Name:

ENGS 108 Fall 2017 Midterm
October 18, 2017
Instructors: George Cybenko and Saeed Hassanpour

Time allotted: 60 minutes. There are 20 problems. Each problem is worth 5 points. Points will be subtracted for incorrect answers if multiple choices are made but incorrect. The exam is open computer, open book and open notes. **No Internet or phone access is allowed. Please be mindful that some students may not have taken this exam yet, due to illness or other circumstance. Do not discuss the exam with anyone other than the teaching assistants and professor until the instructors tell you that you can.**

Dartmouth's Academic Honor Principle is to be observed during this exam.

1. Which machine learning evaluation metric is not appropriate for a skewed test set? (Circle the correct answer.)

   (a) Specificity
   (b) Accuracy
   (c) F1 Score
   (d) Recall

   Answer: b

2. You are using a Convolutional Neural Network (CNN) to determine whether an image is that of a dog, a cat, or a duck. These are the only possibilities. How many neurons should your final layer contain, and what final activation function should you use? (Circle all answers that are correct.)

   (a) 3, ReLU
   (b) 1, Sigmoid
   (c) 3. SoftMax
   (d) 1, SoftMax
   (e) 3, Sigmoid
   (f) 1, ReLU

   Answer: c

3. The real-valued function f has n inputs and is differentiable. When can the gradient of f, ∇f, be approximated as follows:

$$\nabla f(x) \approx \frac{f(x+d) - f(x)}{\delta}$$

   where d is is the n-dimensional vector $d = [\delta, \delta, \delta, ..., \delta]^T$? (Circle all answers that are correct.)

(a) Only when $n = 1$.

(b) Only when $\delta > 0$.

(c) Only when f is a sigmoidal activation function.

(d) Always.

(e) Never.

Answer: a

4. Which model is different from the others? (Circle all answers that are correct.)

(a) MaxEnt classifier

(b) Polytomous logistic regression

(c) Linear regression

(d) Conditional maximum entropy classifier

Answer: c

5. Identify *all* the *parametric* machine learning methods in the list below. (Circle all answers that are correct.)

(a) Support Vector Machine

(b) Shallow neural network

(c) Decision trees

(d) K-Nearest Neighbors

(e) K-means

(f) Logistic regression

Answer: b,f

6. The class of real-valued functions, $\mathscr{F}$, consists of indicator functions of pairs of intervals. That is,

$$\mathscr{F} = \{I_{[a,b]} + I_{[c,d]} - I_{[a,b]\cap[c,d]} \mid -\infty < a \leq b < \infty, -\infty < c \leq d < \infty\}.$$

Here $I_{[a,b]}(x) = 1$ if $a \leq x \leq b$ and $I_{[a,b]}(x) = 0$ otherwise. The Vapnik-Chervonenkis Dimension of $\mathscr{F}$ is: (Enter a number in (a) or circle (b).)

(a) Write your answer here: ______________.

(b) Uncomputable or undefined because $\mathscr{F}$ has an infinite number of members.

Answer: 4

7. Which statements are correct about using AIC and BIC in clustering? (Circle all answers that are correct.)

   (a) AIC penalizes the complexity of the model less strongly than BIC.

   (b) AIC penalizes the complexity of the model more strongly than BIC.

   (c) AIC penalizes the complexity of the model the same as BIC.

   (d) AIC is used in an unsupervised fashion; but, BIC is used in a supervised mode.

Answer: a

8. Given a bound, k, on the number of splits allowed in a decision tree, an optimal decision tree with no more than k splits can always be efficiently computed for a classification problem. (Here "optimal" means the total number of misclassified training samples is minimal and "efficiently" means using polynomial resources in terms of k, the number of features, and samples in the training set.) (Circle your answer.)

   (a) True

   (b) False

Answer: b

9. Linear Support Vector Machines (LSVM), Nonlinear Support Vector Machines (NSVM), autoencoders (AE), nonlinear coordinate transformations (NCT) and Principal Component Analysis (PCA) are techniques that can change the dimensionality of the features used in machine learning problems. Put a check into each box that is possible and typically used in that manner.

	LSVM	NSVM	AE	NCT	PCA
Transform low dimensional features to higher dimensional features					
Transform high dimensional features to lower dimensional features					

Answer:

	LSVM	NSVM	AE	NCT	PCA
Transform low dimensional features to higher dimensional features	X	X		X	
Transform high dimensional features to lower dimensional features			X	X	X

10. What is the simplest method to increase the generalizability of a language model? (Circle all answers that are correct.)

    (a) L1 regularization

    (b) Backoff Smoothing

    (c) Laplace smoothing

    (d) All of the above

    Answer: c

11. Stochastic gradient descent is a powerful optimization technique because: (Circle all answers that are correct.)

    (a) It can handle noise in the training data.

    (b) It can converge using only an appropriate approximation of the true gradient.

    (c) It is computationally more efficient and easier to implement in machine learning problems with large training sets.

    (d) It converges in fewer iterations that Newton's algorithm that uses the Hessian of the loss/error function.

    Answer: b,c

12. Which techniques/concepts are *not* used in backpropagation for deep neural networks, except possibly for validating correctness? (Circle all answers that are correct.)

    (a) Local gradients

    (b) Chain rule

    (c) Numerical gradient

    (d) Recursion

    Answer: c

13. What is the most common non-linear activation used in deep neural networks? (Circle all answers that are correct.)

    (a) Sigmoid

    (b) TanH

    (c) Convolution

    (d) ReLU

Answer: d

14. An “autoencoder” is: (Circle all answers that are correct.)

    (a) similar to an autoimmune system in biological systems.

    (b) similar to data compression in image and speech processing.

    (c) used solely in autonomous systems such as self-driving cars and aircraft.

    (d) useful for finding lower dimensional features in a machine learning problem.

Answer: b,d

15. The ReLU activation function: (Circle all answers that are correct.)

    (a) is differentiable everywhere.

    (b) is bounded.

    (c) has a bounded derivative.

    (d) is the only way to find low dimensional features in a machine learning problem.

Answer: c

16. Which deep learning architecture typically has the most number of layers? (Circle all answers that are correct.)

    (a) VGG

    (b) AlexNet

    (c) ResNet

    (d) GoogLeNet

Answer: c

17. The “one hot” method of encoding a feature is useful: (Circle all answers that are correct.)

    (a) for encoding categorical features.

    (b) because it controls the cooling temperature in simulated annealing searches for optimal parameters.

    (c) for converting real valued features into integers.

    (d) because it allows ReLU activations to be used in deep networks.

Answer: a

18. Assume your CNN has three layers:

    - Layer 1: 6, 5x5x3 filters with stride 1
    - Layer 2: 6, 5x5x6 filters with stride 1
    - Layer 3: 5, 8x8x6 filters with stride 2

    If you apply this network on an image on a 32x32 pixel image with 3 channels (RGB) and one pixel zero padding, what is the size of the output tensor?

    Write your answer here: ___________________________.

    Answer: 10x10x5

19. A 2-class classification problem involves a data set with supervised training data:

    $$\mathscr{D} = \{(f_i, c_i)|\ f_i \in R^n, c_i = 0 \ or \ c_i = 1, 1 \leq i \leq N\}.$$

    It is discovered that one of the features in all of the f_i is less than -10 when $c_i = 1$ but greater than 5 when $c_i = 0$. In this situation, the classification problem : (Circle all answers that are correct.)

    (a) is linearly separable.
    (b) can be easily solved using a quadratic SVM.
    (c) can be easily solved using a linear SVM.
    (d) can be easily solved with a decision tree but only if three or more splits are used.
    (e) is trivial using a simple one nearest neighbor classification approach.

    Answer: a,b,c

20. The computational complexity (that is, the number of operations) of using dynamic programing to solve the minimum edit distance between two strings of sizes n and m is proportional to: (Circle all answers that are correct.)

    (a) $n + m$
    (b) nm
    (c) $\log(nm)$
    (d) $nm\log(nm)$

    Answer: b

21. **(Extra credit.)** You are using a CNN to determine whether an image contains a dog, a cat, *and/or* a duck or none of these animals. These are the same classes as in question 2, but in a multi-label setting now. That is, zero, one, or more class-labels may be associated to each input example. How would you modify the last layer of your network from question 2 to model this problem? Please specify the number of units and the activation function.

    Write your answer below.

    Number of units: ______________

    Activation function: ____

    Answer: 3 units, each sigmoidal