Asam Mahmood – Z23050331

CAP 5615 Introduction to Neural Networks

Question 1 [1.5 pts: 0.25/each]: Please use your own language to briefly explain the following concepts (Must use your own language. No credit if descriptions are copied from external sources)

• Machine Learning: Machine learning is the study of software algorithms that are used to predict, analyze, label data and or to make decisions based on data from a particular use case. Machine learning is subset of Artificial Intelligence. Machine learning has many use cases throughout all industries. A machine learning algorithm consists of a model that is created based on data of a particular use case which is then used to create predictions, labeling of data, or decisions based on that data. The more data the better the model tends to do.

• Hyperplane classification: Hyper planes are like boundaries for decisions that correlate with the classification of data points. The data points that appear on each side of a hyperplane can be classified to different class depending on which side they are one. Also, hyperplanes are more than three dimensional spaces. The number of dimensions depend on the number of attributes/features. If its more than three features it will become a hyperplane.

• Overfitting: Overfitting is when we have higher order line which is when model will perfectly separate data points(the professor mentioned this in the video). Which means it is a badly performing model. It also means that the models behave good with the training data set. But on new test data sets it will perform very badly because it was trained so well with training data.

• Confusion Matrix: A confusion matrix is a table that displays that performance information of a classification model where you can see actual values and predicted values displayed. These values are the true positive, false positive, false negative, and true negative values.

• True positive rate and False positive rate:
True positive is what the model predicted as positive and what is positive in the results.
Other terms for true positive rate are recall or sensitivity. The formula for true positive rate is TPR = TP/TP +FN. ***FN*** stands for false negative values and ***TP*** stands for true positives. These values come from the confusion matrix values derived from the model

False Positive Rate is when the model classifies something as positive buts its truly negative thus making it a false positive. The values to calculate false positive rate also come from the confusion matrix. The formula for False positive rate is FPR = FP / FP + TN. ***FP*** stand for false positive, and ***TN*** stands for true negative.

• ROC (Receiver operating characteristic) Curve:

The ROC Curve is performance measurements for a model that solves a classification problem. It is graph that plots true positive rate and false positive rates. Some traits we can observe from a ROC graph is the area under the curve. The AUC closer to one means that is a very well performing model. The closer it gets to 0 means its not performing as well.

Asam Mahmood – Z23050331

Question 2 :

Figure 1 shows the scatter plot between salmon vs. sea bass, with respect to two features: width and lightness.

• Explain how machine learning algorithms can be trained to separate salmon from sea bass? [0.5 pt].

There are many solutions using machine learning algorithms that can separate salmon from sea bass. Professor went through this in video as well. A couple ways we can do this is by single feature selection(where we can use width of fish a feature and then plot length of fish with count), We can use multiple features(lightness and width of a fish). This will be a two-dimensional feature and we can plot width and lightness on two-dimensional space. Then we can use a decision boundary to separate them.

• Show your solutions (as lines, or math formula) and explain how many parameters need to be estimated in your model [0.5 pt.],

In my solution I would use a linear decision boundary or curve line with a two-dimensional feature.
x = (width, lightness)

Formula 1: $y = ax + b$ straight line (described in video)
Formula 2: ax^2+bx+c

• Explain how does your model classify a future fish as either salmon or seabass [0.5 pt.]

All dots on left of blue are in the salmon region and the right are sea bass region which is the red dots. I used feature width and lightness. Also, My rules are below for how model classifies results results.

R(x)= { salmon if (ax) + b < 0 classify as salmon}
{ seabass if (ax) + b > 0 classify as seabass }

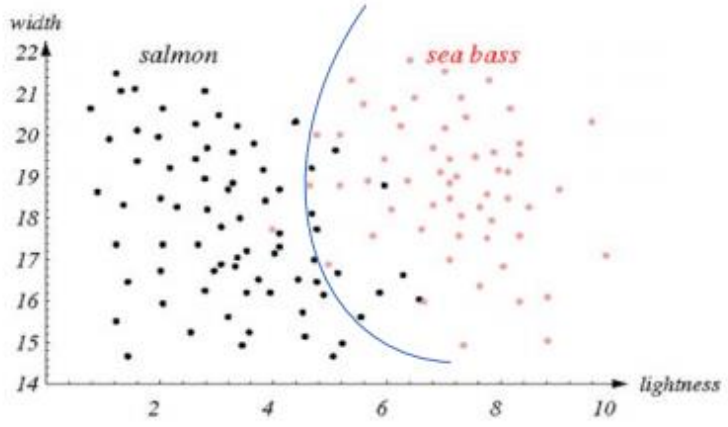


Figure 1

Question 3 [1 pt.]: Figure 2 shows the scatter plot between salmon vs. sea bass, with respect to two features: width and lightness.

• If the solid curve is used to separate salmon from sea bass, what will be the problem when classifying fish into correct category (e.g., the one showing as question mark) [0.5 pt].

Looking at where the question mark lives, we can say model will classify fish as sea bass above the curve but where the question marks live it falls in the salmon region and it will classify the is sea bass because of the type of problem that it is which is ***overfitting***. This happens when data is overly well trained against training data.

• How do you suggest improving the model to make better separation [0.5pt] (explain your solutions and draw lines if necessary).

To resolve this issue and improves the model I would use what the professor described and use linear decision boundary. This is KISS type of way to classify fish and predict the question mark fish as we can see below.

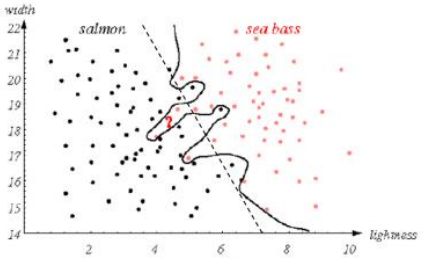


Figure 2

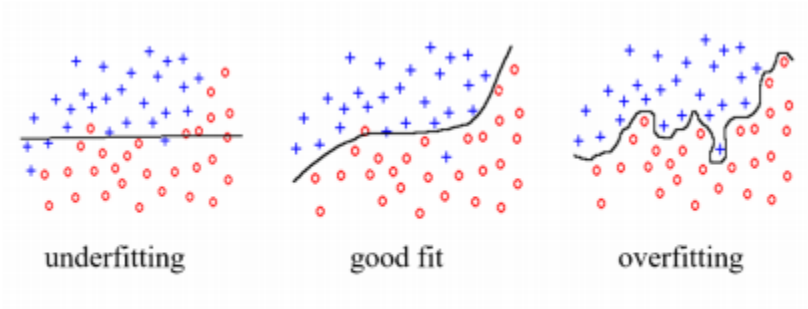


Figure 3

Question 4 [1 pt.]: Figure 3 shows three common scenarios when using decision boundaries to separate samples.

• Explain “under fitting” and “over fitting” [0.5 pt]

Underfitting: Underfitting is when the machine learning model fails to detect the pattern or trend of the data. We can see it happen because its just a straight line across the data in figure 3. It does not capture

the trend when the red dots start shifting upwards. Underfitting can be improved by adding more data to training and remove feature used by the feature selection professor went through in video.

Overfitting: Overfitting is when the model captures a 100% difference of data and fails to capture pattern/trend of data because it is to well-trained with data. You can see this in the example where it captures difference in all blue dots and red dots above. It classifies everything not detecting pattern or trend.

• Explain the risk and model complexity of each of them, respectively [0.5 pt]

Underfitting is very simple, and it happens when there is not enough training data and does not generalize and detect patterns correctly. The risk is that it will have high chances of unreliable predictions. This usually mean that it will have very large bias and very small variance. Overfitting is complex because the model considers a lot of features and determining which feature to give it a better generalization is complex. The overfitting model is not able to capture pattern/trend and is not able to have dependable results. It will have bigger variance and smaller bias because it has more features it considers.

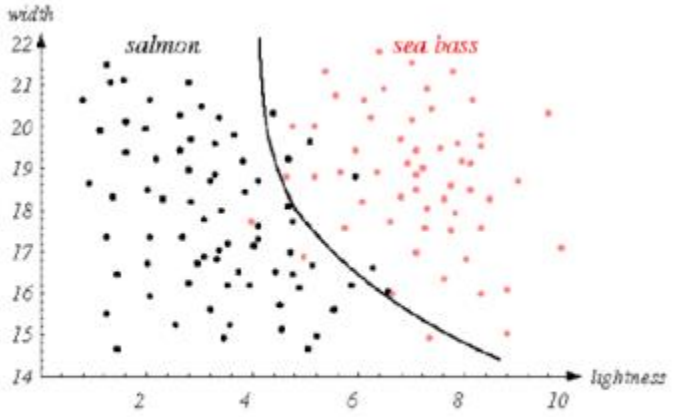


Figure 4

Question 5 [1 pt.]: Figure 4 shows the scatter plot between salmon vs. sea bass, with respect to two features: width and lightness. The black solid line is used to separate samples into two groups (salmon vs. sea bass)

• Explain how many instances will be misclassified, and mark misclassified samples [0.5 pt]

There will be three misclassified sea bass(red dots) that are in the salmon region

There are 6 salmon(black dots) misclassified in the sea bass region

The total amount misclassified are 9 instances in figure 4 above.

Asam Mahmood – Z23050331

• Assume market price of salmon is much higher than sea bass, so we must correctly classify salmon with a 100% accuracy (when classifier predicting a fish as a salmon, it must be a salmon), explain how should the classifier adjust the decision line in this situation [0.5 pt]

If we have to classify salmon a 100% then the black line over to right where it, I will encompass all black dots within the salmon region and all red dots to the sea bass region. When this happens, it will look like a overfitted graph.

		Gold Standard	
Test Outcomes		Genuine Salmon (positive)	Genuine Seabass (negative)
	Predicted Positive	55	12
	Predicted Negative	45	38

		Gold Standard	
Test Outcomes		Genuine Salmon (positive)	Genuine Seabass (negative)
	Predicted Positive	58	18
	Predicted Negative	42	32

		Gold Standard	
Test Outcomes		Genuine Salmon (positive)	Genuine Seabass (negative)
	Predicted Positive	67	22
	Predicted Negative	33	28

Figure 5

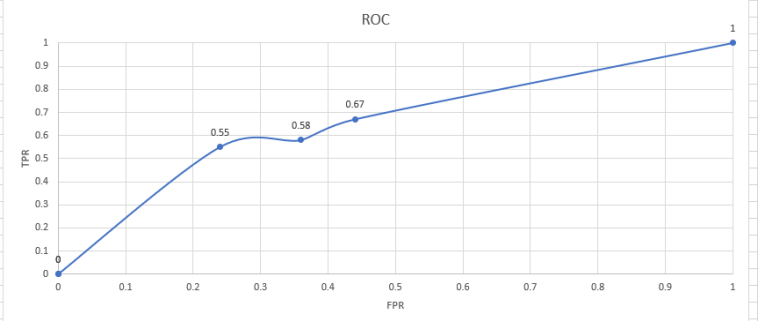
Question 6 [2 pts]: Figure 5 shows the confusion table of a classifier with respect to three testing conditions.

• Calculate classification accuracy of each test, respectively, and also calculate the average classification accuracy [0.5 pt].

• Calculate true positive rate (TPR) and false positive rate (FPR) of each test, and show the results on an ROC curve [0.5 pt].

Asam Mahmood									
		Gold Standard							
		Genuin Salmon (positive)	Genuin Seabass (negative)	True Positive Rate	False Postive Rate	Accuracy Formula	Error Rate	Average Accuracy	
Test Outcomes	Predicted Positive	TP	FP	TP/TP +FN	FP / FP + TN	TP+TN /TN+TP+FN+FP	1-accuarcy	.62+.64+.63333333 / 3	
	Predicted Negative	FN	TN					0.617777767	
		Gold Standard							
		Genuin Salmon (positive)	Genuin Seabass (negative)	True Positive Rate	False Postive Rate	Accuracy	Error Rate		
Test Outcomes	Predicted Positive	55	12	55/55+45	12/12+38	55+38/38+55+45+12	1-.62		
	Predicted Negative	45	38	0.55	0.24	0.62	0.38		
		Gold Standard							
		Genuin Salmon (positive)	Genuin Seabass (negative)	True Positive Rate	False Postive Rate	Accuracy	Error Rate		
Test Outcomes	Predicted Positive	58	18	58/58+42	18/18+32	58+32/58+18+42+32	1-.6		
	Predicted Negative	42	32	0.58	0.36	0.6	0.4		
		Gold Standard							
		Genuin Salmon (positive)	Genuin Seabass (negative)	True Positive Rate	False Postive Rate	Accuracy	Error Rate		
Test Outcomes	Predicted Positive	67	22	67/67+33	22/22+28	67+28/67+22+33+28	1-.6333		
	Predicted Negative	33	28	0.67	0.44	0.633333333	0.36666667		

Asam Mahmood – Z23050331



I did my calculations in excel in red you can see average accuracy for all.
In yellow you can see each individual classification accuracy.
In Blue you can see false positive rates calculated.
In Green you can see true positive rates calculated.
Above you can see graph for roc curve, and I attached csv with calculations

• Explain how to use ROC curve to evaluate the performance of a classifier [0.5 pt]

ROC is a performance measurement tool for evaluating a classifier. It comes in a form of a graph like above. A quick look over you can safely say that the higher positive rate and the lower false positive rate the better the classifier model performed. To compare performance between different classifiers some use the area under the curve to determine which performed better.

• Why ROC curve is better than classification accuracy in assessing the classifier performance [0.5 pt]
ROC is better than classification accuracy because roc takes into account the overall performance of classification model(it is a measure of how model performed in separating classes) . Accuracy works well when classes keeps the same on train and test sets. ROC works better regardless of balance.

Question 7 [2 pts]: A classifier is trained from a given training set, and is validated on a test set with 10 instances. The results generated from the classifier is shown in the following table, where the first column is the index of the test instances, the 2nd column denotes the predicted probability of the test instance belonging to class “1” (there are only two classes: “1” and “0”), and the third column denotes the true label of the test instance. Using decision rule: a test instance is classified as “1” if its predicted probability belonging to class “1” is greater or equal to 0.45, and “0” otherwise.

• Please report the confusion matrix of the classifier [0.5 pt].

• Please report the classification accuracy [0.5 pt], True Positive Rate [0.5 pt], and False Positive Rates [0.5 pt] of the classifier (assuming “1” is the positive).

Asam Mahmood – Z23050331

Index	Predicted probablitly belonging to 1	True Label
1	0.8	1
2	0.2	0
3	0.4	1
4	0.55	1
5	0.45	1
6	0.9	1
7	0.3	0
8	0.4	0
9	0.56	1
10	0.92	1

Question 7 Cont. According to requirements .45 and above is classified as 1 other wise its classified as 0.

Index	Predicted	True Label
1	1	1
2	0	0
3	0	1
4	1	1
5	1	1
6	1	1
7	0	0
8	0	0
9	1	1
10	1	1

	Actual Postive	Actual Negative
Predicted Positive	TP- 6	FP-0
Predicted Negative	FN-1	TN-3

TP: Above I counted 6 true positives which were index's were on 1 in column true table and the classified 1 in predicted column.
FN: Above I counted one false negative at index 3 where it was "1" in true label column and "0" in predicted column.
TN: Above I counted 3 true negatives where its classified "0" in true label and "0" in predicted column.
FP: I count 0 for false positive this is where its classified "0" in the true label column and "1" is classified in predicted column.

Asam Mahmood – Z23050331

Below is calculations (also calculations are in attached csv)

Index	Predicted probablity belonging to 1	True Label
1	1	1
2	0	0
3	0	1
4	1	1
5	1	1
6	1	1
7	0	0
8	0	0
9	1	1
10	1	1

	True Positive Rate	False Postive Rate	Accuracy Formula
	TP/TP +FN	FP / FP + TN	TP+TN /TN+TP+FN+FP

	Actual Postive	Actual Negative			
Predicted Positive	TP- 6	FP-0	6/6+1	0/0+3	6+3/3+6+1+0
Predicted Negative	FN-1	TN-3	0.857142857	0	0.9

Question 8 and 9 are in the attached html also calculations are in attached csv for further viewing.