CSC 411 F 2017 Midterm Test
Duration — 60 minutes
Aids allowed: none

Student Number: |_|_|_|_|_|_|_|_|_|_|_|

Last Name: _____ **First Name:** _____

---

*Do **not** turn this page until you have received the signal to start.*
(Please fill out the identification section above and read the instructions below.)
Good Luck!

This midterm consists of 5 questions on 8 pages (including this one). *When you receive the signal to start, please make sure that your copy is complete.*

1. Do not turn the page until told to do so.
2. If a question asks you to do some calculations, you must *show your work* to receive full credit.
3. You can use either pen or pencil for the exam. **But please be aware that you are not allowed to dispute any credit after the exam is returned if you use a pencil.**
4. Use the back of the page if you need more space on a question. If you require additional paper you may request some from an invigilator.
5. If you use any space for rough work, indicate clearly what you do not want marked.
6. Lastly, enjoy the problems!

1: _____/ 15

2: _____/ 5

3: _____/ 30

4: _____/ 30

5: _____/ 20

TOTAL: _____/100

Question 1. [15 MARKS]

Mark whether the following statements are true or false by placing a tick in the corresponding column for each row.

Statement	True	False
Assume that you have training data with continuous features and targets. Linear regression must be linear in both parameters and features.		
Assume that you have training data with continuous features and binary class labels and that you are not using any feature parameterization. Logistic regression has a linear decision boundary.		
Consider a discrete random variable X. Higher entropy, $H(X)$, implies lower uncertainty about samples drawn from the distribution of X.		
Consider a dataset where each data point has d-dimensional continuous features and discrete class labels. For the dataset's points to be linearly separable, each of its classes have to be separated exactly by hyperplane decision boundaries.		
Assume that you have training data with binary features and discrete class labels. Logistic regression will always have better classification accuracy on test data than Bernoulli Naive Bayes.		

Question 2. [5 MARKS]

Fill in the blanks below.

- Given discrete random variables X and Y. The Information Gain in Y due to X is

  $$IG(Y, X) = H(.....) - H(.....)$$

  where H is the entropy.

- Given model parameters θ and some data observations X. Then the posterior probability $\propto$ ..... $\times$ ..... (You may write this in words)

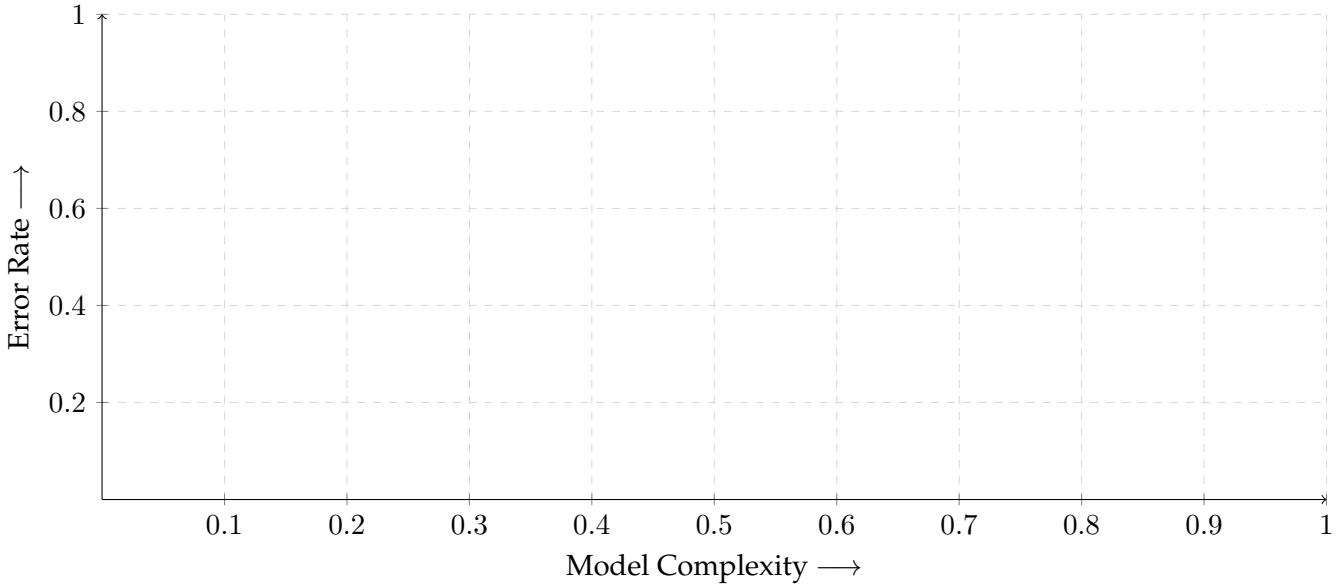
Question 3. [30 MARKS]

This question asks you to show your general understanding of underfitting and overfitting as they relate to model complexity and training set size. A simple example of model complexity might be the number of parameters in a linear regression model. Given the provided axes, where both vertical and horizontal are scaled between 0 and 1, draw your graphs with increasing error upwards and increasing complexity/training set size rightwards.

For this question you will compare to the *Bayes error rate*. The Bayes error rate is the optimal error rate possible given perfect knowledge and no computational limitations. For example, if the targets are deterministic given the features then the Bayes error rate will be zero.

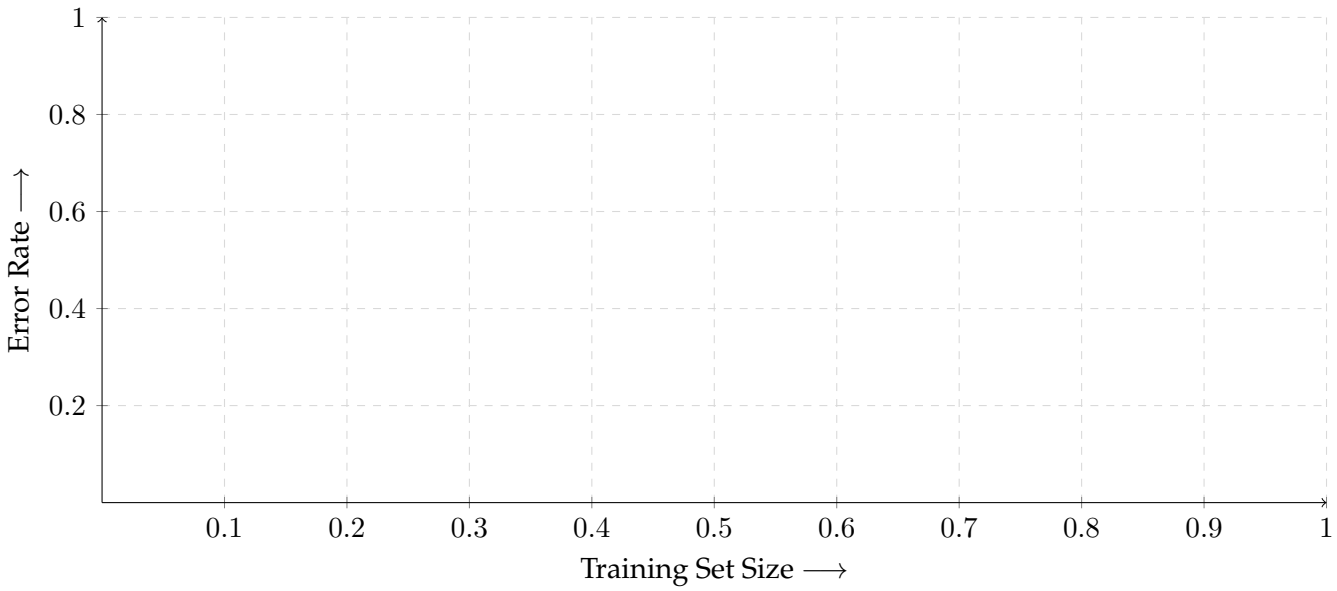
Part (a) [15 MARKS]

For a fixed training set size, sketch a graph of the typical behaviour of training error rate versus model complexity in a learning system. Add to this graph a curve showing the typical behaviour of test error rate (for an infinite test set drawn independently from the same input distribution as the training set) versus model complexity, on the same axes. Mark a horizontal line showing the Bayes error rate, which one of the lines should cross.



Part (b) [15 MARKS]

For a fixed model complexity, sketch a graph of the typical behaviour of training error rate versus training set size in a learning system. Add to this graph a curve showing the typical behaviour of test error rate (again on an iid finite test set) versus training set size, on the same axes. Mark a horizontal line showing the Bayes error rate.

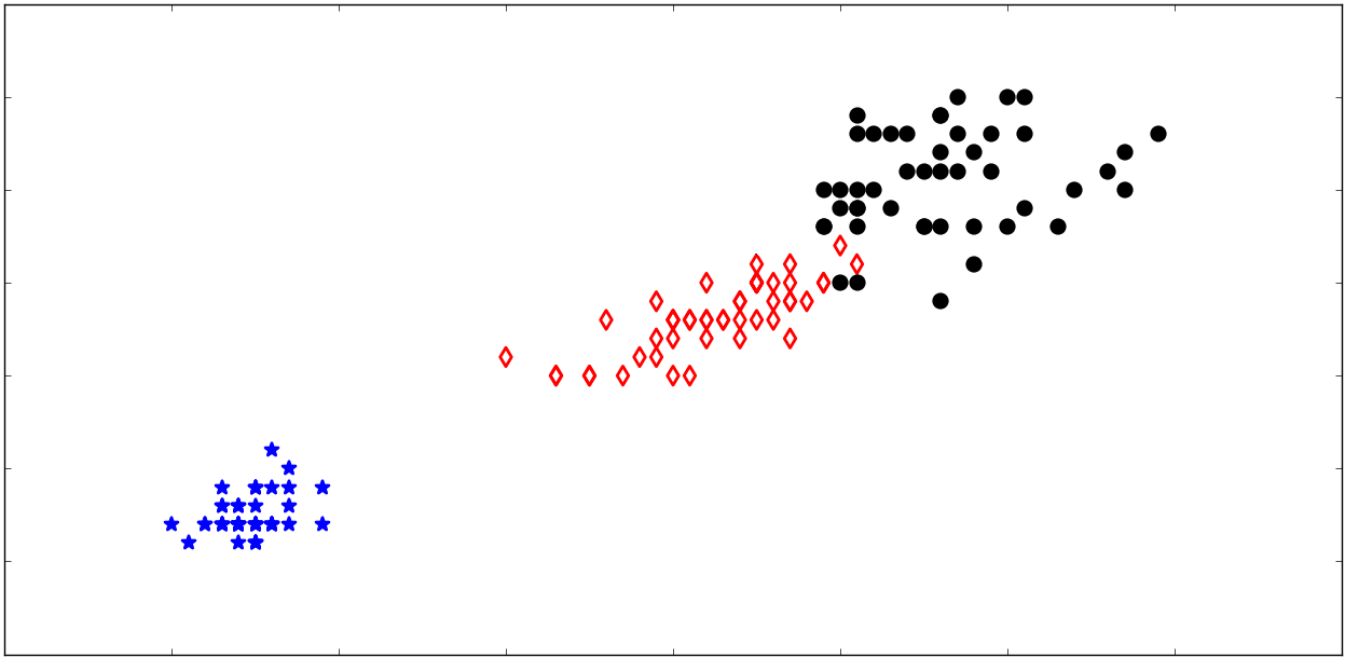


Question 4. [30 MARKS]

Given the following dataset, for each classifier draw the decision boundary

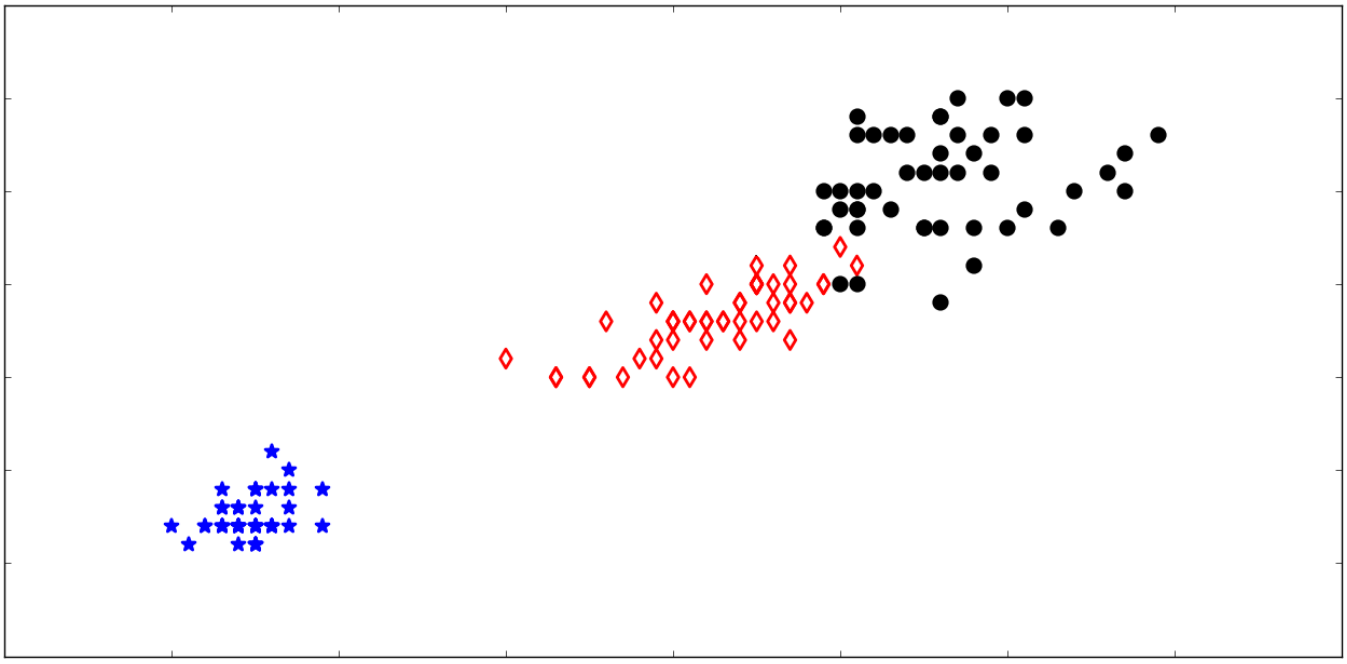
Part (a) [10 MARKS]

Decision tree:



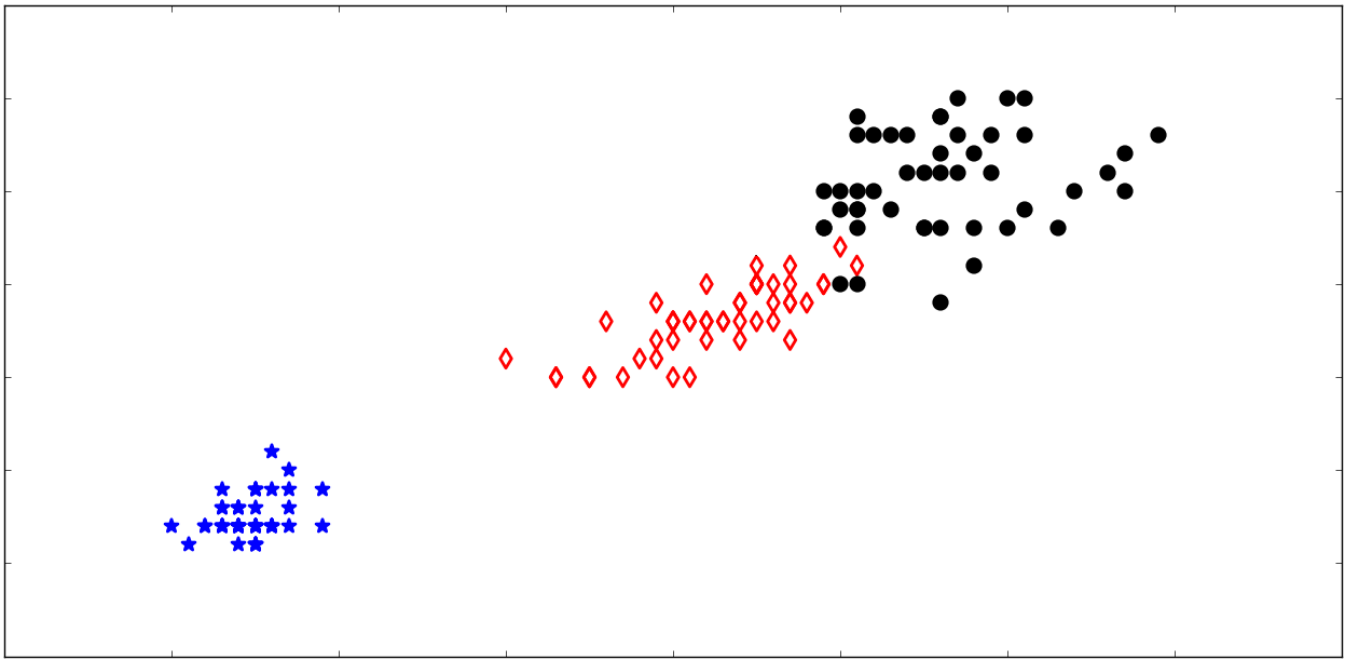
Part (b) [10 MARKS]

1-nearest neighbor:



Part (c) [10 MARKS]

Multi-class logistic regression:



Question 5. [20 MARKS]

We define the Bayes-optimal predictor $h^*(x) = \arg\min_{\hat{y}} \mathbb{E}_y[\ell(y, \hat{y})|x]$. We now assume y only takes a finite number of values $y_1, ..., y_k$. For the L_2 loss this means that

$$h^*(x) = \arg\min_{\hat{y}} \mathbb{E}_y[\ell(y, \hat{y})|x] = \arg\min_{\hat{y}} \sum_{j=1}^{k} (y_j - \hat{y})^2 p(y_j|x)$$

Show that for the L_2 loss $h^*(x) = \mathbb{E}_y[y|x] = \sum_{j=1}^{k} y_j p(y_j|x)$

Print your name in this box.

END OF EXAMINATION