 - \sigma_w(x^{(i)})$

so $E(w) = -\frac{1}{N} \sum_{i=1}^{N} \left[y_1^{(i)} \log(\sigma_w(x^{(i)})) + (1 - y_1^{(i)}) \log(1 - \sigma_w(x^{(i)})) \right]$

hence binary classification is a special case for multiclass cross entropy Error function.

2.1(b)

$$E(W) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{K} y_c^{(i)} \log\left(\frac{e^{w_c^T \phi(x^{(i)})}}{\sum_{j=1}^{K} e^{w_j^T \phi(x^{(i)})}} \right)$$

$$E(W) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{K} y_c^{(i)} \left[w_c^T \phi(x^{(i)}) - \log\left(\sum_{j=1}^{K} e^{w_j^T \phi(x^{(i)})} \right) \right]$$

$$\nabla_{w_l} E(W) = -\frac{1}{N} \sum_{i=1}^{N} \left(y_l^{(i)} \phi(x^{(i)}) - \frac{e^{w_l^T \phi(x^{(i)})}}{\sum_{j=1}^{K} e^{w_j^T \phi(x^{(i)})}} \phi(x^{(i)}) \right)$$

$$\nabla_{w_l} E(W) = \frac{1}{N} \left(\sum_{i=1}^{N} \left(P(Y=l \mid w_l, \phi(x^{(i)})) - y_l^{(i)} \right) \phi(x^{(i)}) \right)$$

$$\nabla(E(W)) = \frac{1}{N} \phi(x)^T (\text{Predict} - Y)$$

$$\text{Predict}_K^{(i)} = P(Y=K \mid w_K, \phi(x^{(i)}))$$

Y is one hot encoded

2.2 (a) Year of release is irrelevant to song popularity

songtitle, artistname, song ID, artist ID are also removed columns as they are not related to music quality

(b) Accuracy of a model always predicting 0 is 84.18 %.

This measure is not that effective as wrongly classified samples are not given enough weightage

(c) F1 score on test set = 0.301

F1 score for a model always predicting 0 will be 0 as number of true positives are 0 since Precision and Recall are affected by wrong classifications hence F1 score is a better metric than accuracy

2.2 (d) logistic regression

Test accuracy for $\mathcal{D}1$ = 84.67 %

" " & Perceptron on $\mathcal{D}1$ = [redacted] %
79.09 %

Perceptron does not try to maximise distance between decision boundary and points. Hence for a test set logistic regression is expected to perform better if it contains point closer to the true decision boundary.