CAP 5615 Introduction to Neural Networks

2021 Summer

Homework 1 [15 Pts, Due: May 29 2021. Late Penalty: -2/day]

[If two homework submissions are found to be similar to each other, both submissions will receive 0 grade]

- Homework solutions must be submitted through Canvas. No email submission is accepted.
- Please try to put all results in one file (e.g., one pdf or word file).
- If you have multiple files, please upload them separately (only **pdf**, **word**, and **html** files are allowed).
- You can always update your submissions. Only the latest version will be graded.]

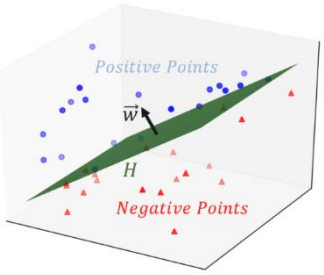
Question 1 [1.5 pts: 0.25/each]: Please use your own language to briefly explain the following concepts (Must use your own language. No credit if descriptions are copied from external sources):

- Machine Learning:

Machine learning is to program and teach computers to use historical data (e.g., past experience) to build models to solve intelligent tasks, such as face recognition, speech recognition. Etc.

- Hyperplane classification:

Hyperplane classification uses a hyperplane to classify samples into two groups, above the hyperplane or underneath the hyperplane. A hyperplane is determined by a linear combination of multiple features values (in a two dimensional space, this would be a linear line). For example, in a 3-d dimensional space, a hyperplane would be a 3-d flat surface.



$w_1 \cdot x_1 + w_2 \cdot x_2 + w_3 \cdot x_3 + b = 0$

A point x=(x1,x2,x3) is classified as Positive, if If $(w_1 \cdot x_1 + w_2 \cdot x_2 + w_3 \cdot x_3 + b) \geq 0$, or Negative, otherwise.

- Overfitting:

In machine learning, overfitting denotes that the trained learning model overfits to the training data, and has low accuracy on the test data. In other words, a learning model may have high accuracies on training data, but have much lower accuracies on test data. More specifically, given two models h and h1, if the error rate of h on the training data is less than the error rate of h1 on the same training data, but the error rate of h1 on the test data is higher than the error rate of h1 on the same training data, then we conclude that h overfit to the data.

- Confusion Matrix:

Confusion matrix consist of cross classification results to show the performance of a model. In a binary classification task, a confusion matrix consists of two rows and two columns. The columns indicate the number of samples belonging to each category (using their class labels), and the rows denote the number of samples being classified into corresponding categories. A confusion matrix is used to derive classification errors, true positive rates (TPR), false positive rates (FPR), and many other performance measures.

		Gold Standard	
Test Outcomes		Genuine Positive	Genuine Negative
	Predicted Positive	True Positive (TP)	False Positive (FP) (type I error)
	Predicted Negative	False Negative (FN) (type II error)	True Negative (TN)

		Gold Standard	
Test Outcomes		Genuine Salmon (positive)	Genuine Seabass (negative)
	Predicted Positive	63	28
	Predicted Negative	37	72

- True positive rate and False positive rate:

True positive rate (TPR) is equal to the true positive (TP) divided by the sum of the true positive and false negative TPR=TP/(TP+FN).
False positive rate (TPR) is equal to the false positive (FP) divided by the sum of the true negative and false positive FPR=FP/(FP+TN).

- ROC (Receiver operating characteristic) Curve:

A ROC curve is a performance metrics to evaluate the performance of classifiers with respect to various test conditions. The x-axis of the ROC covers is the false positive rate, and the y-axis denotes the true positive rate. The best classifier will have its TPR being 1 and its FPR being 0. A random classifier will have 0.5 TPR and 0.5 FPR values. The areas under the roc curve is called AUC, which is commonly used to evaluate the classifier performance.

Question 2 [1.5 pts]: 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].

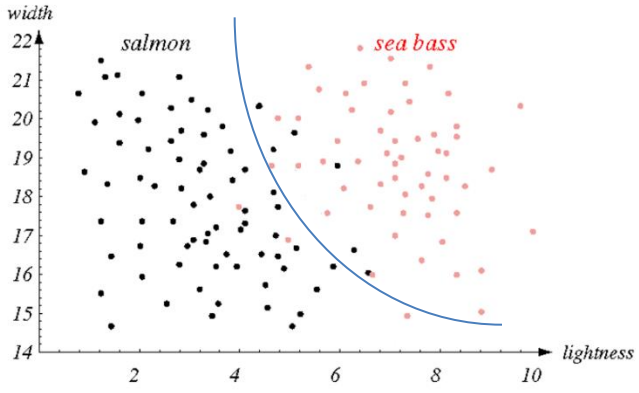
A machine learning method methods can be trained to use a line (e.g., a straight line, a quadratic function, or a higher order function), to separate seabass (red dots) and salmon (black dots) into two regions. One region is below the line, and the other region is above the line, which results in a binary separation of two types.

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

An example of the solution is showing as follows. A quadratic function $y=ax^2+bx+c$ defines the blue line. In this context of width vs. lightness features, we have

FishType=a×**width**^{2}+ b×**lightness**+c

The quadratic function has three parameters (a, b, c), which need to be estimated



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

Based on the defined model, for a fish with width and lightness (w, l) the classification is as following:

If {a×**w**^{2}+ b×**l** +c} ≥ 0, classify the fish as Sea bass
Otherwise, classify the fish as Salmon

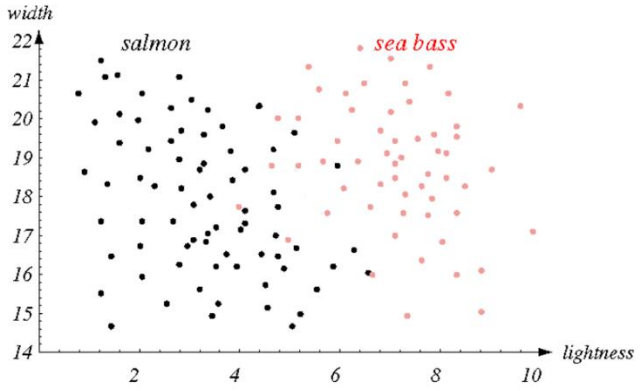


Figure 1

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

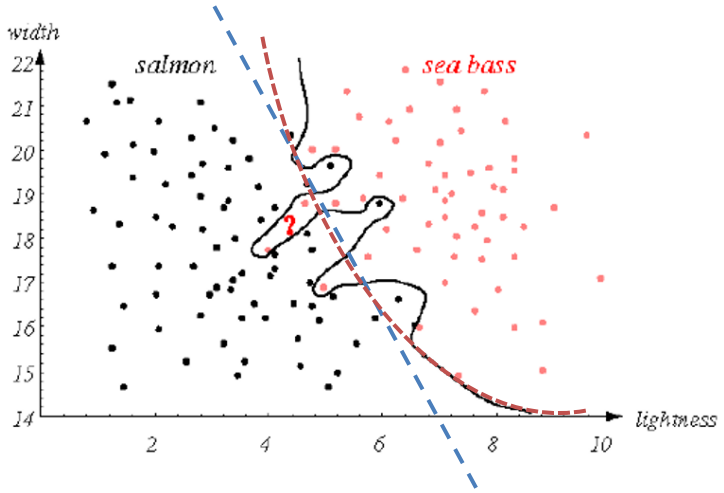
- If the solid curve is used to separate salmon from sea bass, what will be the problem when classifying fish into correct category (e.g., the one showing as question mark) [0.5 pt].

According to the sloid curve, the fish (showing as question mark) will be classified as sea bass, because it is above the decision boundary (the solid line). However, according to the actual distributions of samples, the question mark should be classified as a salmon, because there are more salmons close to the question mark point.

The above problem is caused by over fitting, because an overly sophisticated decision curve is trained to classify all training samples with 100% accuracy (overfitting). Because the decision overfit to the data (i.e, too specific to the training data), it cannot be generalized very well to classify new instances.

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

Overfitting can be significantly reduced by using simpler decision boundaries. In the context of salmon and seabass separation showing in Figure 2, we can use a straight line (showing as blue dashed line) or a quadratic function (showing as brown dashed line). Both lines are much simpler (using much less parameters), and can be generalized to correctly predict the question mark fish.



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]

Under fitting (or underfitting) means that the learned decision boundary is overly simplified, and does not have sufficient specificity to separate the training sample. In Figure 3, a straight line is used to separate blue cross from red circle. This straight line is too simply, and does not capture the distributions of some red circles on the top corner. An under fitting model is often caused by simple model (e.g., a very small decision tree, or a neural network with very few number of nodes).

Overfitting means that the learned decision boundary is overly complicated, and is trying to correctly classify every single instance in the training data (i.e, too specific to the training samples). In Figure 3, a higher order curve is used to separate blue cross from red circle. This curve is too complicated, and is misled by outliers (e.g., the blue cross). An overfitting model is often caused by complex model (such as a very deep and large decision tree, or a neural network with many layers and many nodes).

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

An underfitting model is often too simple, and has very few parameters. The risk of underfitting is that it cannot capture the genuine distribution of the training data, so it's overall classification accuracy is low.

An overfitting model is often complex, and has many parameters. The risk of overfitting is that it was over fit to the training, so it cannot make correct prediction on test instances, if the test

instances are slightly different from training samples. As a result, an overfitting model also has a low classification accuracy.

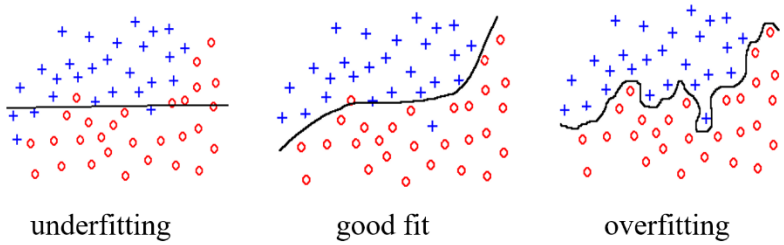


Figure 3

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]

The solid line represents the decision boundaries to separate salmon (black dots) vs. sea bass (red dots). The decision is formed using following rules: If a point is above the line, it is classified as a sea bass, otherwise, it is classified as a salmon. Based on this rule, if a sea bass is underneath the solid line, it will be misclassified. Similarly, if a salmon is above the solid line it will be misclassified. In total, nine instances (3 sea bass and 6 salmons) will be misclassified

- 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]

To ensure that when a classifier predicting a fish as a salmon, it must be a salmon with 100% accuracy, the decision boundary should be adjusted such that no sea bass is underneath the decision boundaries. An example is shown as below, where the original decision boundaries are moving towards the left to form new decision boundaries (blue line)

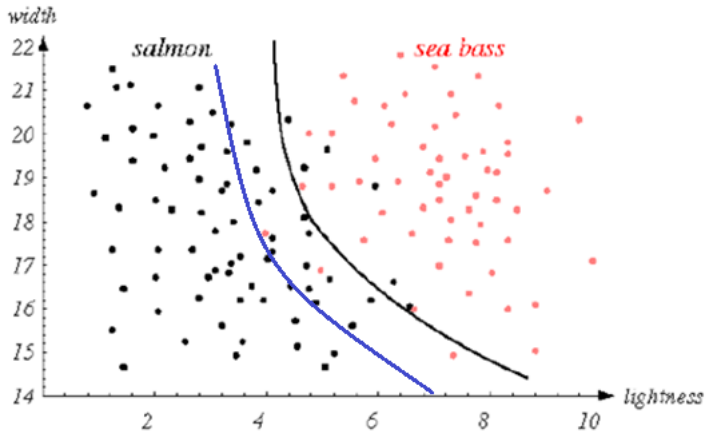


Figure 4

Question 6 [2 pts]: Figure 5 shows the confusion table of a classifier with respect to three testing conditions. **(No partial credit)**

- 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].
- Explain how to use ROC curve to evaluate the performance of a classifier [0.5 pt]
- Why ROC curve is better than classification accuracy in assessing the classifier performance [0.5 pt]

		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**

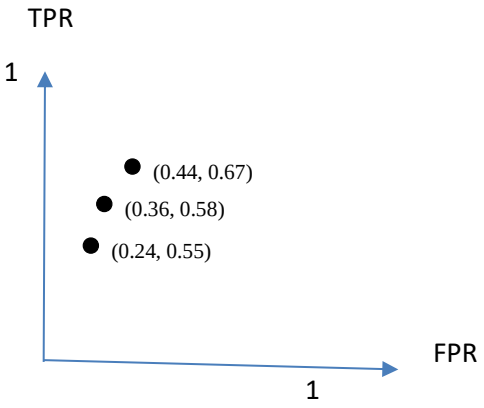
		Gold Standard	
Test Outcomes		Genuine Salmon (positive)	Genuine Seabass (negative)
	Predicted Positive	**67**	**22**
	Predicted Negative	**33**	**28**

Figure 5

Acc1=(55+38)/(55+12+45+38)=0.62
Acc1=(58+32)/(58+18+42+32)=0.6
Acc1=(67+28)/(67+22+33+28)=0.633
Average Accuracy: 0.6177 (or 61.77%)

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

FPR1=12/(12+38)=0.24 TPR1=55/(45+55)=0.55
FPR2=18/(18+32)=0.36 TPR2=58/(58+52)=0.58
FPR3=22/(22+28)=0.44 TPR3=67/(67+33)=0.67



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

ROC curve shows the performance of the classifier with respect to different TPR and FPR values. The higher the TPR and the lower the FPR values, the better the performance of the classifier. In order to compare classifier performance using ROC curve, we can use area under the ROC curve. The maximum area is 1 meaning the best performance classifier. A random classifier will have 0.5 area values under the ROC curve

Why ROC curve is better than classification accuracy in assessing the classifier performance [0.5 pt]

ROC curve is better than classification accuracy, because ROC curves show the overall performance of the classifier in classifying positive and negative samples. If the dataset has very few positive sample, e.g., 99 negative sample and 1 positive samples, a classifier can simply classify all instances as negative resulting in 99% classification accuracy. However, this classifier is useless, because it cannot capture positive sample. In this case, Accuracy is not a good indicator (measure). Alternative, for the same classifier, its TPR value is 0/1=0, and the

FPR value is 0/99=0. This will place the point at origin of the ROC curve, indicating that the classifier is not good for classification.

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).

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	

(No partial credit)

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

Confusion Matrix

		Predicted Lael		
		0	1	
True Label	0	3	0	
	1	1	6	

Accuracy=(3+6)/(3+0+1+6)=0.9
True Positive Rate=6/(6+1)=0.857
False Positive Rate=0/(0+3)=0

For all programming tasks, please submit the Notebook as html files for grading (the notebook must include scrips/code and the results of the script).

Question 8 [3 pts]: Please download housing.header.txt dataset from Canvas, and use a programming language (Python, R, etc.) to implement tasks below (a brief description of this dataset is available from the following URL)

https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html

- Report all samples with respect to the Crim index on a plot (the x-axis show the index of the sample, and the y-axis shows the crim index of the sample).
- Show both histogram of the Crim index and the density of the Crim index on a 1x2 frame.
- Show following four scatter plots in one frame (1x4), crim-medv, Rm-medv, Age-medv,Tax-medv, and explain how are they (Crim, Rm, Age, Tax) correlated to the medium house value (Medv)
- Create a subset which only includes properties with Crim less than 1 (inclusive), and Rm greater than 6 (inclusive).
- Show a scatter plot between Rm and Medv (x-axis shows Rm and y-axis denote Medv), please color all properties with 7 or larger Rm values as “red”, and rest properties as “black”.
- Report the pairwise correlation between every two variables (either as a matrix or as a level plot)
- Please explain which variable is mostly positively correlated to Medv (medium house value), and which variable is mostly negatively correlated to Medv.
- Draw scatterplots to show relationship between each attribute and Medv, respectively.
- Explain how to use scatterplots to find attributes which are positively correlated, negatively correlated, or independent of Medv, respectively

Question 9 [2 pts]: Please download movie100.csv dataset from Canvas, and use a programming language (Python, R, etc.) to implement tasks below. The movie100.csv includes 100 move reviews (from IMDB) and the sentiment (positive vs. negative) of the reviewer (there are 50 positive reviews and 50 negative reviews). Each review (each row) contains two part. The first column is the reviews (text), and the second column is the sentiment (positive or negative).

- Read the movie100.csv, and check average length (in terms of number of words) of the positive and negative reviews, respectively [0.5 pt]
- Check the frequency of the words with respect to positive and negative reviews, respectively. Print the top 20 most frequent words (including words and their frequency), for positive and negative reviews, respectively [1 pt]
- Draw a bar plot to show the frequency of the two-20 most frequent words (the x-axis denotes words, and the y-axis denotes their frequency), for positive and negative reviews, respectively [0.5 pt]