CSC 2515 TEST SAMPLE

Answer exactly 5 questions in Part A. Answer exactly 3 questions in part B. Each question in part A is worth 8 points. Each question in part B is worth 20 points. Please put an X through those questions you do not want us to mark. If you answer more than 5 from A, 3 from B we will randomly select which to grade!

PART A

1. What is a “support vector”?

2. Consider the following learning rule:

\[ w_{ji}^{\text{new}} = w_{ji}^{\text{old}} - \eta \sum_n (o_j^n - t_j^n)x_i^n - 2\alpha w_{ji}^{\text{old}} \]

What is the primary aim of the last term on the right-hand side? What effect does this term have on the network weights?
3. Explain the different strategies used by ensemble methods that use parallel versus sequential training, with respect to how they try to improve the error rate of a single model? Which strategy do Bagging and Boosting respectively use?

4. What is the main difference between discriminative and generative classifiers? Give an example of each type.

5. Consider a learning problem with 2D features. How are the decision tree and 1-nearest neighbor decision boundaries related (i.e., how are they the same, and how are they different)?
1. Mixture Models

Consider a simple form of mixture model, in which each mixture component is a spherical Gaussian density of dimension $d$, and $z$:

$$ p(x|\theta) = \sum_{k=1}^{K} P(z = k|\theta) p(x|z = k, \theta_k) $$

$$ p(x|z = k, \theta_k) = \frac{1}{(2\pi \sigma_k^2)^{d/2}} \exp \left( - \frac{|x - \mu_k|^2}{2\sigma_k^2} \right) $$

(A). Express the likelihood of a dataset of $n$ unlabeled samples $\{x_1, \ldots, x_n\}$ drawn independently from this mixture density.

(B). If there is one component in the mixture ($K = 1$), can standard gradient ascent be used to update the model parameters? What about if $K = 2$? Explain your answers.

(C). Consider increasing the number of components by one. How do you expect this to impact the likelihood of the data?
2. SVMs

<table>
<thead>
<tr>
<th>Data Case</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>0</td>
<td>+</td>
</tr>
<tr>
<td>2</td>
<td>-1</td>
<td>-2</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>+</td>
</tr>
<tr>
<td>4</td>
<td>-1</td>
<td>0</td>
<td>-</td>
</tr>
</tbody>
</table>

(A). Consider fitting a linear SVM to the following data set:

Plot the points below with $x_1$ along the x-axis and $x_2$ along the y-axis. Write down the decision rules the linear SVM will find for class (+) and class (-) in terms of $x_1$ and $x_2$. Draw the decision boundaries. What is/are the support vector(s) for this data set?

(B). What is the kernel trick, and what computational purpose does it serve?
(C). How does a non-linear SVM differ from a linear SVM?

(D). Describe two important differences between a non-linear SVM and a neural network.
3. Neural Networks

Consider a two-layer neural network to learn a function \( f : X \rightarrow Y \), where \( X = (X_1, X_2) \) consists of two attributes. The weights \( w_1, \ldots, w_6 \) can be arbitrary. There are two possible choices for the activation functions for the units in the network:

- **S**: the signed sigmoid function \( S(a) = \text{sign}[\sigma(a) - 0.5] = \text{sign}[\frac{1}{1+\exp(-a)} - 0.5] \)
- **L**: the linear function \( L(a) = ca \)

where in both cases \( a = \sum_i w_i X_i \).

(A). Assign proper activation functions (S or L) to each unit in the following graph so that this network simulates linear regression: \( Y = \beta_1 X_1 + \beta_2 X_2 \).

(B). Assign proper activation functions (S or L) to each unit in the following graph so that this network simulates a binary logistic regression classifier: \( Y = \arg \max_y P(Y = y|X) \), where \( P(Y = 1|X) = \frac{\exp(\beta_1 X_1 + \beta_2 X_2)}{1+\exp(\beta_1 X_1 + \beta_2 X_2)} \), and where \( P(Y = -1|X) = \frac{1}{1+\exp(\beta_1 X_1 + \beta_2 X_2)} \).
(C). Based on the network defined in part B, derive $\beta_1$ and $\beta_2$ in terms of $w_1, ..., w_6$.

(D). Assign proper activation functions (S or L) to each unit in the following graph so that this network simulates an ensemble classifier which combines two binary logistic regression classifiers $f_1 : X \rightarrow Y_1$ and $f_2 : X \rightarrow Y_2$ to produce its final prediction: $Y = \text{sign}[\alpha_1 Y_1 + \alpha_2 Y_2]$. Use the definition of logistic regression classifier from part B.

(E). Based on the network defined in part D, express $\alpha_1$ and $\alpha_2$ in terms of $w_1, ..., w_6$. 
4. Q Learning

Consider this familiar navigation task, shown on the left above. You can move in any of four directions (left/right/up/down). If you hit the blob at B2, you remain in the same state. The rewards are $+10$ for state A4 and $-10$ for B4; these are both absorbing states. The reward for every other state is 0.

(A). Assume that the state transitions are deterministic. Recall that under the simple Q-learning algorithm, the estimated Q values are updated using the following rule:

$$\hat{Q}(s, a) = r + \gamma \max_{a'} \hat{Q}(s', a')$$

Consider applying this algorithm when all the $\hat{Q}$ values are initialized to zero; and $\gamma = 0.9$.

Indicate Q values on the figure after repeatedly cycling through the following episodes:

- A1,A2,A3,B3,B4
- C2,C1,B1,A1,A2,A3,A4
- C4,C3,B3,A3,A4

(B). Assume the robot will now use the policy of always performing the action having the greatest Q value. Indicate this policy on the figure. Is it optimal?
(C). We have seen several examples of greedy approaches to learning in this class, such as decision tree learning and boosting. What are the strengths and weaknesses of a greedy action strategy for reinforcement learning?

(D). Imagine that this grid world is much larger, and contains multiple obstacles and terminal states with positive and negative rewards. Define some features that could be used to enable generalization in this setting.