Student Number | |

The University of Melbourne
School of Computing and Information Systems

Final Examination, Semester 1, 2018
COMP30027 Machine Learning

Reading Time: 15 minutes. **Writing Time:** 2 hours.

This paper has 6 pages including this cover page.

Instructions to Invigilators:

> **Students should be provided with script books, and should answer all questions in the provided script book.** Students may not remove any part of the examination paper from the examination room.

Instructions to Students:

- There are 9 questions in the exam worth a total of 80 marks, making up 60% of the total assessment for the subject. Note that all questions should be answered, and questions are not of equal value.
- All questions should be interpretted as referring to the concepts as described in this subject, whether or not it is explicitly stated. Unless otherwise stated, you are required to show all working for numerical questions; please indicate your final answer clearly.
- Please answer all questions on the ruled pages in the script book provided, starting each numbered question on a new page.
- Please write your student ID in the space above and also on the front of each script book you use. When you are finished, place the exam paper inside the front cover of the script book.
- Your writing should be clear; illegible answers will not be marked.

Authorised Materials: No materials are authorised.

Calculators: Students are permitted to use calculators.

Library: This paper may be held by the Baillieu Library.

Examiners' use only									
1	2	3	4	5	6	7	8	9	Total
21	9	5	9	7	5	7	4	13	80

Section A: Short Answer Questions [21 marks]

Answer each of the questions in this section as briefly as possible. Expect to answer each sub-question in a couple of lines.

Question 1: Short Answer Questions [21 marks]

1. Explain the difference between "supervised" and "unsupervised" learning, and give an example of a typical method for each. [2 marks]
2. What is the relationship between "instances" and "attributes" (also known as "features")? [1 mark]
3. Some Machine Learning methods rely on having "numerical" data. Give an example of how we can use such a learner with "categorical" data. [1 mark]
4. When might the "softmax" function be used in a Machine Learning context? [1 mark]
5. What might it look like, if we had a Machine Learning system with low "model bias", but high "evaluation bias"? Why would such a situation be undesirable? [3 marks]
6. Under what circumstances would a "neural network" be equivalent to a "logistic regression" model? [2 marks]
7. How is "active learning" similar to "self-training", and how are they different? [2 marks]
8. How could a "linear regression" model be used for "classification"? Explain any important data transformations in such a context. [2 marks]
9. We would like to evaluate a Machine Learning system on a development data set with 100 instances, where 80 instances are truly `N` and the rest are truly `Y`. Our system has labelled 20 of the truly `N` instances as `Y`, and 15 of the truly `Y` instances as `N`. Calculate the F-score, with respect to the `Y` label. [2 marks]
10. For what kind of "structured classification" task would one typically use a "hidden Markov model"? What assumption(s) do we make about the accompanying data when using such a model? [3 marks]
11. Use a diagram to show how "hierarchical clustering" is different to "partitional clustering". Which does an "Expectation–Maximisation" method typically produce? [2 marks]

Section B: Methodological Questions [23 marks]

In this section you are asked to demonstrate your conceptual understanding of a subset of the methods that we have studied in this subject.

Question 2: Decision Trees [9 marks]

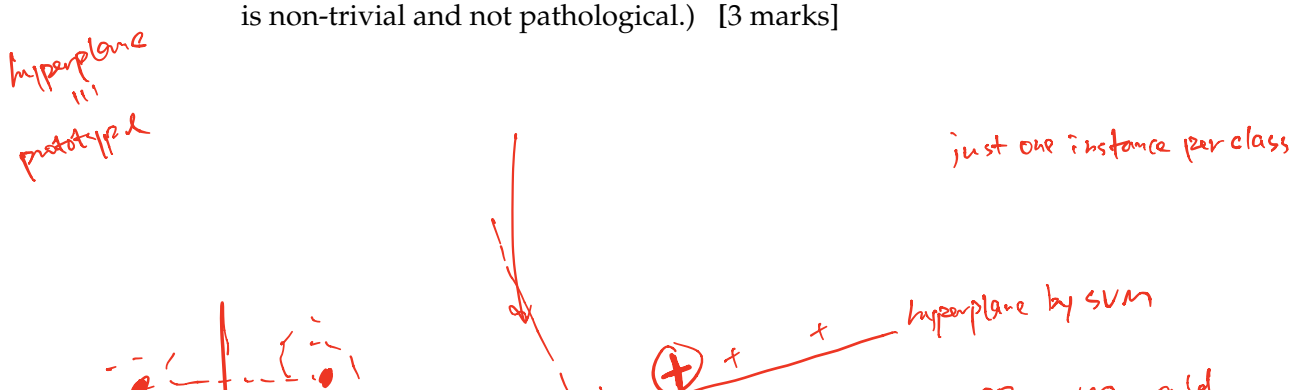
1. "Entropy" is a key concept when building a Decision Tree. Explain what it measures, and how it is used to build a tree. [3 marks]
2. Explain how we use a (trained) Decision Tree to predict the class of a test instance. [2 marks]
3. Naive Bayes and Nearest Prototype have a somewhat similar basis on which they predict the class of a test instance. Decision Trees can be quite different; explain how. [2 marks]
4. Decision Trees are the basis for a number of "ensemble methods". Choose one, and explain: [2 marks]
   (a) How is it different to a plain Decision Tree;
   (b) What improvement(s) it would typically provide, over using a plain Decision Tree.

Question 3: Gradient Descent [5 marks]

1. Given an example of where we might use the method of Gradient Descent, according to how it was described in this subject. [1 mark]
2. Explain what Gradient Descent is used for, and what problem it is designed to solve, based on your answer to Question 3(1). [2 marks]
3. The "learning rate" is a value that must be chosen; what is the risk of choosing a very small value for the learning rate? A very large value? [2 marks]

Question 4: Support Vector Machines [9 marks]

1. What is a "support vector"? How are support vectors used to predict the class of a test instance? [3 marks]
2. For some datasets, a Support Vector Machine (SVM) can make the same predictions as Nearest Prototype (NP). Explain why, and explain under what circumstances an SVM would produce a superior model to NP. [3 marks]
3. By referring to the "parameters" and/or "hyper-parameters" in a typical model, under what circumstances would an SVM be equivalent to Logistic Regression? (Assume that the data is non-trivial and not pathological.) [3 marks]



Section C: Numeric Questions [23 marks]

In this section you are asked to demonstrate your understanding of a subset of the methods that we have studied in this subject, in being able to perform numeric calculations. Questions 5 through 8 make use of the following training data set, with a single test instance labelled as `?`:

ebc	ibu	label
L	1.00	ale
L	0.10	ale
M	0.15	ale
M	0.45	stout
H	0.30	stout
H	0.45	stout
M	0.11	stout
M	0.80	?

Question 5: Nearest Neighbour [7 marks]

1. Choose a sensible distance metric based on the data above, and use it to predict the label of the test instance according to the method of 1-Nearest Neighbour. [4 marks]
2. Using the results of the previous question, use the method of 5-Nearest Neighbour to predict the test instance, employing an inverse linear distance voting scheme (or you may use a different voting scheme for partial marks). [3 marks]

Question 6: Discretisation [5 marks]

Discretisation is used to transform a "numerical" attribute into a "categorical" attribute.
According to the given training data, discretise `ibu` into three categories, using the following strategies:

1. Equal-width [2 marks]
2. k-means, with seeds 0.10, 0.30, and 0.45 [3 marks]

(You do not have to write all of the calculations for this question; a depiction of the process consistent with the correct calculations is sufficient.)

Question 7: Naive Bayes [7 marks]

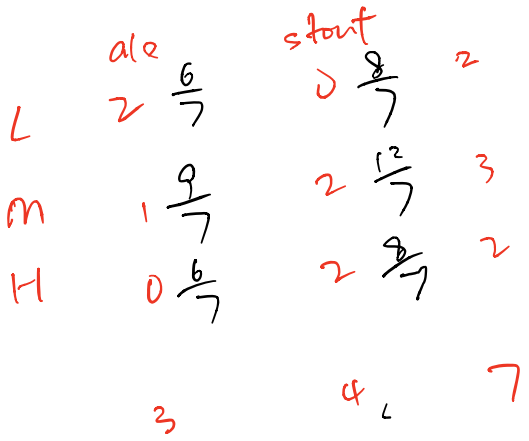
1. Using the data set before discretisation, predict the label of the test instance, using the method of Naive Bayes, consistent with this subject. (n.b. The use of "gaussians" is not recommended.) [4 marks]

2. Using a discretised representation of the data, predict the test instance, using Naive Bayes with "Laplace smoothing". [3 marks]

Question 8: Feature Selection [4 marks]

1. Given this two-class problem, and a discretised representation of the data, construct the contigency matrices for `ebc` and `ibu`. [2 marks]

2. Which one of these attributes would be preferred by a typical "feature filtering" method? (You do not have to show all of your working; a description consistent with the methods from this subject is sufficient.) [2 marks]



Section D: Design and Application Questions [13 marks]

In this section you are asked to demonstrate that you have gained a high-level understanding of the methods and algorithms covered in this subject, and can apply that understanding.

Expect to respond using about one-third of a page to one full page, for each of the three points below. These questions will require significantly more thought than those in Sections A–C and should be attempted only after having completed the earlier sections.

Question 9: Do customers like our services? [13 marks]

You have been contracted by a medium–sized business to help them to assess public sentiment regarding several of the services they provide. The business has noticed that many people mention their services in short messages, posted to a public micro-blogging platform.

You aren't a text-processing expert (or are you?), but other data scientists on the team have provided you with a representation of the messages as an "embedding" (a dense, real–valued vector). At the start of your contract, a large number of messages have been collected, each of which is related to one of the corresponding services. However, only a handful of messages have been labelled as either `positive`, `negative`, or `neutral` in nature.

Each day, you have access to a few interns, who are capable of reading a few more messages and labelling the corresponding sentiment.

Your objective is to build a system that can allow the business to reliably assess current public opinion toward all of their services. You conjecture that a Machine Learning system can be constructed, which can label a large number of new messages related to each of the services, according to their sentiment as one the three labels (`positive`, `neutral`, `negative`).

Detail the following points:

- the type of machine learner(s) that you will use, and why, making sure to state any required assumptions; [4 marks]
- how you will evaluate such a system, to demonstrate its effectiveness to–date; [5 marks]
- how you will best make use of the interns (beyond having them purchase you coffee). [4 marks]

End of Exam

Section A: Short Answer Questions [21 marks]

Answer each of the questions in this section as briefly as possible. Expect to answer each sub-question in a couple of lines.

Question 1: Short Answer Questions [21 marks]

1. Explain the difference between "supervised" and "unsupervised" learning, and give an example of a typical method for each. [2 marks]
2. What is the relationship between "instances" and "attributes" (also known as "features")? [1 mark]
3. Some Machine Learning methods rely on having "numerical" data. Give an example of how we can use such a learner with "categorical" data. [1 mark]
4. When might the "softmax" function be used in a Machine Learning context? [1 mark]
5. What might it look like, if we had a Machine Learning system with low "model bias", but high "evaluation bias"? Why would such a situation be undesirable? [3 marks]
6. Under what circumstances would a "neural network" be equivalent to a "logistic regression" model? [2 marks]
7. How is "active learning" similar to "self-training", and how are they different? [2 marks]
8. How could a "linear regression" model be used for "classification"? Explain any important data transformations in such a context. [2 marks]
9. We would like to evaluate a Machine Learning system on a development data set with 100 instances, where 80 instances are truly `N` and the rest are truly `Y`. Our system has labelled 20 of the truly `N` instances as `Y`, and 15 of the truly `Y` instances as `N`. Calculate the F-score, with respect to the `Y` label. [2 marks]
10. For what kind of "structured classification" task would one typically use a "hidden Markov model"? What assumption(s) do we make about the accompanying data when using such a model? [3 marks]
11. Use a diagram to show how "hierarchical clustering" is different to "partitional clustering". Which does an "Expectation–Maximisation" method typically produce? [2 marks]

1. Supervised: labelled training data: SVM.
   unsupervised: data not labelled or not known to learner: k-means

2. Instances are described by a combination of features.
   Together they form a matrix of data, where rows are instances and features are columns.

3. SVM relies on numeric data since it considers a vector space.
   Use one-hot-encoder to convert a categorical feature into a vector of binary values.
   e.g. data has a feature "colour" having 3 possible values {R, G, B}.
   To use SVM on this data, convert an instance with "colour = R" into "colour = (1,0,0)"; similarly, convert "colour = G" into "colour = (0,1,0)", "colour = B" into "colour = (0,0,1)".

4. In a multi-class classification problem using neural network/logistic regression.
   In soft k-means.

5. A classifer with (n-1) instances for training, 1 instance for evaluation. With sufficient training data, it is likely that class distribution is well fitted. i.e. low model bias. Only one evaluation instance means accuracy is either 0 or 1, High evaluation bias means accuracy on test set is not representative of real performance.

6. A single-layer perceptron, i.e. one neuron, with sigmoid activation function and a binary step to convert response to class.

7. Both are semi-supervised learning.
   Active learning requires an oracle to be present during training.
   Training in self-learning is automated.

8. Map linear response to range (0, 1) using a logistic function.
   Use 0.5 as the boundary between two classes.

9.

		Predicted Y	Predicted N	
Actual	Y	5	15	20
	N	20	60	80

$\text{Precision}(Y) = \frac{5}{5+20} = 0.2$

$\text{Recall}(Y) = \frac{5}{5+15} = 0.25$

$F_1(Y) = \frac{2 \times 0.2 \times 0.25}{0.2+0.25} = \frac{2}{9}$

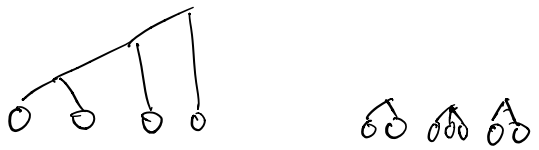
10. Sequential data. Find the most probable observation sequence.

Assumption

states only depend on the immediate preceding state.

observation only depends on the current state.

11. hierarchical partitional



EM usually produces partitional clustering.

Question 2: Decision Trees [9 marks]

1. "Entropy" is a key concept when building a Decision Tree. Explain what it measures, and how it is used to build a tree. [3 marks]
2. Explain how we use a (trained) Decision Tree to predict the class of a test instance. [2 marks]
3. Naive Bayes and Nearest Prototype have a somewhat similar basis on which they predict the class of a test instance. Decision Trees can be quite different; explain how. [2 marks]
4. Decision Trees are the basis for a number of "ensemble methods". Choose one, and explain: [2 marks]
   (a) How is it different to a plain Decision Tree;
   (b) What improvement(s) it would typically provide, over using a plain Decision Tree.

1. Entropy measures skewness of class distribution.
We use information gain to decide partition
Info gain is difference between entropy of class label
before partition and weighted average entropy after partition.

2. Follow a series of decision rules using test instance's
feature values until the deepest level of decision tree and choose final
majority class as predicted label.

3. NB and NP are similar because both consider how likely an instance
belongs to any class and compare these "likelihoods" to decide the
most likely class for an instance. NB is based on probability, NP
is based on distance in a vector space.
DT does not consider how likely an instance belongs to a class.
Instead, it summarises a set of rules that link feature patterns
to a class and simply checks which rules are satisfied by test
instance. No comparison involved.

4(a) Random Forest. Build multiple random trees, each using only
a sample of instances and features.
(b) Reduce variance by using multiple base classifiers.
Less impact from high-arity features, extreme values.

Question 3: Gradient Descent [5 marks]

1. Given an example of where we might use the method of Gradient Descent, according to how it was described in this subject. [1 mark]
2. Explain what Gradient Descent is used for, and what problem it is designed to solve, based on your answer to Question 3(1). [2 marks]
3. The "learning rate" is a value that must be chosen; what is the risk of choosing a very small value for the learning rate? A very large value? [2 marks]

1. Linear regression to fit a "best-fit-line"

2. Used for optimisation of a loss function measuring error of predictions

   Loss function needs to be well defined and differentiable.

3. Small learning rate: convergence is slow or not reached

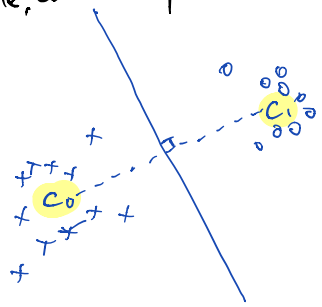
   Large learning rate: miss the optimal value; exploding gradient.

Question 4: Support Vector Machines [9 marks]

1. What is a "support vector"? How are support vectors used to predict the class of a test instance? [3 marks]
2. For some datasets, a Support Vector Machine (SVM) can make the same predictions as Nearest Prototype (NP). Explain why, and explain under what circumstances an SVM would produce a superior model to NP. [3 marks]
3. By referring to the "parameters" and/or "hyper-parameters" in a typical model, under what circumstances would an SVM be equivalent to Logistic Regression? (Assume that the data is non-trivial and not pathological.) [3 marks]

1. Vectors that are at the boundary of different classes.

Support vectors constrain the margin between two classes. They help decide the position of hyperplane, which separates two classes with maximum margin.



2. When hyperplane formed by SVM is identical, or very close to the middle line between two prototypes (shown in the sketch).

One simple scenario: each class only has one instance.

SVM has embedded feature selection => if some features are meaningful then SVM is more likely to make use of such features thus better performance.

SVM can use kernel trick to handle non-linear separable data but NP cannot.

3. Linear kernel for SVM.
Regularisation for Logistic regression. (most likely L2 regularisation).

	ebc	ibu	label	d	
0	L	1.00	ale	0.7	✓
0	L	0.10	ale	1.2	
0.5	M	0.15	ale	0.65	✓
0.5	M	0.45	stout	0.35	✓
1	H	0.30	stout	1.0	
1	H	0.45	stout	0.85	✓
0.5	M	0.11	stout	0.69	✓
0.5	M	0.80	?		

Question 5: Nearest Neighbour [7 marks]

1. Choose a sensible distance metric based on the data above, and use it to predict the label of the test instance according to the method of 1-Nearest Neighbour. [4 marks]
2. Using the results of the previous question, use the method of 5-Nearest Neighbour to predict the test instance, employing an inverse linear distance voting scheme (or you may use a different voting scheme for partial marks). [3 marks]

1. Treat (L, M, H) = (0, 0.5, 1) as numeric type.
Choose manhattan distance.
distance calculation shown on the plot.
=> nearest neighbour is (0.5, 0.45) => class prediction is stout.

2. 5 nearest neighbours:

distance	class	inversed linear distance
0.7	ale	(0.85-0.7)/(0.85-0.35) = 0.3
0.65	ale	(0.85-0.65)/(0.85-0.35) = 0.4
0.35	stout	(0.85-0.35)/(0.85-0.35) = 1
0.85	stout	(0.85-0.85)/(0.85-0.35) = 0
0.69	stout	(0.85-0.69)/(0.85-0.35) = 0.32

Voting(ale) = 0.3 + 0.4 = 0.7
Voting(stout) = 1 + 0 + 0.32 = 1.32
=> predicted class is stout.

	ebc	ibu	label
0	L	1.00	ale
0	L	0.10	ale
0.5	M	0.15	ale
0.5	M	0.45	stout
1	H	0.30	stout
1	H	0.45	stout
0.5	M	0.11	stout
0.5	M	0.80	?

Question 6: Discretisation [5 marks]

Discretisation is used to transform a "numerical" attribute into a "categorical" attribute. According to the given training data, discretise `ibu` into three categories, using the following strategies:

1. Equal-width [2 marks]
2. k-means, with seeds 0.10, 0.30, and 0.45 [3 marks]

(You do not have to write all of the calculations for this question; a depiction of the process consistent with the correct calculations is sufficient.)

1. width = $\frac{1.00 - 0.10}{3} = 0.30$.

level 1 : $[0.10, 0.40)$

level 2 : $[0.40, 0.70)$

level 3 : $[0.70, 1.00]$

2. Iteration 1:

centroid 0.10 contains $\{0.10, 0.15, 0.11\}$

centroid 0.30 contains $\{0.30\}$

centroid 0.45 contains $\{1.00, 0.45, 0.45\}$

updated centroids: $\{0.12, 0.30, 0.63\}$

Iteration 2:

centroid 0.12 contains $\{0.10, 0.15, 0.11\}$

centroid 0.30 contains $\{0.30, 0.45, 0.45\}$

centroid 0.63 contains $\{1.00\}$

updated centroids: $\{0.12, 0.40, 1.00\}$

Iteration 3:

centroid 0.12 contains {0.10, 0.15, 0.11} → level 1

centroid 0.40 contains {0.30, 0.45, 0.45} → level 2

centroid 1.00 contains {1.00} → level 3.

updated centroids: {0.12, 0.40, 1.00}

stop.

=>

sbq	equal width	k-means
1.00	3	3
0.10	1	1
0.15	1	1
0.45	2	2
0.30	1	2
0.45	2	2
0.11	1	1
0.80	3	2 or 3

	ebc	ibu	label
0	L	3 1.00	ale
0	L	1 0.10	ale
0.5	M	1 0.15	ale
0.5	M	2 0.45	stout
1	H	1 0.30	stout
1	H	2 0.45	stout
0.5	M	1 0.11	stout
0.5	M	3 0.80	?

Question 7: Naive Bayes [7 marks]

1. Using the data set before discretisation, predict the label of the test instance, using the method of Naive Bayes, consistent with this subject. (n.b. The use of "gaussians" is not recommended.) [4 marks]

2. Using a discretised representation of the data, predict the test instance, using Naive Bayes with "Laplace smoothing". [3 marks]

1. Prior : $P(ale) = \frac{3}{7}$, $P(stout) = \frac{4}{7}$.

Likelihoods: $P(ebc = L \mid ale) = \frac{2}{3}$, $P(ebc = M \mid ale) = \frac{1}{3}$, $P(ebc = H \mid ale) = 0$

$P(ebc = L \mid stout) = 0$, $P(ebc = M \mid stout) = \frac{1}{2}$, $P(ebc = H \mid stout) = \frac{1}{2}$

Assume uniform distribution of ibu.

$ibu \mid ale \overset{d}{=} Unif(0.10, 1.00)$, $ibu \mid stout \overset{d}{=} Unif(0.11, 0.45)$.

$$P(ale \mid ebc = M, ibu = 0.80) = P(ale) P(ebc = M \mid ale) P(ibu = 0.80 \mid ale)$$

$$= \frac{3}{7} \times \frac{1}{3} \times \frac{1}{1.00 - 0.10} = \frac{3}{7} \times \frac{1}{3} \times \frac{10}{9} = \frac{10}{63} = 0.16.$$

$$P(stout \mid ebc = M, ibu = 0.80) = P(stout) P(ebc = M \mid stout) P(ibu = 0.8 \mid stout)$$

$$= \frac{4}{7} \times \frac{1}{2} \times 0 = 0$$

$\Rightarrow$ ale.

2. Use equal width:

Prior: $P(\text{ale}) = \frac{3}{7}$, $P(\text{stout}) = \frac{4}{7}$.

Likelihoods: $P(\text{ebc}=L|\text{ale}) = \frac{2+1}{3+3} = \frac{3}{6}$, $P(\text{ibu}=1|\text{ale}) = \frac{2+1}{3+3} = \frac{3}{6}$

$P(\text{ebc}=M|\text{ale}) = \frac{1+1}{3+3} = \frac{2}{6}$, $P(\text{ibu}=2|\text{ale}) = \frac{0+1}{3+3} = \frac{1}{6}$

$P(\text{ebc}=H|\text{ale}) = \frac{0+1}{3+3} = \frac{1}{6}$. $P(\text{ibu}=3|\text{ale}) = \frac{1+1}{3+3} = \frac{2}{6}$

$P(\text{ebc}=L|\text{stout}) = \frac{0+1}{4+3} = \frac{1}{7}$ $P(\text{ibu}=1|\text{stout}) = \frac{2+1}{4+3} = \frac{3}{7}$

$P(\text{ebc}=M|\text{stout}) = \frac{2+1}{4+3} = \frac{3}{7}$ $P(\text{ibu}=2|\text{stout}) = \frac{2+1}{4+3} = \frac{3}{7}$

$P(\text{ebc}=H|\text{stout}) = \frac{2+1}{4+3} = \frac{3}{7}$. $P(\text{ibu}=3|\text{stout}) = \frac{0+1}{4+3} = \frac{1}{7}$.

$$P(\text{ale}|\text{ebc}=M, \text{ibu}=3) = P(\text{ale})P(\text{ebc}=M|\text{ale})P(\text{ibu}=3|\text{ale})$$
$$= \frac{3}{7} \times \frac{2}{6} \times \frac{2}{6} = \frac{1}{21} = 0.048.$$

$$P(\text{stout}|\text{ebc}=M, \text{ibu}=3) = P(\text{stout})P(\text{ebc}=M|\text{stout})P(\text{ibu}=3|\text{stout})$$
$$= \frac{4}{7} \times \frac{3}{7} \times \frac{1}{7} = \frac{12}{343} = 0.035$$

$\Rightarrow$ predicted label is ale.

	ebc	ibu	label
0	L	31.00	ale
0	L	10.10	ale
0.5	M	10.15	ale
0.5	M	20.45	stout
1	H	20.30	stout
1	H	20.45	stout
0.5	M	10.11	stout
	M	0.80	?

Question 8: Feature Selection [4 marks]

1. Given this two-class problem, and a discretised representation of the data, construct the contigency matrices for `ebc` and `ibu`. [2 marks]
2. Which one of these attributes would be preferred by a typical "feature filtering" method? (You do not have to show all of your working; a description consistent with the methods from this subject is sufficient.) [2 marks]

1.

ebc	label: ale	label: stout	total
L	2	0	2
M	1	2	3
H	0	2	2
total	3	4	7

ibu (k-mean)	label: ale	label: stout	total
1	2	1	3
2	0	3	3
3	1	0	1
total	3	4	7

2. χ^2:

Expectation matrix:

ebc	label: ale	label: stout	total
L	6/7	8/7	2
M	9/7	12/7	3
H	6/7	8/7	2
total	3	4	7

ibu	label: ale	label: stout	total
1	9/7	12/7	3
2	9/7	12/7	3
3	3/7	4/7	1
total	3	4	7

$$\chi^2(abc) = \sum_{ij} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

$$= \frac{(2-6/7)^2}{6/7} + \frac{(0-8/7)^2}{8/7} + \frac{(1-6/7)^2}{6/7}$$

$$+ \frac{(2-12/7)^2}{12/7} + \frac{(0-6/7)^2}{6/7} + \frac{(2-8/7)^2}{8/7}$$

$$= \frac{1}{7}\left(\frac{8^2}{6} + \frac{8^2}{8} + \frac{1^2}{6} + \frac{2^2}{12} + \frac{1^2}{6} + \frac{6^2}{8}\right) = \frac{143}{42} = 3.4$$

$$\chi^2(ibu) = \sum_{ij} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

$$= \frac{(2-9/7)^2}{9/7} + \frac{(1-12/7)^2}{12/7} + \frac{(0-9/7)^2}{9/7}$$

$$+ \frac{(3-12/7)^2}{12/7} + \frac{(1-3/7)^2}{3/7} + \frac{(0-4/7)^2}{4/7}$$

$$= \frac{1}{7}\left(\frac{5^2}{12} + \frac{5^2}{12} + \frac{9^2}{9} + \frac{9^2}{12} + \frac{4^2}{3} + \frac{4^2}{4}\right) = \frac{11}{28} = 4.18$$

=> ibu better.

abc

	label: ale	stout	total
L	2	0	2
M	1	2	3
H	0	2	2
total	3	4	7

k-means

	label: ale	stout	total
1	2	1	3
2	0	3	3
3	1	0	1
total	3	4	7

$$MI(a,c) = \sum_{i,j} P(i,j) \log_2 \frac{P(i,j)}{P(i)P(j)}$$

$$= \frac{2}{7}\log_2 \frac{2/7}{\frac{2}{7}\times\frac{3}{7}} + \frac{1}{7}\log_2 \frac{1/7}{\frac{3}{7}\times\frac{3}{7}} + \frac{2}{7}\log_2 \frac{2/7}{\frac{3}{7}\times\frac{4}{7}}$$

$$+ \frac{2}{7}\log_2 \frac{2/7}{\frac{2}{7}\times\frac{4}{7}}$$

$$= \frac{1}{7}\left[2\log_2\left(\frac{14}{6}\right) + \log_2\left(\frac{7}{9}\right) + 2\log_2\left(\frac{14}{12}\right) + 2\log_2\left(\frac{14}{8}\right)\right] = 0.59$$

$$MI(i,u) = \sum_{i,j} P(i,j) \log_2 \frac{P(i,j)}{P(i)P(j)}$$

$$= \frac{2}{7}\log_2 \frac{2/7}{\frac{3}{7}\times\frac{3}{7}} + \frac{1}{7}\log_2 \frac{1/7}{\frac{3}{7}\times\frac{4}{7}} + \frac{3}{7}\log_2 \frac{3/7}{\frac{3}{7}\times\frac{4}{7}}$$

$$+ \frac{1}{7}\log_2 \frac{1/7}{\frac{1}{7}\times\frac{3}{7}}$$

$$= \frac{1}{7}\left[2\log_2\left(\frac{14}{9}\right) + \log_2\left(\frac{7}{12}\right) + 3\log_2\left(\frac{21}{12}\right) + \log_2\left(\frac{7}{3}\right)\right] = 0.59$$

Section D: Design and Application Questions [13 marks]

In this section you are asked to demonstrate that you have gained a high-level understanding of the methods and algorithms covered in this subject, and can apply that understanding.

Expect to respond using about one-third of a page to one full page, for each of the three points below. These questions will require significantly more thought than those in Sections A–C and should be attempted only after having completed the earlier sections.

Question 9: Do customers like our services? [13 marks]

You have been contracted by a medium–sized business to help them to assess public sentiment regarding several of the services they provide. The business has noticed that many people mention their services in short messages, posted to a public micro-blogging platform.

You aren't a text-processing expert (or are you?), but other data scientists on the team have provided you with a representation of the messages as an "embedding" (a dense, real–valued vector). At the start of your contract, a large number of messages have been collected, each of which is related to one of the corresponding services. However, only a handful of messages have been labelled as either `positive`, `negative`, or `neutral` in nature.

Each day, you have access to a few interns, who are capable of reading a few more messages and labelling the corresponding sentiment.

Your objective is to build a system that can allow the business to reliably assess current public opinion toward all of their services. You conjecture that a Machine Learning system can be constructed, which can label a large number of new messages related to each of the services, according to their sentiment as one the three labels (`positive`, `neutral`, `negative`).

Detail the following points:

- the type of machine learner(s) that you will use, and why, making sure to state any required assumptions; [4 marks]
- how you will evaluate such a system, to demonstrate its effectiveness to–date; [5 marks]
- how you will best make use of the interns (beyond having them purchase you coffee). [4 marks]

End of Exam

i) Semi-supervised learning.
Maybe active learning.
We have small number of labelled data, a large number of unlabelled data. and more data coming ⇒ suitable for semi-supervised learning.
Interns can help in the training as oracles under the assumption that they are reliable in labelling data correctly.
Learners can be neural network which is suitable for representation learning

ii) keep some labelled instances unknown to learner.
For instance, ask interns to label instances for the current week.
Use these as test set.
Could use a confusion matrix to represent results.
Some class maybe more useful, e.g. negative reviews may be the most important, so we can focus on precision of negative class and aim to improve it.
Error analysis can be done.
Check for overfitting/underfitting through learning curve
.....

iii) They can act as oracles for active learning, to decide those instances posted as queries by the learner.
They can select manually instances from each service and see the sentiments toward any specified service.
They can fine-tune the learner
.....