Name of Candidate: .....

Student id: .....

Signature: .....

THE UNIVERSITY OF NEW SOUTH WALES

COMP9417 Machine Learning and Data Mining –
SAMPLE Examination Questions
VERSION WITH ANSWERS

Here are some questions from past papers which are ***SOMEWHAT*** representative of the type of material to be covered in the final exam.

Note that since the questions in this sample are taken from different past papers they are ***NOT*** of equal value. However, the total marks for a particular question ***APPROXIMATELY*** corresponds to the number of minutes it should take to answer the question.

Candidates may bring authorised calculators to the examination, but no other materials will be permitted.

Question 1 [15 marks]

Regression

A) [3 marks] Variance is a useful measure of the scatter or *spread* of values of some random variable X around its mean $E(X)$. Variance can be remembered as the "mean of the squares minus the square of the mean", but which of the following is the correct definition of variance ?
(1) $E(X^2 - E(X))$
(2) $E((X - E(X))^2)$
(3) $E(X^2 - E(X))^2$
(4) $E(E(X^2) - E(X))$
(5) $E(E(X^2) - E(X))^2$

B) [3 marks] The sum of the residuals (i.e., the differences between the actual and predicted values of the linear regression function) for the least-squares solution is:
(1) negative
(2) zero
(3) positive
(4) non-negative
(5) non-positive

C) [3 marks] Covariance of two random variables x, y is determined in relation to their differences from their respective means $\bar{x}$, $\bar{y}$. Covariance is observed when, for all instances x_i, y_i of the random variables:
(1) $x_i < \bar{x}, y_i > \bar{y}$ or $x_i < \bar{x}, y_i < \bar{y}$
(2) $x_i < \bar{x}, y_i < \bar{y}$ or $x_i < \bar{x}, y_i > \bar{y}$
(3) $x_i > \bar{x}, y_i > \bar{y}$ or $x_i > \bar{x}, y_i < \bar{y}$
(4) $x_i < \bar{x}, y_i < \bar{y}$ or $x_i > \bar{x}, y_i > \bar{y}$
(5) $x_i > \bar{x}, y_i < \bar{y}$ or $x_i > \bar{x}, y_i > \bar{y}$

D) [3 marks] Which of the following statements about the correlation of two random variables x, y is true?
(1) positive correlation between x and y means x causes y
(2) zero correlation between x and y means that x does not cause y
(3) negative correlation between x and y means x has no relationship wth y
(4) non-zero correlation between x and y means x and y have some relationship
(5) correlation of r between x and y means $y = r \times x$

E) [3 marks] Which of the following do you consider to be correct statements ?

(1) linear regression can fit non-linear dependencies of y on $\mathbf{x}$ if the parameters $\mathbf{w}$ are non-linear
(2) linear regression cannot fit non-linear dependencies of y on $\mathbf{x}$
(3) linear regression can fit any dependency of y on $\mathbf{x}$ using logarithmic transformations of $\mathbf{x}$
(4) linear regression can fit any dependency of y on $\mathbf{x}$ using polynomial transformations of $\mathbf{x}$
(5) linear regression can fit linear dependencies of y on non-linear transformations of $\mathbf{x}$

Question 1 ANSWER

A) [3 marks]
ANSWER: (2) $E((X - E(X))^2)$

B) [3 marks]
ANSWER: (2) zero (see lecture " Regression", slide 67)

C) [3 marks]
ANSWER: (4) $x_i < \bar{x}, y_i < \bar{y}$ or $x_i > \bar{x}, y_i > \bar{y}$

D) [3 marks]
ANSWER: (4) non-zero correlation between x and y means x and y have some relationship

E) [3 marks]
ANSWER: (5) linear regression can fit linear dependencies of y on non-linear transformations of $\mathbf{x}$

Question 2 [30 marks]

Comparing Representations for Learning

Consider the following truth table which gives an "m–of–n function" for three Boolean variables, where "1" denotes true and "0" denotes false. In this case the target function is: "exactly two out of three variables are true".

X	Y	Z	Class
0	0	0	false
0	0	1	false
0	1	0	false
0	1	1	true
1	0	0	false
1	0	1	true
1	1	0	true
1	1	1	false

A) **[10 marks]**
Consider the task of learning a decision tree to predict "Class". To select an attribute to split the data at the root of the tree, you must calculate the information gain for each of "X", "Y" and "Z", based on the the examples in the table. (i) For each attribute, state the information gain, and (ii) state which attribute, in your opinion, should be selected and why.

B) **[6 marks]**
Once the attribute selection is made, try to complete the tree by hand so that it is complete and correct for the examples in the table. [Hint: use an *if–then–else* representation.]

C) **[10 marks]**
Suppose we define a simple measure of distance between two equal length strings of Boolean values, as follows. The distance between two such strings B_1 and B_2 is:

$$\text{distance}(B_1, B_2) = |(\sum B_1) - (\sum B_2)|$$

where $\sum B_i$ is simply the number of variables with value 1 in string B_i. For example:

$$\text{distance}(\langle 0,0,0\rangle, \langle 1,1,1\rangle) = |0-3| = 3$$

and

$$\text{distance}(\langle 1,0,0\rangle, \langle 0,1,0\rangle) = |1-1| = 0$$

What is the LOOCV ("Leave–one–out cross-validation") error of 2–Nearest Neighbour using our distance function on the examples in the table ? [Show your working.]

D) **[4 mark]**
Compare these models. Which do you conclude provides a better representation for this particular problem ? Briefly summarise your reasoning.

Question 2 ANSWER

A) The entropy of the sample is

$$\begin{aligned}\text{SampleEntropy} &= -\frac{5}{8}\log_2\frac{5}{8} - \frac{3}{8}\log_2\frac{3}{8} \\ &= 0.954\end{aligned}$$

(i) For each attribute, "X", "Y" and "Z", the split entropy is:

$$\begin{aligned}\text{SplitEntropy} =& \frac{1}{2}(-\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}) \\ &+\frac{1}{2}(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}) \\ =& ((0.5 * 0.81) + (0.5 * 1.0)) \\ =& 0.905\end{aligned}$$

So InformationGain = SampleEntropy − SplitEntropy = 0.954 − 0.905 = 0.049, i.e., every attribute has *identical* information gain.

(ii) This means that some arbitrary, e.g., random, choice of attribute for the split has to be made, since none has higher information gain than any other.

B) Here is one complete and correct tree (there can be others):

```
If X=0 Then
    If Y = 0 Then class=false
    Else If Y = 1 Then
        If Z = 0 Then class=false
        Else If Z = 1 Then class=true
Else If X = 1 Then
    If Y = 0 Then
        If Z = 0 Then class=false
        Else If Z = 1 Then class=true
    Else If Y = 1 Then
        If Z = 0 Then class=true
        Else If Z = 1 Then class=false
```

C) This distance function is unusual for k-NN, because it can result in multiple examples (> 2) that are the same distance from the query. We need to make a design decision; here we simply take the majority vote of all neighbours that are the shortest distance from the query.

Example	Nearest neighbours	Majority Vote	Actual Class	Error
1	2,3,5	false	false	0
2	3,5	false	false	0
3	2, 5	false	false	0
4	6,7	true	true	0
5	2,3	false	false	0
6	4, 7	true	true	0
7	4,6	true	true	0
8	4,6,7	true	false	1

So the LOOCV error is $\frac{1}{8}$.

D) We can observe that to represent an "m–of–n function" will result in very complex decision trees or rule sets due to the problem of replicating sub-trees or rules to express all the cases of the function, especially as m, n increase, but k-NN does not suffer from this representational issue, so on this criterion it is a better approach.

We can make the comment, however, that a linear threshold function learner like a perceptron would be a still better choice for this target concept.

Question 3 [15 marks]

Perceptron Learning

(a) [12 marks] Consider the perceptron training rule applied to learn a perceptron for the two-input Boolean function $x_1 \leftarrow x_2$, which can be written as (x_1 OR (NOT x_2)) .

Assume that input values for variables x_1 and x_2 are either 0 or 1. So the set of labelled examples for this function is the set of four possible combinations of values for the variables x_1 and x_2, each labelled with the output of the function y, which will be $+1$ if the function is "true" for those inputs, and -1 otherwise.

Set the learning rate $\eta = 1.0$. Suppose you apply the perceptron training rule, taking as input the four labelled examples just described.

Let's say that after some iterations the following weights have been learned: $w_0 = 1.2, w_1 = 1.0, w_2 = 0.5$. Now, starting with this set of weights, apply the perceptron training rule for 2 epochs to the set of examples.

Specifically, on each epoch, for each example, you will need to compute the output for the perceptron given the current weights and inputs, the weight update value Δw_i for each of the weights w_i and then you need to apply these updates. To get the marks you need to show, at each step, the current weights, inputs, output and class label.

(b) [3 marks] When all weight updates for the examples are applied, has the perceptron learned the target function ? Briefly explain your answer.

Question 3 ANSWER

Start by enumerating the examples for what we are told is the target function, (x_1 OR (NOT x_2)).

x_1	x_2	Class
0	0	+1
0	1	-1
1	0	+1
1	1	+1

We are told that the learning rate $\eta = 1.0$ and we are given the starting weights $w_0 = 1.2, w_1 = 1.0, w_2 = 0.5$.

Running the perceptron training algorithm for 2 epochs should show the following inputs, weights and outputs:

w_0	w_1	w_2	x_0	x_1	x_2	$\mathbf{w} \cdot \mathbf{x}$	Output	Class
Epoch 0								
1.2	1.0	0.5	1	0	0	1.2	+1	+1
1.2	1.0	0.5	1	0	1	1.7	+1	−1
0.2	1.0	-0.5	1	1	0	1.2	+1	+1
0.2	1.0	-0.5	1	1	1	0.7	+1	+1
Epoch 1								
0.2	1.0	-0.5	1	0	0	0.2	+1	+1
0.2	1.0	-0.5	1	0	1	−0.3	−1	−1
0.2	1.0	-0.5	1	1	0	1.2	+1	+1
0.2	1.0	-0.5	1	1	1	0.7	+1	+1

If the output differs from the class then the perceptron has made a mistake.

Since there are no mistakes on the second epoch, we see that the perceptron has learned the target function on this dataset.

Question 4 [20 marks]

Naive Bayes

A) **[2 marks]**
Explain the difference between the *maximum a posteriori* hypothesis h_{MAP} and the *maximum likelihood* hypothesis h_{ML}.

B) **[4 marks]**
Would the Naive Bayes classifier be described as a generative or as a discriminative probabilistic model ? Explain your reasoning informally in terms of the conditional probabilities used in the Naive Bayes classifier.

C) **[10 marks]**
Using the multivariate Bernoulli distribution to model the probability of some type of weather occurring or not on a given day, from the following data calculate the probabilities required for a Naive Bayes classifier to be able to decide whether to play or not.

Use pseudo-counts (Laplace correction) to smooth the probability estimates.

Day	Cloudy	Windy	Play tennis
1	1	1	no
2	0	1	no
3	1	1	no
4	0	1	no
5	0	1	yes
6	1	0	yes

D) **[4 marks]**
To which class would your Naive Bayes classifier assign each of the following days ?

Day	Cloudy	Windy	Play tennis
7	0	0	?
8	0	1	?

Question 4 ANSWER

A) In the Bayesian setting, both h_{MAP} and h_{ML} are single *most probable* hypotheses from the hypothesis space, given training data. However, h_{MAP} is found by taking the product of the likelihood and the prior, whereas h_{ML} is found simply by taking the likelihood.

B) The Naive Bayes classifier is used to determine the most probable class Y given the data X, so we can view this as computing the probability $P(Y|X)$. There are algorithms that learn this probability directly, such as logistic regression (see slide 120 in the lecture on Classification). Such models are characterised as *Discriminative*. On the other hand, the Naive Bayes classifier actually learns the joint probability $P(X|Y)P(Y) = P(Y, X)$. These models are termed *Generative* since they can be used to "generate" (sample) the examples for learning (see slide 67 in the lecture on Classification). Naive Bayes applies Bayes Theorem to actually do the classification.

C) Using the multivariate Bernoulli, treat examples as bit vectors (don't forget for the probability smoothing to add two "pseudo-examples" for each class, one with all bits set to 1 and the other with all 0). We have 2 examples of class 'yes' and 4 examples of class 'no' in the data. Adding bit vectors for each class results in $(2, 4)$ for 'no' and $(1, 1)$ for 'yes'. We need to divide the counts by the number of examples in each class, but first add the counts for the probability smoothing, giving $(\frac{3}{6}, \frac{5}{6}) = (0.5, 0.83)$ for 'no' and $(\frac{2}{4}, \frac{2}{4}) = (0.5, 0.5)$ for 'yes'.

D) To classify an example we need to compute the probabilities of the data given each class, then predict the class with the higher probability.

Day	Cloudy	Windy	Play tennis ?	Probability
7	0	0	no	$(1 - 0.5) \times (1 - 0.83) = 0.09$
			yes	$(1 - 0.5) \times (1 - 0.5) = 0.25$
8	0	1	no	$(1 - 0.5) \times 0.83 = 0.41$
			yes	$(1 - 0.5) \times 0.5 = 0.25$

So for example 7 the prediction is 'yes' and for example 8 it is 'no'.

Question 5 [20 marks]

Neural Learning

A) **[4 marks]**
A *linear unit* from neural networks is a linear model for numeric prediction that is fitted by gradient descent. Explain the differences between the *batch* and *incremental* (or *stochastic*) versions of gradient descent.

B) **[4 marks]**
Stochastic gradient descent would be expected to deal better with local minima during learning than batch gradient descent – true or false ? Explain your reasoning.

A) **[12 marks]**
Suppose a single unit has output o the form:

$$o = w_0 + w_1 x_1 + w_1 x_1^2 + w_2 x_2 + w_2 x_2^2 + \cdots + w_n x_n + w_n x_n^2$$

The problem is to learn a set of weights w_i that minimize squared error. Derive a batch gradient descent training rule for this unit.

Question 5 ANSWER

A) For a linear unit, batch and stochastic gradient descent differ as follows:

- in batch gradient descent the gradient is computed over *all* the examples in the training set before the weight update is applied
- in stochastic gradient descent the gradient is computed for a *single* example, and then the weight update is applied

B) With batch gradient descent, the direction of the gradient is computed over all the training data, so in some sense this is the true gradient. If the algorithm is located near some local minimum then it will move in the direction of the steepest descent and converge at that minimum. However, in stochastic gradient descent, where some example is selected at random, we would expect that the gradient computed for that example may not be the true gradient, so the algorithm may instead move in a different direction, thus avoiding the local minimum. So our answer is 'true'.

C) The approach to deriving the required training rule pretty much follows the method in the lecture slides. We have a single unit with output o of the form:

$$o = w_0 + w_1 x_1 + w_1 x_1^2 + w_2 x_2 + w_2 x_2^2 + \cdots + w_n x_n + w_n x_n^2$$

Using homogeneous coordinates we can write this as:

$$o = \Sigma_{i=0}^{n} w_i (x_i + x_i^2)$$

Assuming the same error function and gradient definition as before (slides 15-18 on the lecture on Neural Learning) we can derive the following:

$$\begin{aligned}
\frac{\partial E}{\partial w_i} &= \frac{\partial}{\partial w_i} \frac{1}{2} \sum_d (t_d - o_d)^2 \\
&= \frac{1}{2} \sum_d \frac{\partial}{\partial w_i} (t_d - o_d)^2 \\
&= \frac{1}{2} \sum_d 2(t_d - o_d) \frac{\partial}{\partial w_i} (t_d - o_d) \\
&= \sum_d (t_d - o_d) \frac{\partial}{\partial w_i} (t_d - \Sigma_{i=0}^{n} w_i (x_i + x_i^2)) \\
\frac{\partial E}{\partial w_i} &= \sum_d (t_d - o_d)(-x_{i,d} - x_{i,d}^2)
\end{aligned}$$

Question 6 [20 marks]

Ensemble Learning

A) **[8 marks]**
As model complexity increases from low to high, what effect does this have on:

1) Bias ?
2) Variance ?
3) Predictive accuracy on training data ?
4) Predictive accuracy on test data ?

B) **[2 marks]**
Is decision tree learning relatively stable ? Describe decision tree learning in terms of bias and variance in no more than two sentences.

C) **[2 marks]**
Is linear regression relatively stable ? Describe linear regression in terms of bias and variance in no more than two sentences.

D) **[2 marks]**
Bagging reduces bias. True or false ? Give a one sentence explanation of your answer.

E) **[2 marks]**
Boosting reduces variance. True or false ? Give a one sentence explanation of your answer.

F) **[2 marks]**
Specify two ways in which boosting uses weights in ensemble learning.

G) **[2 marks]**
Are Random Forests designed to extend bagging, or boosting ?

Question 6 ANSWER

A) These questions should be answered in the setting of sampling theory, which means all errors are taken over a large (possibly infinite) set of training sets of the same size drawn at random from the same target distribution, where the error can be broken down into bias and variance. As model complexity increases from low to high:

   1) Bias will reduce, since the probability of systematic error due to a mismatch between the target model class and the learner's model class will be reduced;
   2) Variance will increase, since the amount of error due to variation over the training samples drawn repeatedly from the target distribution will increase due to increased flexibility in fitting any particular set of training samples;
   3) Predictive accuracy on training data will be increased, since in general increasing model complexity will lead to better fitting of the training data;
   4) Predictive accuracy on test data will increase, since in general increasing model complexity will lead to overfitting of the training data, unless this is controlled by regularisation.

B) Decision tree learning is relatively unstable. This is because a decision tree can be grown to completely fit a training set (except for noise in labelling examples). This can be understood as decision tree learning being in general low bias and high variance.

C) Linear regression is relatively stable. This is because (assuming datasets are "not too small") small changes in the examples appearing in the training set will not change the predictions that much. This can be understood as linear regression learning being in general high bias and low variance.

D) Bagging is designed to reduce variance, not reduce bias, since bagging uses a majority vote of base learning methods, each of which is treated independently of the other base learners and is not changed by the ensemble, so bias is not changed.

E) Boosting is designed to reduce bias, not reduce variance, since multiple instances of a boosted base learning method are combined using an error-based weighted voting scheme on which predictions are made, so most of the error reduction is due to learning one or more base classifiers focused on a particular subset of examples which the other base classifiers do not fit well, rather than variance reduction.

F) Firstly, every *classifier* in the ensemble is assigned a weight based on the (weighted) error of the classifier on the dataset. Secondly, on every boosting iteration each *example* is assigned a weight based on the (weighted) classifier error, and whether the example is correctly classified or not.

G) Random Forests learn an ensemble of trees, and are designed to extend *bagging* by forcing increased diversity with the aim of further reducing variance. This is implemented by using a method within the ensemble to randomly select, on each iteration, a subset of features onto which the training set from which the tree must be learned is projected.

Question 7 [20 marks]

Learning Theory

A) **[7 marks]**
Suppose we have a consistent learner with a hypothesis space restricted to conjunctions of exactly 8 attributes, each with values $\{\text{true}, \text{false}, \text{don't care}\}$. What is the size of this learner's hypothesis space ? Give the formula for the number of examples sufficient to learn with probability at least 95% an approximation of any hypothesis in this space with error of at most 10%. [Note: you are *not* required to compute the solution.]

B) **[8 marks]**
An instance space X is defined using m Boolean attributes. Let the hypothesis space H be the set of decision trees defined on X (you can assume two classes). What is the largest set of instances in this setting which is shattered by H ? [Show your reasoning.]

C) **[3 marks]**
Consider the following description of a concept learning algorithm similar to the HALVING ALGORITHM. The algorithm is initialised with the complete set H of all hypotheses in its (finite) hypothesis space, and the mistake count is set to zero. H is guaranteed to contain the target hypothesis. The training set is guaranteed to be noise-free. On each iteration, the algorithm is given as input a single data instance from the training set, for which it predicts whether or not the instance is in the concept by a *majority vote* of the hypotheses in H. If this prediction is *incorrect*, the mistake count is incremented. All hypotheses that predicted incorrectly are removed from H. What is the worst-case mistake bound of this algorithm ?

THIS QUESTION CONTINUES ON THE NEXT PAGE.

D) [**2 marks**]
Informally, which of the following are consequences of the No Free Lunch theorem:

a) averaged over all possible training sets, the variance of a learning algorithm dominates its bias

b) averaged over all possible target concepts, no learning algorithm has a better off-training set error than any other

c) averaged over all possible target concepts, the bias of a learning algorithm dominates its variance

d) averaged over all possible training sets, no learning algorithm has a better off-training set error than any other

e) averaged over all possible target concepts and training sets, no learning algorithm is independent of the choice of representation in terms of its classification error

Question 7 ANSWER

A) If there are 8 attributes, each with 3 values, and the hypothesis space H can be formed by conjunctions of attributes, then the size of hypothesis space H is 3^8.

Recall the formula for the sample complexity for a consistent learner:

$$m \geq \frac{1}{\epsilon}(\log(|H|) + \log(\frac{1}{\delta}))$$

Here $\epsilon = 0.1$ and $\delta = 1 - 0.95 = 0.05$, so the expression for the number of examples required

$$m \geq \frac{1}{0.1}(\log(3^8) + \log(\frac{1}{0.05}))$$

B) There are $n = 2^m$ possible instances in the instance space X. The number of possible dichotomies is the number of possible subsets of the instance space (each subset can be labelled as 'positive' and its complement labelled 'negative'), which is 2^n.

Each subset of instances can be uniquely defined by a conjunction of Boolean attributes, which can be represented in a decision tree by the path from root node to leaf node. So any dichotomy can be represented by a decision tree defined in this way.

So the size of the largest set of instances in this setting that can be shattered by H is the size of the instance space X, which is n.

C) The worst-case mistake-bound for this algorithm is $\lfloor \log_2 |H| \rfloor$ where H is the (initial) hypothesis space.

Informally, this can be explained as a kind of "binary chop" procedure. For each instance, the algorithm makes a classification based on a majority vote of the hypotheses in the hypothesis space. If the predicted class is the same as the actual class, there is no mistake, otherwise there is. However, in *either* case all hypotheses that predicted incorrectly are eliminated. So on each mistake, at least half of the hypotheses will be eliminated (because of majority voting).

D) Answer is b) averaged over all possible target concepts, no learning algorithm has a better off-training set error than any other

END OF PAPER