CS760 Final Exam Practice Questions

December 16, 2021

Instructions: Answer the following questions. You may write out solutions on extra paper, but label your answers clearly. Read all of the questions first. Even if you cannot obtain a final answer, make sure to write your setup and explain how you would obtain the answer. Partial credit will be considered. Do not spend too much time on any single problem—some problems are harder than others.

1 True/False Questions

1. Perceptron is a generative model. [True/False] $\boxed{False}$
2. A Naive Bayes classifier with n Boolean features and a Boolean label need to estimate 2n+1 parameters. [True/False] $\boxed{True}$
3. Function approximation in Q learning may speed up learning in some problems, because the function approximation can be generalized to previously unseen states [True/False] $\boxed{True}$
4. For an infinite horizon MDP with a finite number of states and actions and with a discount factor γ, where $0 < \gamma < 1$, value iteration is guaranteed to converge. [True/False] $\boxed{True}$
5. LSTM cells use gates to better control information flow and lengthen memory. [True/False] $\boxed{True}$
6. Exploration is unnecessary in finding optimal state-action value pairs. [True/False] $\boxed{False}$
7. PCA can only be applied to square data matrices. [True/False] $\boxed{False}$
8. Standard Q learning (using Q tables) only applies to discrete action and state spaces. [True/False] $\boxed{True}$
9. Transformers usually show better performance than LSTM in longer sequences. [True/False] $\boxed{True}$
10. Hierarchical clustering of n points can produce a tree of depth n. [True/False] $\boxed{False}$

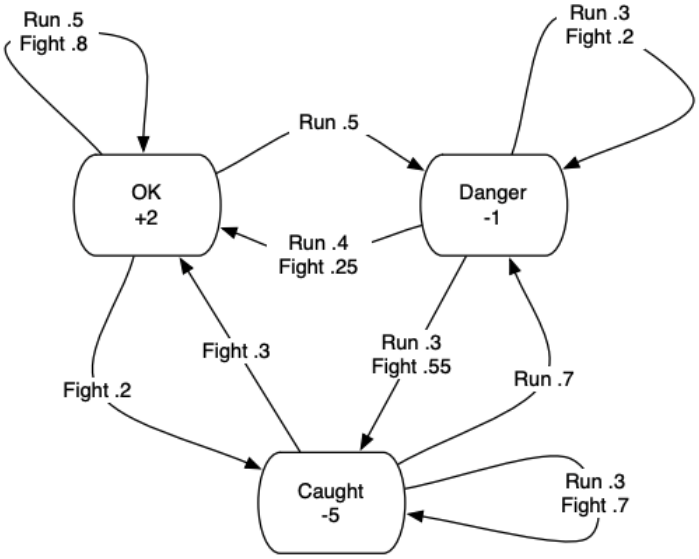

Figure 1: MDP of Dangerous School

2 Reinforcement Learning

A person is being chased around the school yard by bullies and must choose whether to Fight or Run. There are three states:

- Ok (O), where they are fine for the moment.
- Danger (D), where the bullies are right on their heels.
- Caught (C), where the bullies catch up with them.

1. Fill out the table with the results of value iteration with a discount factor $\gamma = 0.9$

k	$V_k(O)$	$V_k(D)$	$V_k(C)$
1	2	-1	-5
2	2.54	-1.9	-6.98

Initialize $V_0(O) = V_0(D) = V_0(C) = 0$ and apply value iteration.

$$V_{i+1}(s) = r(s) + \gamma max_a \sum_{s'} P(s'|s,a)V_i(s')$$

Then, $V_1(O) = 2$, $V_1(D) = -1$, $V_1(C) = -5$. And,

$$
\begin{aligned}
V_2(O) &= 2 + \max(0.9((0.5 * 0.2) + (0.5 * -1)), 0.9((0.8 * 0.2) + (0.2 * -5)) \\
&= 2 + \max(0.45, 0.54) \\
&= 2.54 \\
V_2(D) &= -1 + \max(0.9((0.4 * 0.2) + (0.3 * -1) + (0.3 * -5)), 0.9((0.25 * 2) + (0.2 * -1) + (0.55 * -1))) \\
&= -1 + \max(-0.9, -2.205) \\
&= -1.9 \\
V_2(C) &= -5 + \max(0.9((0.7 * -1) + (0.3 * -5)), -5 + 0.9((0.3 * 2) + (0.7 * -5))) \\
&= -5 + \max(-1.98, -2.61) \\
&= -6.98
\end{aligned}
$$

2. At $k = 2$ with $\gamma = 0.9$ what policy would you select? Is it necessarily true that this is the optimal policy? At $k = 3$ what policy would you select? Is it necessarily true that this is the optimal policy?

Take $\pi^*(s) = argmax_a \sum_{s'} P(s'|s, a)V_k(s)$. Then for $k = 2$ and $k = 3$,

- From O choose Fight
- From D choose Run
- From C choose Run

Value iteration that has not converged is not guaranteed to find the optimal policy, so this policy is not necessarily optimal.

k	$V_k(O)$	$V_k(D)$	$V_k(C)$
1	2	-1	-5
2	2.54	-1.9	-6.98
3	2.57	-2.48	-8.08
4	2.40	-2.93	–8.75

3 Clustering

1. Consider the 2D data points shown in Figure 2. Each dot represents a data point here. You can shift the x-axis by 0.5 so that all co-ordinates are integers. (Horizontal gap b/w consecutive points is of 1-unit and the points on top of each-other are 3-units apart)

   (a) Run k-means (k=2) clustering on this data with given cluster centers (circled points) choices of initial cluster centers.

   (b) Give values of initial cluster centers that will lead to highly imbalanced clusters.

   (a) There will be two clusters, one with the points points left to the left center (including the one below it) and another cluster with points right to the right center (including the point below it)

   (b) If cluster centers are two leftmost (or rightmost) consecutive horizontal points, then one cluster will have just two points and other cluster will have rest of the points in the first epoch. However, in each epoch we get new centers and due to this, if you run k-means until convergence you will get balanced clusters.

2. Will k-means algorithm work well in settings like the ones shown in Figure 3. Justify? If not, how will you make use of the kernel trick in k-means to upgrade it?

   K-means algorithm will fail to cluster the points in shown in the figure, since K-means creates linear boundaries between clusters. To use kernel-method here, we transform each data point $\mathbf{x}_i$ to $\phi(\mathbf{x}_i)$ and each cluster center $\mathbf{c}_j$ to $\phi(\mathbf{c}_j)$. Then consider the euclidean distance between two points in this space, $||\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)||_2^2 = k(\mathbf{x}_i, \mathbf{x}_i) - 2k(\mathbf{x}_i, \mathbf{x}_j) + k(\mathbf{x}_j, \mathbf{x}_j)$. Using this we can do k-means clustering in the ϕ space induced by some kernel function $k(\cdot, \cdot)$.

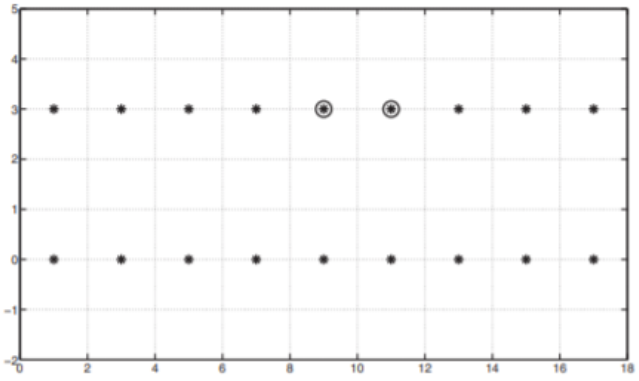

Figure 2: Data For K-means Clustering

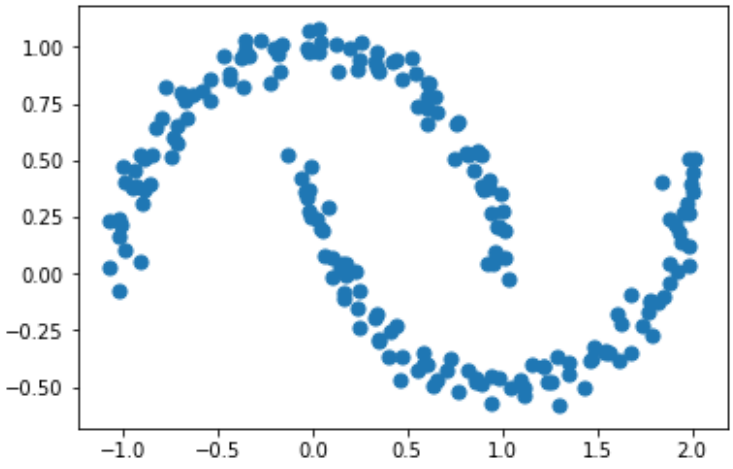

Figure 3: Data For K-means Clustering

4 Maximum Margin Classifier

Consider two 1-D data points, $(x_1 = 0, y_1 = -1)$ and $(x_2 = \sqrt{2}, y_2 = 1)$. (Here y_i are the labels and x_i are the features.) Let $\phi : \mathbb{R} \mapsto \mathbb{R}^3$, be given by $\phi(x) = [1, \sqrt{2}x, x^2]^T$. First, using ϕ transform x from $\mathbb{R}$ to $\mathbb{R}^3$ i.e. now you are working with a classification problem in $\mathbb{R}^3$. Then, answer the following questions considering the primal formulation of SVM in the new space ($\mathbb{R}^3$). (Assume you also have the bias term.)

1. Give a vector $w \in \mathbb{R}^3$ that is parallel to the optimal $w^* \in \mathbb{R}^3$
   $\phi(x_1) = [1, 0, 0]$ and $\phi(x_2) = [1, 2, 2]$
   It is easy to see that w is parallel to [0,2,2]

2. What is the value of the margin (γ) achieved by this w. $\gamma = \sqrt{2}$

3. Let the margin be $\gamma = \frac{1}{||w||}$. Then find w which satisfies the conditions of above two steps.
   $w = (0, \frac{1}{2}, \frac{1}{2})$

4. Using your estimate of w in previous step, find the bias term.
   bias $b = -1$

5 Dimensionality Reduction

Let $\{\mathbf{x}_i\}_{i=1}^n$ be given data points in $\mathbb{R}^d$ and let $\{\mathbf{v}_i\}_{i=1}^K$ be the first K principal directions (components) with $z_{ij} = \mathbf{x}_i^T \mathbf{v}_j$. Then for some data point $\mathbf{x}_i$ show that,

$$r(\mathbf{x}_i) := ||\mathbf{x}_i - \sum_{j=1}^{K} z_{ij}\mathbf{v}_j||_2^2 = ||\mathbf{x}_i||_2^2 - \sum_{j=1}^{K} \mathbf{v}_j^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{v}_j$$

Now assume that $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. What is $\mathbb{E}_{\mathbf{x}}[r(\mathbf{x})]$? Find $\{\mathbf{v}_j\}_{j=1}^K$ that minimize this expectation. What can you say qualitatively about reducing dimensionality in this setting?

$$\begin{aligned}
r(\mathbf{x}_i) &= (\mathbf{x}_i - \sum_{j=1}^{K} z_{ij}\mathbf{v}_j)^T (\mathbf{x}_i - \sum_{j=1}^{K} z_{ij}\mathbf{v}_j) \\
&= \mathbf{x}_i^T \mathbf{x}_i - 2\sum_j z_{ij}\mathbf{x}_i^T \mathbf{v}_j + \sum_k \sum_j z_{ij} z_{ik} \mathbf{v}_j^T \mathbf{v}_k \\
&= ||\mathbf{x}_i||_2^2 - 2\sum_j z_{ij}\mathbf{x}_i^T \mathbf{v}_j + \sum_{j \neq k} z_{ij} z_{ik} \mathbf{v}_j^T \mathbf{v}_k + \sum_j z_{ij}^2 \mathbf{v}_j^T \mathbf{v}_j \\
&= ||\mathbf{x}_i||_2^2 - 2\sum_j z_{ij}\mathbf{x}_i^T \mathbf{v}_j + \sum_j z_{ij}^2 \qquad \text{Using } \mathbf{v}_j^T \mathbf{v}_j = 1 \text{ and } \mathbf{v}_j^T \mathbf{v}_k = 0 \\
&= ||\mathbf{x}_i||_2^2 - 2\sum_j z_{ij}\mathbf{x}_i^T \mathbf{v}_j + \sum_j (\mathbf{x}_i^T \mathbf{v}_j)^T (\mathbf{x}_i^T \mathbf{v}_j) \\
&= ||\mathbf{x}_i||_2^2 - \sum_{j=1}^{K} \mathbf{v}_j^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{v}_j
\end{aligned}$$

$\mathbb{E}_{\mathbf{x}}[r(\mathbf{x})] = d - \sum_{j=1}^K \mathbf{v}_j^T \mathbf{v}_j$ It is minimized when $K = d$. That means $\{\mathbf{v}_j\}_{j=1}^d$ have to be mutually orthogonal, unit norm vectors in $\mathbb{R}^d$. One such choice is standard basis vectors in $\mathbb{R}^d$ i.e. $\mathbf{v}_j = \mathbf{e}_j$. Here $\mathbf{e}_j \in \mathbb{R}^d$ with j^{th} co-ordinate =1 and 0 at other co-ordinates.