Student Number | |

The University of Melbourne
School of Computing and Information Systems

Final Examination, Semester 1, 2017
COMP30027 Machine Learning

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

This paper has 7 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 120 marks, making up 60% of the total assessment for the subject.

- Please answer all questions in the script book provided, starting each 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

Section A: Short Answer Questions [28 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 [28 marks]

1. What is the primary difference between "supervised" and "unsupervised" learning? [1.5 marks]
2. For each of the following features, indicate which of "numeric", "ordinal", and "categorical" best captures its type: [4 marks]
   (a) blood pressure level, with possible values {**low**, **medium**, **high**}
   (b) age, with possible values **[0,120]**
   (c) weather, with possible values {**clear**, **rain**, **snow**}
   (d) abalone sex, with possible values {**male**, **female**, **infant**}
3. Describe a strategy for measuring the distance between two data points comprising of "categorical" features. [1.5 marks]
4. What is the relationship between "accuracy" and "error rate" in evaluation? [1.5 marks]
5. With the aid of a diagram, describe what is meant by "maximal marginal" in the context of training a "support vector machine". [1.5 marks]
6. What makes a feature "good", i.e. worth keeping in a feature representation? How might we measure that "goodness"? [3 marks]
7. For each of the following models, state whether it is canonically applied in a "classification", "regression" or "clustering" setting: [6 marks]
   (a) multi-layer perceptron with a softmax final layer
   (b) soft k-means
   (c) multi-response linear regression
   (d) logistic regression
   (e) model tree
   (f) support vector regression
8. With the aid of an example, briefly describe what a "hyperparameter" is. [3 marks]
9. With the use of an example, outline what "stacking" is. [1.5 marks]
10. What is the convergence criterion for the "EM algorithm"? [1.5 marks]
11. Outline the basis of "purity" as a form of cluster evaluation. [1.5 marks]
12. What is the underlying assumption behind active learning based on "query-by-committee"? [1.5 marks]

Section B: Methodological Questions [30 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: Random Forests [7 marks]

"Random forests" are based on decision trees under different dimensions of "randomisation". With reference to the following toy training dataset, provide a brief outline of two (2) such "random processes" used in training a random forest. (You should give examples as necessary; it is not necessary to draw the resulting trees, although you may do so if you wish.)

Instance ID	*Feature 1*	*Feature 2*	*Feature 3*	*Class*
A:	1	1	1	`True`
B:	0	2	0	`True`
C:	3	0	0	`False`
D:	1	1	2	`False`
E:	2	0	2	`False`

Question 3: HMMs [9 marks]

This question is on "hidden Markov models".

1. In the "forward algorithm", $\alpha_t(j)$ is used to "memoise" a particular value for each state j and observation t. Describe what each $\alpha_t(j)$ represents. [3 marks]
2. In the "Viterbi algorithm", two memoisation variables are used describe for each combination of state j and observation t: (1) $\beta_t(j)$ (which plays a similar role to $\alpha_t(j)$ in the forward algorithm); and (2) $\psi_t(j)$. Describe what each $\psi_t(j)$ represents. [3 marks]
3. Why do we tend to use "log probabilities" in the Viterbi algorithm but not the forward algorithm? [3 marks]

Question 4: Model learning [14 marks]

We have discussed that the objective in much of supervised machine learning is to derive a model that explains ("fits") a set of (labelled) data. In a basic "curve fitting" scenario, we will assume that all of the needed data is available to us at the time of building the model; the objective is to fit a model to that data. In machine learning, in contrast, we have only a subset of possible data available to us when we build the model. This contrast has certain implications for how we approach machine learning, which we explore in this question.

Discuss the implications of knowing that we do not have all possible data for a given problem, in terms of the following aspects:

1. Is our primary objective in machine learning to derive a model that fits the subset of the data that we do have? Why or why not? [3 marks]
2. Explain how we can use our limited data, in a machine learning context, to demonstrate whether or not our objective has been met. [3 marks]
3. Identify and explain one important problem that can emerge with respect to this primary objective, even if we are successful in deriving a good model for the data that we have. Name one specific technique discussed in class that can be applied to mitigate this problem, and explain how it does so. [3 marks]
4. Define "bias" and "variance", indicating how we might detect each one. Discuss how bias and variance relate to each other in the context of our primary objective. [5 marks]

Section C: Numeric Questions [42 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 and 6 both make use of the following training data set:

early	tie	label
N	Y	dinner
Y	N	tea
Y	N	dinner
N	N	dinner
Y	N	dinner
N	N	tea

Table 1: Training Dataset for Questions 5 and 6

Question 5: Naive Bayes [9 marks]

Given the training dataset from Table 1:

1. Using the method of "maximum likelihood estimation" (without smoothing), compute $P(\texttt{dinner})$ and $P(\texttt{dinner}|\texttt{tie} = \texttt{Y})$. [1.5 marks]
2. Apply the method of "Naive Bayes" as it was discussed in the lectures, to predict the label of the test instance {early=N, tie=N}; show your workings. [7.5 marks]

Question 6: Decision Trees [15 marks]

1. Briefly explain — in at most two sentences — the basic logic behind the "ID3" algorithmic approach toward building decision trees. This should be focussed on labelling the nodes and leaves; you do not have to explain edge-cases. [3 marks]
2. The criterion for labelling a node, as explained in the lectures, was based around the idea of "entropy" — what does entropy tell us about a node of a decision tree? [3 marks]
3. A very similar alternative to entropy is the so-called GINI coefficient, defined as follows:

$$GINI = 1 \sum_{j \in C} [p(j)]^2$$

   Calculate the GINI coefficient based on the labels in Table 1. [3 marks]
4. Extend the notion of "Information Gain" to "GINI Gain", and demonstrate why tie would be chosen as the root of the decision tree on the given data. [3 marks]
5. Why will the resulting decision tree have difficulty classifying test instances like {early=N, tie=N}? [3 marks]

Question 7: Nearest Prototype and k-Nearest Neighbour [10.5 marks]

Given the following training dataset:

abv	opacity	label
4.8	0.20	ale
5.2	0.10	ale
5.0	0.33	ale
4.7	0.02	lager
5.1	0.23	lager
4.6	0.05	lager

1. Generate the "prototype" for each class in the training data. [3 marks]
2. For the test instance {abv=5.1, opacity= 0.23}, determine which of the prototypes is more similar. (You will need to choose an appropriate metric, and show your work; do not just use inspection.) [3 marks]
3. What should happen if we instead use the "1-Nearest Neighbour" method to predict the label of that test instance? (You do not need to show your work.) [1.5 marks]
4. Give one possible reason why each of these two methods could reasonably claim to be making a better prediction for this instance. [3 marks]

Question 8: Evaluation [7.5 marks]

Assume that our development set contains 100 instances (truly) labelled as one of three classes as follows: 50 acro instances, 30 base instances, and 20 claw instances. We then build a classifier, apply it to this dataset, and observe the following confusion matrix:

		Actual		
		acro	base	claw
Predicted	acro	28	5	7
	base	10	10	0
	claw	0	8	12

1. Determine the "classification accuracy" of the system described above. [1.5 marks]
2. Calculate the "micro-averaged precision" and "macro-averaged precision" of this system. (Show your workings; you should simplify these to a single fraction or decimal value.) [3 marks]
3. In some contexts these three values are equal; here they are not. Briefly explain why. [3 marks]

Section D: Design and Application Questions [20 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: [20 marks]

You are tasked with building a system which labels images as to whether they contain a given product type (e.g. `car` or `mobile phone`), based on a large set of labelled training instances. A given image may contain multiple or no product types, with the expectation that most (but not all) test instances will contain at least one product.

Each image has been transformed into an "embedding" (i.e. a dense real-valued vector representation of the image). This transformation process should be treated as a "black-box", which will be applied consistently to every image in the collection.

At the start of the project, you are provided with training instances for each of 500 product types, but as the project progresses, the set of product types is to be expanded in increments of around 100 new types, and new training instances are to be provided for each of the newly-added product types, up to a final total of around 1000 product types.

You will additionally be provided with extra training instances for a subset of the pre-existing product types, e.g. to capture newly-released models of cars or mobile phones. However, it is desirable to have a model that does not reverse positive predictions for a pre-existing category; rather, you are instructed to employ this extra training data to find further instances of the corresponding product type.

Finally, for each product type, you are provided with a "priority level" (high, medium or low) for how critical it is that your model has good coverage in correctly identifying instances of that product type.

Outline the following:

- the type of machine learning algorithm that you will use, and why; state any assumptions you are making in your answer.
- how you will deal with updates to the label set and also extra training instances for pre-existing labels, making specific mention of how you will maintain consistency in your model predictions, but still have your model find new instances of a given product type.
- how you will evaluate your model.

End of Exam

Section A: Short Answer Questions [28 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 [28 marks]

1. What is the primary difference between "supervised" and "unsupervised" learning? [1.5 marks]
   ↳ not labelled or not known
2. For each of the following features, indicate which of "numeric", "ordinal", and "categorical" best captures its type: [4 marks]
   (a) blood pressure level, with possible values {**low**, **medium**, **high**} ordinal
   (b) age, with possible values **[0,120]** numeric
   (c) weather, with possible values {**clear**, **rain**, **snow**} categorical
   (d) abalone sex, with possible values {**male**, **female**, **infant**} categorical
3. Describe a strategy for measuring the distance between two data points comprising of "categorical" features. [1.5 marks]
4. What is the relationship between "accuracy" and "error rate" in evaluation? [1.5 marks]
5. With the aid of a diagram, describe what is meant by "maximal marginal" in the context of training a "support vector machine". [1.5 marks]
6. What makes a feature "good", i.e. worth keeping in a feature representation? How might we measure that "goodness"? [3 marks]
7. For each of the following models, state whether it is canonically applied in a "classification", "regression" or "clustering" setting: [6 marks]
   (a) multi-layer perceptron with a softmax final layer
   (b) soft k-means
   (c) multi-response linear regression
   (d) logistic regression
   (e) model tree
   (f) support vector regression
8. With the aid of an example, briefly describe what a "hyperparameter" is. [3 marks]
9. With the use of an example, outline what "stacking" is. [1.5 marks]
10. What is the convergence criterion for the "EM algorithm"? [1.5 marks]
11. Outline the basis of "purity" as a form of cluster evaluation. [1.5 marks]
12. What is the underlying assumption behind active learning based on "query-by-committee"? [1.5 marks]

1. supervised: training instances are labelled and known to the learner.
   unsupervised: not labelled or not known to the learner.

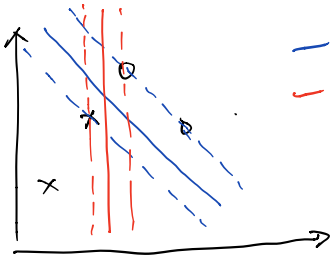
2. 1) ordinal 2) numeric 3) categorical 4) categorical

3. Hamming distance: e.g. $d('fgnb', 'fgsb') = 1$ (equal length)
   $d('ceb', 'ceeeb') = 3$

   Jaccard similarity: $sim(A,B) = \frac{|A \cap B|}{|A \cup B|}$

   Dice similarity: $sim(A,B) = \frac{2|A \cap B|}{|A|+|B|}$

4. Accuracy = 1 - error rate.

5. 

   — hyperplane with maximal margin
   — hyperplane without maximal margin.

   The width of the margin, i.e. distance from hyperplane to support vectors for both classes is maximized.

6. Good feature is well correlated with class label or response.
   Mutual info, χ^2, R^2 are ways to measure correlation.

7. a) Multi-class classification.
   b) Clustering.
   c) Classification.
   d) Classification.
   e) Regression.
   f) Regression.

8. A special parameter determined prior to training, used to control the training process.
   e.g. number of neighbours k, in kNN.

9. Stacking refers to training a meta classifier, using predictions of several base classifiers as input, e.g.

10. Convergence of maximum log-likelihood to a very small range.

11. Proportion of majority class in the cluster. Overall purity is weighted average of purity within each cluster.

12. Disputes among classifiers imply the instance is difficult, thus useful for training.

Question 2: Random Forests [7 marks]

"Random forests" are based on decision trees under different dimensions of "randomisation". With reference to the following toy training dataset, provide a brief outline of two (2) such "random processes" used in training a random forest. (You should give examples as necessary; it is not necessary to draw the resulting trees, although you may do so if you wish.)

Instance ID	*Feature 1*	*Feature 2*	*Feature 3*	*Class*
A:	1	1	1	`True`
B:	0	2	0	`True`
C:	3	0	0	`False`
D:	1	1	2	`False`
E:	2	0	2	`False`

1. Each random tree uses a random sample of the total training instances, which could also be used along with bagging.
   e.g. Suppose each random tree uses 4 instances to train.
   tree 1 may use {A, B, C, D}, tree 2 may use {A, C, C, E},
   tree 3 may use {B, C, E, E}.

2. Each random tree uses a random subset of all features
   e.g. tree 1 may only consider {Feature 1, Feature 2}
   tree 2 may only consider {Feature 3, Feature 2}
   tree 3 may only consider {Feature 1, Feature 2}.

Question 3: HMMs [9 marks]

This question is on "hidden Markov models".

1. In the "forward algorithm", $\alpha_t(j)$ is used to "memoise" a particular value for each state j and observation t. Describe what each $\alpha_t(j)$ represents. [3 marks]
2. In the "Viterbi algorithm", two memoisation variables are used describe for each combination of state j and observation t: (1) $\beta_t(j)$ (which plays a similar role to $\alpha_t(j)$ in the forward algorithm); and (2) $\psi_t(j)$. Describe what each $\psi_t(j)$ represents. [3 marks]
3. Why do we tend to use "log probabilities" in the Viterbi algorithm but not the forward algorithm? [3 marks]

1. Probability of observing all observations up to and including t ending up at state j.

2. Given observations up to and including t at state j, the most probable next state for j.

3. Product of probabilities in the Viterbi algorithm can be converted to sum of log probabilities.
Forward algorithm calculates sum of product of probabilities, which are not suitable for log transformation.

Question 4: Model learning [14 marks]

We have discussed that the objective in much of supervised machine learning is to derive a model that explains ("fits") a set of (labelled) data. In a basic "curve fitting" scenario, we will assume that all of the needed data is available to us at the time of building the model; the objective is to fit a model to that data. In machine learning, in contrast, we have only a subset of possible data available to us when we build the model. This contrast has certain implications for how we approach machine learning, which we explore in this question.

Discuss the implications of knowing that we do not have all possible data for a given problem, in terms of the following aspects:

1. Is our primary objective in machine learning to derive a model that fits the subset of the data that we do have? Why or why not? [3 marks]
2. Explain how we can use our limited data, in a machine learning context, to demonstrate whether or not our objective has been met. [3 marks]
3. Identify and explain one important problem that can emerge with respect to this primary objective, even if we are successful in deriving a good model for the data that we have. Name one specific technique discussed in class that can be applied to mitigate this problem, and explain how it does so. [3 marks]
4. Define "bias" and "variance", indicating how we might detect each one. Discuss how bias and variance relate to each other in the context of our primary objective. [5 marks]

1. No. Primary objective should be to train a model that can generalise to unseen data, or any data that is within our domain.

2. Use hold-out or cross validation to split data into training set and testing set. Use traing set to fit the model and use testing set, which is unknown to the model, to evaluate performance. It gives a measurement of model's generalisation power.

3. Overfitting.
Add regularisation. In the setting of "curve fitting" scenario, we can use L1 LASSO regression or L2 ridge regression. Regularization controls model complexity, aiming to obtain a good model, but as simple as possible. In the setting of "curve fitting", this is done by minimising both loss function as well as size of parameters.

4. Bias: how well the model fits the training data. Detected by calculating error. A high bias implies constantly poor performance
Variance: how well the model generalise unseen data. Detected by testing model trained by different training data. A high variance

implies model behaves obviously different among different training data, i.e. sensitive to training data.

Bias and variance usually have a trade-off.

Higher model complexity increases variance since it captures minor details of training data, but decreases bias as it fits training data better. In the "curve fitting" scenario, it may be a very complicated curve passing exactly through each instance.

Lower model complexity decreases variance but increases bias.

A good model tries to lower both or find a balance between bias and variance to achieve good fit as well as good generalization.

early	tie	label
N	Y	dinner
Y	N	tea
Y	N	dinner
N	N	dinner
Y	N	dinner
N	N	tea

Question 5: Naive Bayes [9 marks]

Given the training dataset from Table 1:

1. Using the method of "maximum likelihood estimation" (without smoothing), compute $P(\texttt{dinner})$ and $P(\texttt{dinner}|\texttt{tie} = \texttt{Y})$. [1.5 marks]
2. Apply the method of "Naive Bayes" as it was discussed in the lectures, to predict the label of the test instance {early=N, tie=N}; show your workings. [7.5 marks]

1. $P(\text{dinner}) = \frac{\text{frequency of "dinner"}}{\text{number of instances}} = \frac{4}{6} = \frac{2}{3}$.

$P(\text{dinner}|\text{tie=Y}) = \frac{\text{frequency of ("dinner" and "tie=Y")}}{\text{number of instances where "tie=Y"}}$

$= \frac{1}{1} = 1$

2. Prior: $P(\text{dinner}) = \frac{2}{3}$, $P(\text{tea}) = \frac{1}{3}$.

Likelihoods: $P(\text{early=Y}|\text{dinner}) = \frac{2}{4} = \frac{1}{2}$, $P(\text{early=N}|\text{dinner}) = \frac{2}{4} = \frac{1}{2}$.

$P(\text{tie=Y}|\text{dinner}) = \frac{1}{4}$, $P(\text{tie=N}|\text{dinner}) = \frac{3}{4}$

$P(\text{early=Y}|\text{tea}) = \frac{1}{2}$, $P(\text{early=N}|\text{tea}) = \frac{1}{2}$

$P(\text{tie=Y}|\text{tea}) = \frac{0}{2} = 0$ $P(\text{tie=N}|\text{tea}) = \frac{2}{2} = 1$

$P(\text{dinner}|\text{early=N, tie=N}) = P(\text{early=N}|\text{dinner})\,P(\text{tie=N}|\text{dinner})\,P(\text{dinner})$

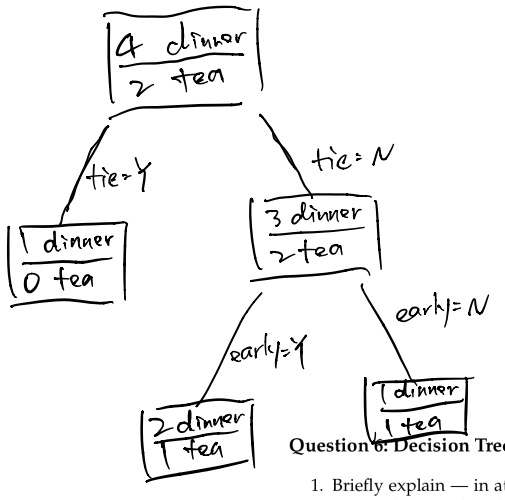
$= \frac{1}{2} \times \frac{3}{4} \times \frac{2}{3} = \frac{1}{4}$

$P(\text{tea}|\text{early=N, tie=N}) = P(\text{early=N}|\text{tea})\,P(\text{tie=N}|\text{tea})\,P(\text{tea})$

$= \frac{1}{2} \times 1 \times \frac{1}{3} = \frac{1}{6}$

$\Rightarrow$ prediction is dinner.

early	tie	label
N	Y	dinner
Y	N	tea
Y	N	dinner
N	N	dinner
Y	N	dinner
N	N	tea



Question 6. Decision Trees [15 marks]

1. Briefly explain — in at most two sentences — the basic logic behind the "ID3" algorithmic approach toward building decision trees. This should be focussed on labelling the nodes and leaves; you do not have to explain edge-cases. [3 marks]
2. The criterion for labelling a node, as explained in the lectures, was based around the idea of "entropy" — what does entropy tell us about a node of a decision tree? [3 marks]
3. A very similar alternative to entropy is the so-called GINI coefficient, defined as follows:

$$GINI = 1 - \sum_{j \in C} [p(j)]^2$$

   Calculate the GINI coefficient based on the labels in Table 1. [3 marks]
4. Extend the notion of "Information Gain" to "GINI Gain", and demonstrate why `tie` would be chosen as the root of the decision tree on the given data. [3 marks]
5. Why will the resulting decision tree have difficulty classifying test instances like {`early=N`, `tie=N`}? [3 marks]

1. At each node, partition training data using the feature with highest information gain. Repeat partition recursively until every leaf node is pure or features are used up.

2. Entropy is calculated with respect to labels. It indicates unpredictedness of the label and skewness of class distribution at a node. High entropy has less skewed class distribution, thus more unpredictable.

3. $P(\text{dinner}) = \frac{2}{3}$, $P(\text{tea}) = \frac{1}{3}$.

$$GINI = 1 - \sum_{j \in C} [p(j)]^2 = 1 - \left([P(\text{dinner})]^2 + [P(\text{tea})]^2 \right)$$

$$= 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2 \right]$$

$$= 1 - \frac{5}{9}$$

$$= \frac{4}{9}.$$

4. After partitioning by early:

$$GINI(early=N) = 1 - \sum_{j \in C} [p(j)]^2$$
$$= 1 - \left([P(dinner|early=N)]^2 + [P(tea|early=N)]^2 \right)$$
$$= 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2 \right]^2$$
$$= \frac{4}{9}.$$

$$GINI(early=Y) = 1 - \sum_{j \in C} [p(j)]^2$$
$$= 1 - \left([P(dinner|early=Y)]^2 + [P(tea|early=Y)]^2 \right)$$
$$= 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2 \right]$$
$$= \frac{4}{9}.$$

$$\text{Mean GINI} = P(early=N)\,GINI(early=N) + P(early=Y)\,GINI(early=Y)$$
$$= \frac{3}{6} \times \frac{4}{9} + \frac{3}{6} \times \frac{4}{9} = \frac{4}{9}.$$

GINI gain = GINI (before partition) − Mean GINI (after partition)

$$= \frac{4}{9} - \frac{4}{9} = 0.$$

Similarly, $GINI(tie=Y) = 1 - [1^2 + 0^2] = 0$

$$GINI(tie=N) = 1 - \left[\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2 \right] = \frac{12}{25}.$$

$$\text{Mean GINI} = \frac{1}{6} \times 0 + \frac{5}{6} \times \frac{12}{25} = \frac{2}{5}.$$

$$\text{GINI gain} = \frac{4}{9} - \frac{2}{5} = \frac{2}{45} > 0$$

$\Rightarrow$ tie is a better feature to partition data.

5. We have a draw in the training data for an instance like $\{early=N, tie=N\}$

Question 7: Nearest Prototype and k-Nearest Neighbour [10.5 marks]

Given the following training dataset:

abv	opacity	label
4.8	0.20	ale
5.2	0.10	ale
5.0	0.33	ale
4.7	0.02	lager
5.1	0.23	lager
4.6	0.05	lager

1. Generate the "prototype" for each class in the training data. [3 marks]
2. For the test instance {abv=5.1, opacity= 0.23}, determine which of the prototypes is more similar. (You will need to choose an appropriate metric, and show your work; do not just use inspection.) [3 marks]
3. What should happen if we instead use the "1-Nearest Neighbour" method to predict the label of that test instance? (You do not need to show your work.) [1.5 marks]
4. Give one possible reason why each of these two methods could reasonably claim to be making a better prediction for this instance. [3 marks]

1. Prototype(ale) = $\left(\text{abv} = \frac{4.8+5.2+5.0}{3},\ \text{opacity} = \frac{0.20+0.10+0.33}{3}\right)$

$= (\text{abv} = 5.0,\ \text{opacity} = 0.21)$

Prototype(lager) = $\left(\text{abv} = \frac{4.7+5.1+4.6}{3},\ \text{opacity} = \frac{0.02+0.23+0.05}{3}\right)$

$= (\text{abv} = 4.8,\ \text{opacity} = 0.10)$

2. Choose Euclidean distance as a metric.

$d(\text{prototype(ale)}, \text{test instance}) = \sqrt{(5.0-5.1)^2+(0.21-0.23)^2}$

$= \sqrt{0.1^2+0.02^2}$

$d(\text{prototype(lager)}, \text{test instance}) = \sqrt{(4.8-5.1)^2+(0.10-0.23)^2}$

$= \sqrt{0.3^2+0.13^2} > \sqrt{0.1^2+0.02^2}$

Prototype of ale is more similar.

3. Test instance would be classified as lager since a training instance labelled lager is exactly the same as test instance, thus is the nearest neighbour.

4. Ale: Test instance is closer to a typical ale. In fact, the exact instance labelled as lager looks like an outlier with obviously higher opacity.

Lager: We have an exactly the same instance present in training data, already labeled as lager.

Question 8: Evaluation [7.5 marks]

Assume that our development set contains 100 instances (truly) labelled as one of three classes as follows: 50 `acro` instances, 30 `base` instances, and 20 `claw` instances. We then build a classifier, apply it to this dataset, and observe the following confusion matrix:

		Actual			
		`acro`	`base`	`claw`	
Predicted	`acro`	28	5	7	40
	`base`	10	10	0	20
	`claw`	0	8	12	20
		38	23	19	80

1. Determine the "classification accuracy" of the system described above. [1.5 marks]
2. Calculate the "micro-averaged precision" and "macro-averaged precision" of this system. (Show your workings; you should simplify these to a single fraction or decimal value.) [3 marks]
3. In some contexts these three values are equal; here they are not. Briefly explain why. [3 marks]

1. Accuracy: $\frac{28+10+12}{100} = \frac{50}{100} = 0.5$.

2. Micro precision:

$$Precision_\mu = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} TP_i + FP_i} = \frac{28+10+12}{(28+5+7)+(10+10+0)+(0+8+12)} = \frac{50}{80} = 0.625.$$

Macro precision:

$$Precision_M = \frac{\sum_{i=1}^{C} Precision(i)}{C} = \frac{\frac{28}{28+5+7} + \frac{10}{10+10+0} + \frac{12}{0+8+12}}{3}$$

$$= \frac{\frac{28}{40} + \frac{10}{20} + \frac{12}{20}}{3} = \frac{28+20+24}{3\times 40} = \frac{72}{120} = \frac{3}{5} = 0.6.$$

3. Macro treats every class equally by taking average of each class's precision.
If class distribution is even, then two averaged precision should be the same.
In fact, a weighted average of individual precision (instead of treating them equally) is another way to calculate micro average.
Also, here 20 instances are not classified, which also causes the difference.

Section D: Design and Application Questions [20 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: [20 marks]

You are tasked with building a system which labels images as to whether they contain a given product type (e.g. `car` or `mobile phone`), based on a large set of labelled training instances. A given image may contain multiple or no product types, with the expectation that most (but not all) test instances will contain at least one product.

Each image has been transformed into an "embedding" (i.e. a dense real-valued vector representation of the image). This transformation process should be treated as a "black-box", which will be applied consistently to every image in the collection.

At the start of the project, you are provided with training instances for each of 500 product types, but as the project progresses, the set of product types is to be expanded in increments of around 100 new types, and new training instances are to be provided for each of the newly-added product types, up to a final total of around 1000 product types.

You will additionally be provided with extra training instances for a subset of the pre-existing product types, e.g. to capture newly-released models of cars or mobile phones. However, it is desirable to have a model that does not reverse positive predictions for a pre-existing category; rather, you are instructed to employ this extra training data to find further instances of the corresponding product type.

Finally, for each product type, you are provided with a "priority level" (high, medium or low) for how critical it is that your model has good coverage in correctly identifying instances of that product type.

Outline the following:

i) • the type of machine learning algorithm that you will use, and why; state any assumptions you are making in your answer.

ii) • how you will deal with updates to the label set and also extra training instances for pre-existing labels, making specific mention of how you will maintain consistency in your model predictions, but still have your model find new instances of a given product type.

iii) • how you will evaluate your model.

End of Exam

i) Deep learning, e.g. Convolutional Neural Network (CNN), supervised.
Suitable for representation learning, such as image embedding in this case;
to deal with complex data. Allowing multiple-responses for each instance.
Assumption: sufficient training data;

OR k-NN. Assumption: embedding represents some similarity between products through their positions on vector space.

OR SVM. Assumption: embedding creats linear separability between different products.

ii) New class labels:
Build a new classifier for each new product type.

New instances:
With new instances coming, train a new classifier.
but only apply it when the previous classifier predicts a negative response.

iii) Assign weights to classes based on given priority.
Overall accuracy is now weighted among labels, awarding more for correctly labelling instances with high priority classes.
Since revewing positive predictions is not desirable, recall seems more important since it's likely that we do not want to miss any chance at identifying product type in images. False positive is not so penalised, so use F-score that prioritise recall for evaluation metric.

Use hold-out / CV to evaluate performance.

Calculate average of incremental accuracy.