Midterm for CSC2515, Machine Learning Fall 2019 Wednesday, Oct. 30, 4:10–5:40pm

Name: _____

Student number: _____

This is a closed-book test. It is marked out of 100 marks. Please answer ALL of the questions. Here is some advice:

- The questions are NOT arranged in order of difficulty, so you should attempt every question.
- Questions that ask you to "briefly explain" something only require short (1-3 sentence) explanations. Don't write a full page of text. We're just looking for the main idea.
- None of the questions require long derivations. If you find yourself plugging through lots of equations, consider giving less detail or moving on to the next question.
- Many questions have more than one right answer.
- If you need extra room, you may continue your answer on the scratch pages at the end. To receive points, you must indicate you are doing so.

.

Q1:	_____	/	6
Q2:	_____	/	6
Q3:	_____	/	6
Q4:	_____	/	12
Q5:	_____	/	8
Q6:	_____	/	10
Q7:	_____	/	12
Q8:	_____	/	10
Q9:	_____	/	12
Q10:	_____	/	8
Q11:	_____	/	10

Final mark: _____ / 100

1. **[6 points]** TRUE or FALSE: for any two neighboring Voronoi cells C_1 and C_2 corresponding to two training examples $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$, the boundary between C_1 and C_2 is part of the decision boundary of the 1-nearest-neighbor classifer. Briefly justify your answer.

2. **[6 points]** In lecture, we originally introduced binary linear classifiers in terms of a linear function followed by a threshold:

$$z = \mathbf{w}^T\mathbf{x} + b$$
$$y = \begin{cases} 1 & \text{if } z \geq r \\ 0 & \text{if } z < r \end{cases}$$

   We saw that it is safe to assume the threshold $r = 0$. Briefly explain why we may assume this without limiting the expressive power of the model.

3. **[6 points]** Give one example of a situation where you would prefer to use L^2 regularization rather than L^1. Briefly justify your answer.

4. [**12 points**] Carol and Dave are each trying to predict stock prices using neural networks. They formulate this as a regression problem using squared error loss. Carol trains a single multilayer perceptron on a certain training set and uses its predictions on the test set. Dave trains 5 different networks (using exactly the same architecture, training data, etc. as Carol) starting with different random initializations, and averages their predictions on the test set.

   For each of the following questions, please briefly and informally justify your answer. You do not need to provide a mathematical proof.

   (a) [**4 points**] Compared with Carol's approach, is the Bayes error for Dave's approach HIGHER, LOWER, or THE SAME?

   (b) [**4 points**] Compared with Carol's approach, is the bias for Dave's approach HIGHER, LOWER, or THE SAME?

   (c) [**4 points**] Compared with Carol's approach, is the variance for Dave's approach HIGHER, LOWER, or THE SAME?

5. **[8 points]** In lecture, we considered using linear regression for binary classification. Here, we use a linear model

$$y = \mathbf{w}^\top \mathbf{x} + b$$

and squared error loss $\mathcal{L}(y, t) = \frac{1}{2}(y-t)^2$. We saw this has the problem that it penalizes confident correct predictions. Will it fix this problem if we instead use absolute value loss, $\mathcal{L}(y, t) = |y - t|$? Justify your answer mathematically.

6. **[10 points]** Recall the soft-margin SVM optimization problem:

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \ & \frac{1}{2}\|\mathbf{w}\|_2^2 + \gamma \sum_{i=1}^{N} \xi_i \\ \text{s.t.} \ & t^{(i)}(\mathbf{w}^\top \mathbf{x}^{(i)} + b) \geq 1 - \xi_i \qquad i = 1, \ldots, N \\ & \xi_i \geq 0 \qquad\qquad\qquad\qquad\quad i = 1, \ldots, N \end{aligned}$$

What will go wrong if we eliminate the $\frac{1}{2}\|\mathbf{w}\|_2^2$ term from the cost function? Justify your answer mathematically.

7. [**12 points**] When we analyzed KNN, we assumed the training examples were sampled densely enough so that the true conditional probability $p(t \mid \mathbf{x})$ is approximately constant in the vicinity of a query point $\mathbf{x}_\star$. Suppose it is a binary classification task with targets $t \in \{0, 1\}$ and $p(t = 1 \mid \mathbf{x}_\star) = 0.6$.

   (a) [**6 points**] What is the asymptotic error rate at $\mathbf{x}_\star$ for a 1-nearest-neighbor classifier? (By asymptotic, I mean as the number of training examples $N \to \infty$.) Justify your answer.

   (b) [**6 points**] Approximately what is the asymptotic (as $N \to \infty$) error rate at $\mathbf{x}_\star$ for a K-nearest-neighbors classifier when K is very large? Justify your answer.

8. **[10 points]** Consider a pair of random variables X and Y whose joint distribution is as follows:

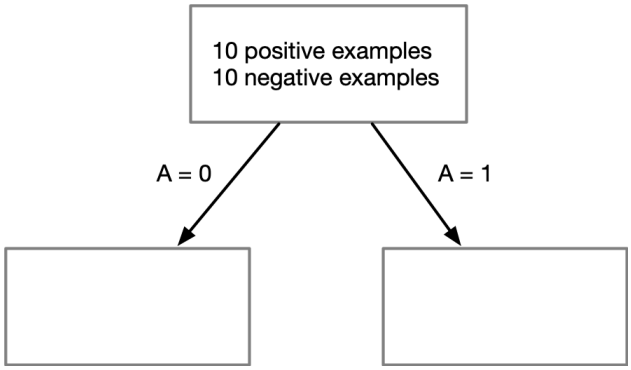
	$Y=0$	$Y=1$
$X=0$	0	0.5
$X=1$	0.25	0.25

(a) **[4 points]** Compute the joint entropy $\mathcal{H}(X, Y)$.

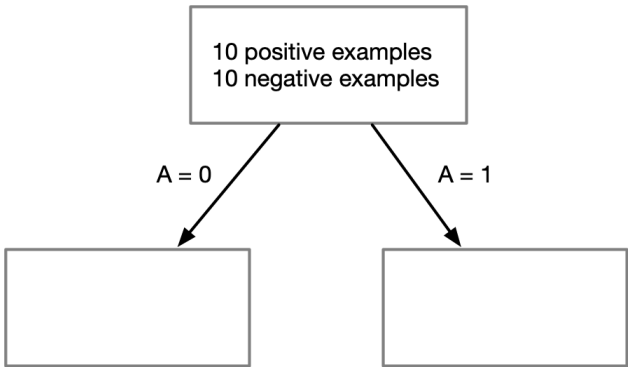
(b) **[6 points]** Compute the conditional entropy $\mathcal{H}(Y \mid X)$.

9. [**12 points**] Suppose we are deciding how to split a particular node in a decision tree. Currently 20 training examples (10 positive and 10 negative) belong to this node. If we split on a certain binary-valued attribute A, then some examples will go down the left branch and others will go down the right branch.

   (a) [**6 points**] What is the *minimum* possible value of the information gain for a split? Give an example of a split that achieves this value by indicating the numbers of training examples that go down each branch. Justify your answer.



   (b) [**6 points**] What is the *maximum* possible value of the information gain for a split? Give an example of a split that achieves this value by indicating the numbers of training examples that go down each branch. Justify your answer.



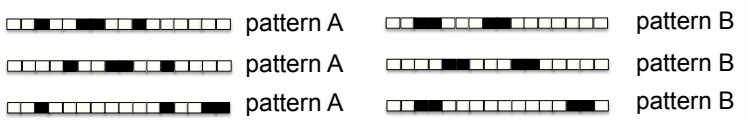
10. **[8 points]** Consider the following convolution layer in a conv net. The input has a spatial dimension of 12×12, and has 10 channels. The convolution kernels are 3×3, the stride is 2, and we use "valid" convolution, which means that each output neuron only looks at image regions that lie entirely within the spatial bounds of the input. The output dimension is 5×5, with 20 channels.

    For this question, you don't need to show your work or justify your answer, but doing so may help you get partial credit.

    (a) **[4 points]** How many weights are required for this convolution layer?

    (b) **[4 points]** Suppose we instead make this a locally connected layer, i.e. don't use weight sharing. How many weights are required?

11. [**10 points**] Recall the example from lecture that we showed linear classifiers were unable to classify. The task is to distinguish the following two patterns in all 16 possible translations with wrap-around. The inputs are 16-dimensional binary vectors, where black indicates 1 and white indicates 0.



Now let's classify these patterns using a conv net.

- First we have a convolution layer with a single convolution kernel of size 3 with weights $\mathbf{w} = (w_1\ w_2\ w_3)^\top$ and bias b. Unlike in ordinary convolution layers, this one will use wrap-around. The output of this layer has 16 units.
- We apply the ReLU activation function to this layer.
- We pool together the activations by taking the sum. Call this value z.
- We threshold the result at a value r. If $z \geq r$, it is classified as B, otherwise it's classified as A.

Your task is to choose the weights $\mathbf{w}$, bias b, and threshold r to correctly separate all instances of the patterns. You are not required to show your work, but explaining your reasoning may help you get partial credit.

(Scratch work or continued answers)