Atahan İyiekici
21801985
CS-464 HW1

CS-464 Homework 1

Question 1)

Question 1.1

$$P(S_m|H) = 0.87 \qquad P(S_m|L) = 0.21 \qquad P(S_m|F) = 0.04$$

$$P(S_u|H) = 0.13 \qquad P(S_u|L) = 0.79 \qquad P(S_u|F) = 0.96$$

$$P(H) = 0.64 \qquad P(L) = 0.24 \qquad P(F) = 0.12$$

The probabilities above are given in the question, To find $P(S_m)$ we will use the equation below

$$P(S_m) = P(S_m|H)P(H) + P(S_m|H^c)P(H^c)$$

$$P(S_m) = P(S_m|H)P(H) + P(S_m|L)P(L) + P(S_m|F)P(F)$$

$$P(S_m) = (0.87)(0.64) + (0.21)(0.24) + (0.04)(0.12) = 0.612$$

Question 1.2

$$P(H|S_m) = \frac{P(S_m|H)P(H)}{P(S_m)}$$

$$P(H|S_m) = \frac{(0.87)(0.64)}{0.612} = 0.9098$$

Question 1.3

$$P(H|S_u) = \frac{P(S_u|H)P(H)}{P(S_u)}$$

$$= \frac{P(S_u|H)P(H)}{1 - P(S_m)}$$

$$P(H|S_u) = \frac{(0.13)(0.64)}{1 - 0.612} = 0.2144$$

Atahan İyiekici
21801985
CS-464 HW1

Question 2)

Question 2.1

Question 2.1.1

Distribution of classes in training data can be seen in the figure 1 below

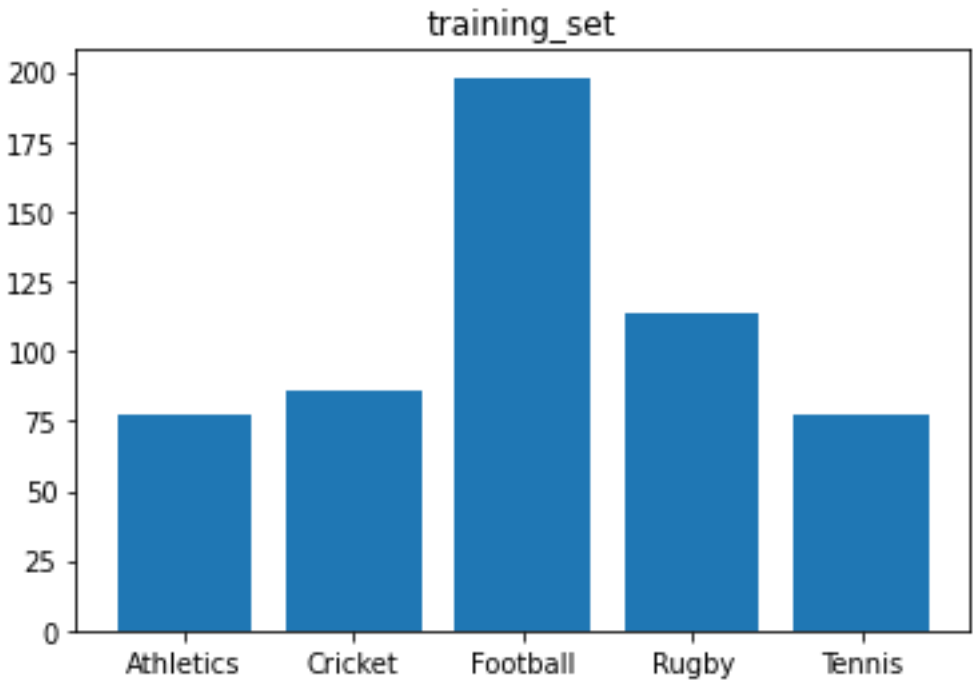

Figure 1: Distribution of the classes in traning set

The distribution of classes are 77, 86, 198, 114, 77 for the classes introduced above, respectively.

Question 2.1.2

The training dataset is not balanced. It can be seen in the Figure 1 that it skewed towards the class “Football”. Having an imbalanced training set affects model significantly. Since the model's prediction probabilities will be skewed towards the dominant class if the dataset is unbalanced, the model may predict in favor of the dominant class which is “Football” in this case. This is due to the fact that, after making a forecast about any non-dominating class, the likelihood of getting it wrong increases more than the probability of getting it right. Laplace Smoothing that we will do in the following parts will be a solution to decrease the fact of this problem and, it can increase the accuracy.

Question 2.1.3

Distribution of classes in validation data can be seen in the figure 2 below

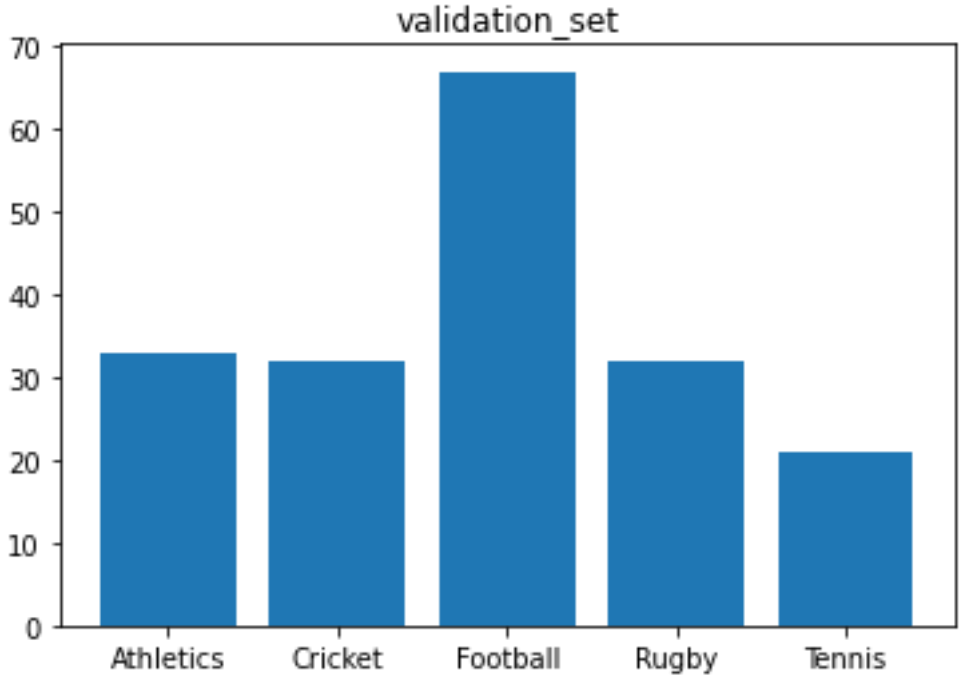

Figure 2: Distribution of the classes in validation set

The distribution of classes are 33, 32, 67, 32, 21 for the classes introduced above, respectively.

As it can be seen in the Figure 2, Even though, it looks very similar to training set distribution there are some differences between those. If we had a bad split, this situation arises a low accuracy problem. The term for the class probability (π) is misleading.

Question 2.1.4

Yes, the reported accuracy will be affected if there is a skew towards one class. The wrong or bad split of training and test sets could effect the accuracy. If the split of sets are proper and there was no skew towards any class, the accuracy would be much higher because the probability would be higher for the dominant class in this case and the bias factor would affect better.

Atahan İyiekici
21801985
CS-464 HW1

Question 2.2

$Predictions/Real - Class$	Class 0	Class 1	Class 2	Class 3	Class 4
Class 0	24	33	45	30	19
Class 1	0	5	0	0	0
Class 2	0	0	22	0	0
Class 3	0	0	0	3	0
Class 4	0	0	0	0	4

Accuracy without smoothing= 31.351%

Accuracy= 0.31351

True Prediction (out of 185)= 58

False Prediction (out of 185)= 127

Question 2.3

$Predictions/Real - Class$	Class 0	Class 1	Class 2	Class 3	Class 4
Class 0	25	0	0	0	0
Class 1	0	35	0	0	1
Class 2	0	0	67	0	0
Class 3	0	2	0	36	0
Class 4	0	0	0	0	22

Accuracy with smoothing= 97.297% (smoothing factor = 1)

Accuracy= 0.97297

True Prediction (out of 185)= 178

True Prediction (out of 185)= 8

Atahan İyiekici
21801985
CS-464 HW1

Question 2.4

The effect of Dirichlet prior α is huge, the accuracy difference is nearly 60 percent which is a great number.