

Final Exam

December 18, 2019

Name:

Student Number:

Instructions

- DO NOT TURN OVER THIS PAGE UNTIL YOUR ARE TOLD TO DO SO
- Write your name and student number in the boxes provided.
- You have 2 hours to complete this exam.
- Write your solutions in the space provided. We will only consider work written within these pages.
- This exam has 8 problems that are not necessarily in order of difficulty. The problems are worth 100 points total.
- A correct answer does not guarantee full credit and a wrong answer does not guarantee loss of credit. You should concisely indicate your reasoning and show all relevant work. The grade on each problem is based on our judgment of your level of understanding as reflected by what you have written.
- A basic, non-programmable calculator is permitted.
- Write clearly! If we can't read it, we can't grade it.

Problem 1 - Feature selection - 8 marks

1. Before training a model, we typically remove all correlated features. Give two reasons why we do this. (2 marks)
2. If we provide a K-nearest neighbor model with a large number of features and some of those features are not important, the model will perform poorly. Why? (2 marks)
3. List two models (that we learned in this course) that have internal feature selection methods and describe how they internally select features. (4 marks)

Problem 2 - Model interpretation and explanation - 14 marks

Dep. Variable:	TenYearCHD	No. Observations:	2924			
Model:	Logit	Df Residuals:	2908			
Method:	MLE	Df Model:	15			
Date:	Sun, 12 May 2019	Pseudo R-squ.:	0.1207			
Time:	12:21:57	Log-Likelihood:	-1102.6			
converged:	True	LL-Null:	-1253.9			
		LLR p-value:	1.612e-55			
	coef	std err	z	P> z	[0.025	0.975]
const	-8.4105	0.805	-10.444	0.000	-9.989	-6.832
male	0.5351	0.122	4.400	0.000	0.297	0.773
age	0.0693	0.008	9.224	0.000	0.055	0.084
education	-0.0485	0.055	-0.888	0.375	-0.156	0.059
currentSmoker	0.1305	0.174	0.750	0.453	-0.211	0.472
cigsPerDay	0.0176	0.007	2.542	0.011	0.004	0.031
BPMeds	0.3125	0.272	1.147	0.251	-0.222	0.847
prevalentStroke	0.6820	0.597	1.143	0.253	-0.488	1.852
prevalentHyp	0.2834	0.154	1.840	0.066	-0.018	0.585
diabetes	0.1162	0.342	0.340	0.734	-0.553	0.786
totChol	0.0021	0.001	1.678	0.093	-0.000	0.005
sysBP	0.0125	0.004	2.896	0.004	0.004	0.021
diaBP	-0.0024	0.007	-0.329	0.742	-0.017	0.012
BMI	0.0046	0.014	0.321	0.748	-0.024	0.033
heartRate	-0.0012	0.005	-0.261	0.794	-0.010	0.008
glucose	0.0065	0.002	2.642	0.008	0.002	0.011

Figure 1: Statsmodel output.

Consider the output in Figure 1. Use a statistical significance level of 0.05 and answer the following questions

1. What type of model has been fit? (1 mark)
2. What is the target? (1 mark)
3. Which features are positively associated with the target? (2 marks)

4. Which features are negatively associated with the target? (2 marks)
5. What is the interpretation for the `cigsPerDay` variable? (2 marks)

Dep. Variable:	TenYearCHD	No. Observations:	2924			
Model:	Logit	Df Residuals:	2917			
Method:	MLE	Df Model:	6			
Date:	Fri, 15 Mar 2019	Pseudo R-squ.:	0.1121			
Time:	12:40:55	Log-Likelihood:	-1113.3			
converged:	True	LL-Null:	-1253.9			
		LLR p-value:	8.568e-58			
	coef	std err	z	P> z	[0.025	0.975]
const	-8.6870	0.521	-16.674	0.000	-9.708	-7.666
age	0.0698	0.007	9.703	0.000	0.056	0.084
male	0.6078	0.115	5.306	0.000	0.383	0.832
diabetes	0.6943	0.253	2.747	0.006	0.199	1.190
sysBP	0.0170	0.002	7.092	0.000	0.012	0.022
totChol	0.0022	0.001	1.748	0.081	-0.000	0.005
currentSmoker	0.4242	0.116	3.666	0.000	0.197	0.651

Figure 2: Statsmodel output.

Consider the output in Figure 2. Answer the following questions.

6. What is the target value for a 55 year old non-smoking male with diabetes, a systolic BP of 140, and a total cholesterol of 110? (2 mark)
7. If the person in question 6 were to start smoking, by how much does his target value increase? (1 mark)

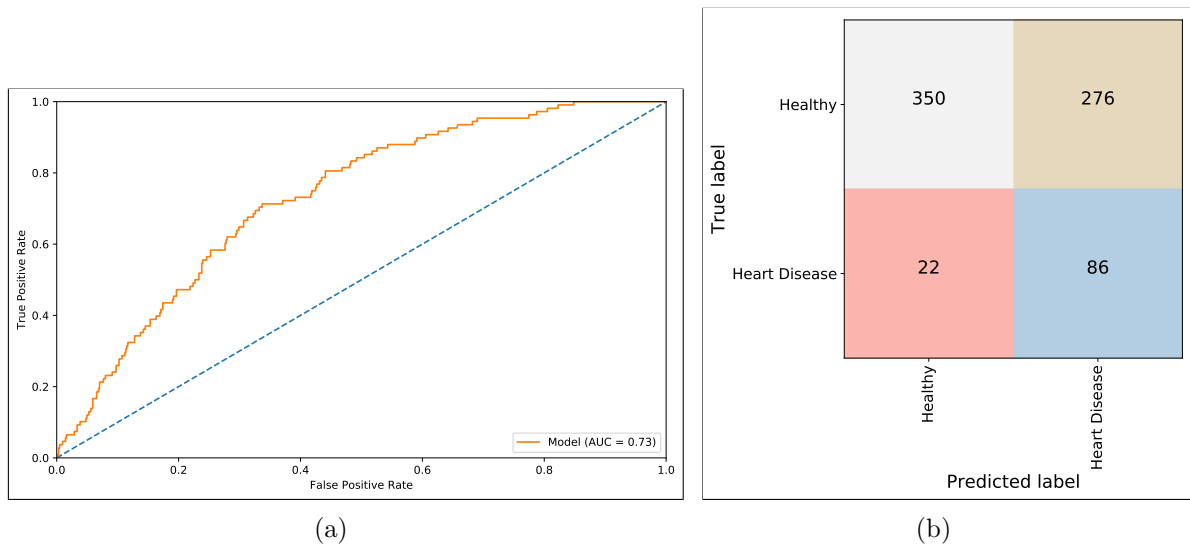


Figure 3: (a) ROC curve for the model in Figure 2, and (b) corresponding confusion matrix for a pre-determined threshold.

Consider Figure 3 and answer the following questions.

8. What does the dashed line represent in Figure 3(a)? (1 mark)
9. Calculate the false negative rate using Figure 3(b). (1 mark)
10. Indicate the point on the ROC curve shown in Figure 3(a) that corresponds to the confusion matrix in Figure 3(b). (1 mark)

Problem 3 - Clustering - 10 Marks

1. Why is it important to normalize the data before clustering? (2 marks)
2. Explain the difference between k-means and hierarchical clustering. (2 marks)

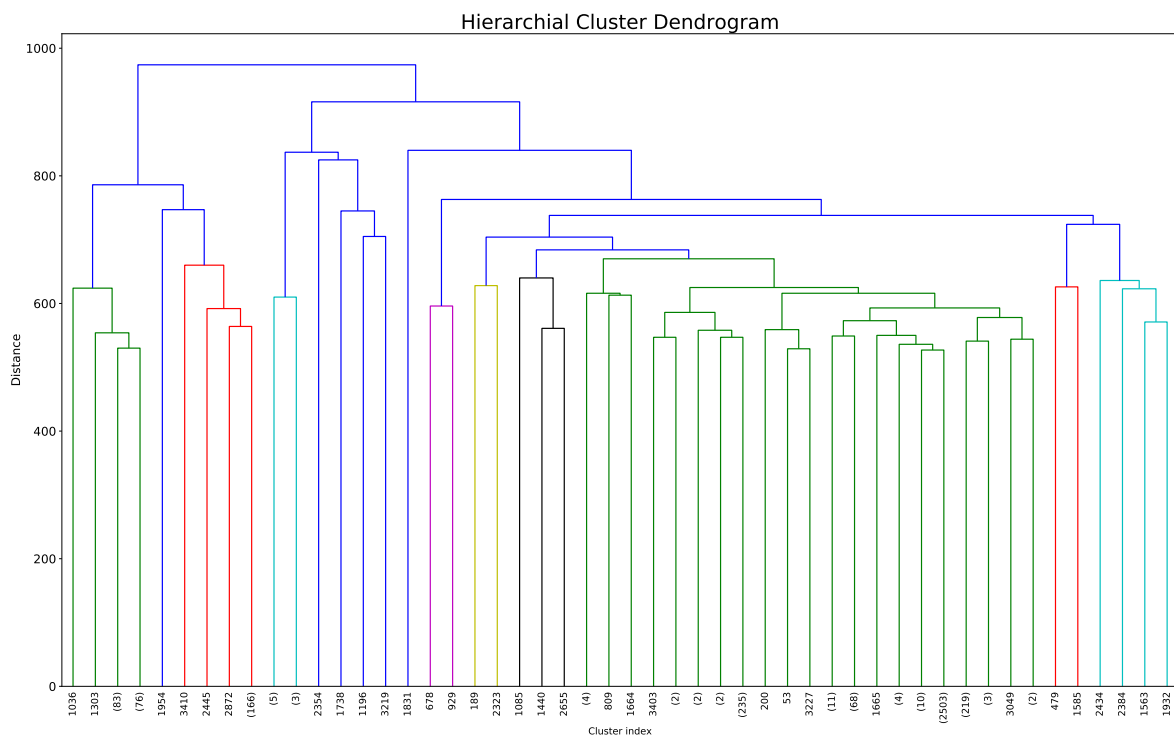


Figure 4: Hierarchical clustering dendrogram.

Consider the dendrogram shown in Figure 4 and answer the following question.

3. How many clusters are formed if the max inter-cluster distance is 800? (1 mark)

	Drama	Fantasy	Film-Noir	Horror	Musical	Mystery	Romance	Sci-Fi
0	0.0588	0.5091	0.0000	0.0041	0.1724	0.0183	0.0669	0.1095
1	1.0000	0.0336	0.0203	0.0226	0.0324	0.0418	0.2123	0.0312
2	0.2752	0.0357	0.0014	0.0533	0.0500	0.0186	0.0000	0.0343
3	0.1288	0.1318	0.0030	0.0942	0.0010	0.0387	0.0396	0.3637
4	0.6096	0.0334	0.0416	0.2698	0.0024	0.2453	0.0619	0.1011
5	0.3750	0.0554	0.0012	0.0094	0.0767	0.0130	1.0000	0.0177
6	0.0000	0.0237	0.0260	0.3059	0.0587	0.0418	0.0045	0.0903

Figure 5: K-means cluster centroids.

Suppose we run K-means clustering on a movie dataset with 8 features, where each feature represents a genre. We find 7 clusters with the centroids shown in Figure 5.

4. What type of movies can you expect to find in cluster 0? (1 mark)
5. A new movie is classified only as a drama and romance. What cluster does it belong in and what is the euclidean distance between the new movie and its cluster? (2 marks)
6. If we run the K-means algorithm 10 times, should we expect to find the same clusters each time? Why or why not? (2 marks)

Problem 6 - SVM and deep learning - 20 marks

1. What is the mathematical formulation/definition for the perceptron? (2 marks)
2. Provide a graphical representation/visualization of the perceptron. Label each component. (4 marks)
3. What is a support vector? Provide a graphical representation. (4 marks)

4. What is the difference between an artificial neural network (ANN) and a convolutional neural network (CNN)? (2 marks)
5. What is the difference between a feed forward neural network and a recurrent neural network? (2 marks)
6. What type of practical application is particularly suited to a recurrent neural network? Why? (2 marks)
7. What is the role of the activation function? List one activation function and draw it. (4 marks)

Problem 7 - Model training - 6 marks

We learned the following algorithms this year:

- Linear and logistic regression
- k-means and hierarchical clustering
- K-nearest neighbors
- CART and random forest
- Ensembles (boosting, bagging, stacking)
- SVM
- Neural networks and deep learning

1. List the methods that are typically solved to optimality (i.e., we do not use a heuristic). (2 marks)

2. Provide two advantages and two disadvantages of an optimal model. (4 marks)

Problem 8 - 2 marks

What is your favorite machine learning model and why?