Name and CU email address: _____

Exam 3 (Final Exam) CSCI 4622: Machine Learning
Fall 2019 Instructor: C. Monteleoni

This exam has 7 questions, for a total of 40 points and 8 bonus points. Bonus points can be used to increase your score. Note: a total score of 40 will be considered 100%.

Definitions:

- $|a|$ denotes the absolute value of scalar a.

For a vector $x \in \mathbb{R}^d$, the following norms are defined as follows:

- L2 norm: Note: when there is no subscript to the norm, we will assume it is L2.

$$\|x\| = \|x\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2}$$

- L1 norm:

$$\|x\|_1 = \sum_{i=1}^{d} |x_i|$$

For grading. Please do not write here:

Question	Points	Bonus Points	Score
1	10	0	
2	8	0	
3	3	0	
4	2	0	
5	6	0	
6	11	4	
7	0	4	
Total:	40	8	

1. Total for Question 1: 10

For each of the following statements, check the box indicating whether the statement is True or False.

(a) (1 point) Feature selection is a form of dimensionality reduction.
$\surd$ **True** ☐ False

(b) (1 point) In most cases, when a model is overfit, its training error is greater than its true test error.
☐ True $\surd$ **False**

(c) (1 point) In training standard Generative Adversarial Networks (GANs), the training data is input in to the Generator network.
☐ True $\surd$ **False**

(d) (1 point) Binary classification is a special case of multiclass classification.
$\surd$ **True** ☐ False

(e) (1 point) The sequence of observations (output symbols) sampled from a hidden Markov model (HMM) satisfies the first-order Markov property.
☐ True $\surd$ **False**

(f) (1 point) It is NP-hard to minimize the k-center clustering objective.
$\surd$ **True** ☐ False

(g) (1 point) The k-means clustering algorithm minimizes the k-means clustering objective.
☐ True $\surd$ **False**

(h) (1 point) The L1 measure defined on $\mathbb{R}^d$ is a valid *metric*.
$\surd$ **True** ☐ False

(i) (1 point) The L2 measure defined on $\mathbb{R}^d$ is a valid *metric*.
$\surd$ **True** ☐ False

(j) (1 point) For a data set in $\mathbb{R}^d$, for $k < d$, the k-PCA embedding of the data set preserves all the information in the data set.
☐ True $\surd$ **False**

2. Total for Question 2: 8

For each question below, choose **one** answer.

(a) (1 point) Which Autoencoder architecture will map the input data to a higher dimensional space?

☐ A dense autoencoder

√ **A wide autoencoder**

(b) (1 point) Which clustering objective favors locating a cluster near an outlier?

√ **The k-center objective**

☐ The k-means objective

(c) (1 point) In multiclass learning with K classes, which problem framework involves learning $\binom{K}{2}$ binary classifiers?

√ **All-Pairs**

☐ One-Vs-All

(d) (2 points) The first step of standard spectral clustering is to build the Affinity Matrix. For n data points in $\mathbb{R}^d$, the Affinity Matrix will have the following dimensions:

☐ $n \times d$

√ $n \times n$

☐ $d \times d$

(e) (3 points) Once a Variational Autoencoder (VAE) has been trained, it can be used to generate samples:

☐ that are of the same dimension as the input examples

☐ that are of the latent dimension

√ **It can generate samples at both of the dimensions listed above.**

☐ None of the above

3. Total for Question 3: 3

Consider the problem of multiclass classification of optical character recognition data of the form (x, y), where each data point, $x \in \mathbb{R}^{784}$ represents 784 pixels, each pixel can take 256 values, and the label, y, is a digit in $\{0, .., 9\}$.

An "output code" matrix can be used for solving a multiclass classification problem by solving multiple binary classification problems.

(a) (1 point) How many rows does **any** output code matrix for this problem have?

(a) **10**

(b) (1 point) Now consider an output code matrix for this problem, in which the columns are all possible one-vs-all problems. How many columns does this matrix have?

(b) **10**

(c) (1 point) How many zeros are there in the output code matrix in part b?

(c) **0 (none)**

4. (2 points) Total for Question 4: 2

Suppose that we use the following error correcting output code classification scheme to determine what language a document is written in.

	h_0	h_1	h_2	h_3
Dutch	1	1	0	0
German	0	0	1	0
Danish	0	1	1	1

If the result for a particular test document is $h_0 = 1$, $h_1 = 1$, $h_2 = 1$, and $h_3 = 0$ then the most probable language for the document is:

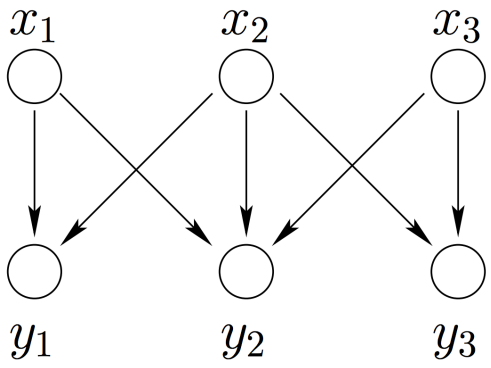
A. Dutch

B. German

C. Danish

4. _____

5. (6 points) Total for Question 5: 6



a) Bayesian network (directed)

Figure a) shows a Bayesian Network; nodes indicate random variables and edges indicate conditional probability distributions. Write the factorization of the joint probability distribution implied by this network.

Solution: $P(x_1)P(x_2)P(x_3)P(y_1|x_1,x_2)P(y_2|x_1,x_2,x_3)P(y_3|x_2,x_3)$

6. Total for Question 6: 11

Total for Question 6: 4 (bonus)

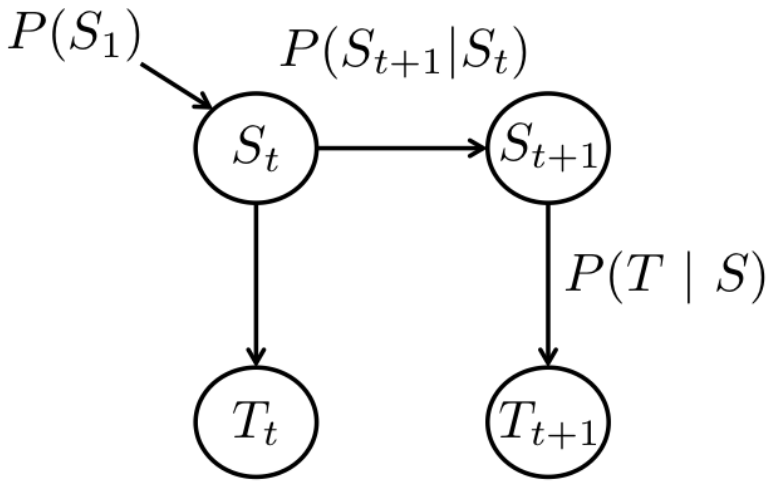


Figure 1: Hidden Markov Model.

We want to be able to figure out if a patient is sick by touching his forehead. His forehead can either be Hot (H), Cold (C) or Warm (W). Each day, the patient is either sick or not sick, and we will model his daily health with a first-order Hidden Markov Model. More formally, each day the patient is either in state $S = 1$ or $S = 0$. If the patient is Sick ($S = 1$), he has a Hot forehead with probably $\frac{2}{3}$, a Cold forehead with probability $\frac{1}{6}$ and a Warm forehead with probability $\frac{1}{6}$. On a day when he is Not Sick ($S = 0$), he has a Hot forehead with probability $\frac{1}{4}$, a Cold forehead with probability $\frac{1}{4}$ and a Warm forehead with probability $\frac{1}{2}$. If the patient is Sick today, then there's a $\frac{2}{3}$ chance he will be Sick tomorrow. If the patient is Not Sick today, then there's a $\frac{1}{2}$ chance he will not be sick tomorrow. The initialization probability of the patient being Sick is $\frac{1}{2}$.

(a) (7 points) Figure 1 shows a (homogeneous) first-order Hidden Markov Model. Label it by filling in the names of the appropriate random variables and probability distributions in the dotted boxes. As defined above, let S be a binary random variable for whether the patient is sick, and let T be a discrete random variable taking the values H, C, and W. The labels you need to apply to the graph are $P(T|S), P(S_{t+1}|S_t), T_t, S_{t+1}, P(S_1), T_{t+1}$, and S_t.

Solution: Figure has been labeled in solution.

(b) (4 points) Write out the state transition matrix for the hidden Markov chain.

Solution:

$$P(S_{t+1}|S_t) = \begin{Bmatrix} \frac{2}{3} & \frac{1}{3}; \frac{1}{2} & \frac{1}{2} \end{Bmatrix}$$

(c) (4 points (bonus)) Suppose that on the first day we feel the patient's forehead and it's Cold. What is the probability that the patient is Sick, given this observation?

Solution: We are being asked to solve for $P(S = 1|T = C)$. We can apply Bayes rule.

$$P(S = 1|T = C) = \frac{P(T = C|S = 1)P(S = 1)}{P(T = C)}$$

From above we have that on the first day (the beginning of time) $P(S = 1) = \frac{1}{2}$, therefore $P(S = 0) = P(N) = \frac{1}{2}$. And we are also given that $P(T = C|S = 1) = \frac{1}{6}$. It remains to solve for the denominator, which we can do as follows. We were also given that $P(T = C|S = 0) = \frac{1}{4}$. From this and the above we can compute $P(T = C) = P(T = C|S = 1)P(S = 1) + P(T = C|S = 0)P(S = 0) = \frac{1}{2}(\frac{1}{4} + \frac{1}{6}) = \frac{5}{24}$. We can now plug in for all the probabilities in Bayes rules, as follows:

$$P(S = 1|T = C) = \frac{P(T = C|S = 1)P(S = 1)}{P(T = C)} = \frac{\frac{1}{6} \cdot \frac{1}{2}}{\frac{5}{24}} = \frac{2}{5}$$

So the patient is Sick with probability $\frac{2}{5}$ and Not sick with probability $\frac{3}{5}$.

7. (4 points (bonus)) **Bonus Question** Total for Question 7: 4 (bonus)

Consider a data set containing just the following two labeled data points, $(x_1, +1), (x_2, -1)$, for any $x_1, x_2 \in \mathbb{R}^d$, where $x_1 \neq x_2$. We will use this data set to perform k-nearest-neighbor classification on other (unlabeled) data points, $x_t \in \mathbb{R}^d$, where $k = 1$, and distance is the L2 metric (Euclidean distance), $d(x_t, x_s) = \|x_t - x_s\|$.

What is the form of the resulting binary classifier ($f : \mathbb{R}^d \to \{-1, +1\}$)? You do not have to prove your answer, however be precise and state both the shape of the resulting decision boundary, and how to compute it (geometrically) from the two data points. (Hint: you may gain intuition by drawing a figure).

Solution: The resulting classifier is a linear separator, i.e. a hyperplane in $\mathbb{R}^d$. To find the separating hyperplane, take the chord connecting the two points, and find the midpoint. The hyperplane is through the midpoint, and normal to the chord. Positive classifications are in the half-space containing x_1.