CM146: Introduction to Machine Learning **Winter 2020**

Final Exam

March 20^{th}, 2020

- This is an open notes exam. You can refer to the class slides and your lecture notes.
- Everything you need in order to solve the problems is supplied in the body of this exam.
- Mark your answers on $8.5'' \times 11''$ blank sheets.
- This exam has a total of **11 pages** including the cover sheet.
- Your submission should have as many sheets as this pdf (including the cover sheet). For example, if questions 1 and 2 are on page 2, please mark your answer for question 1 and 2 on page 2 of your answer.
- Clearly write your UID and page number on each sheet of your submission.
- Scan each sheet, convert to a single PDF and submit to gradescope.
- You have **2 hours** to work on the exam (start time – 8:30 am, end time – 10:30am).
- You will have **30 minutes** to scan and upload.
- On the cover sheet:
  - Legibly write your name and UID.
  - Write and sign this statement: **I certify that I completed this exam entirely on my own without reference to any prohibited source materials and without consulting anyone.**

Good Luck!

Name and UID		/2
True/False		/20
Multiple choice		/40
HMM		/5
Clustering		/8
Total		/75

1 True or False (20 pts)

Choose either True or False for each of the following statements.

1. (2 pts) A decision tree learned with $MaxDepth = \infty$ and no pruning is always guaranteed to have zero training error.

   True False

   Solution: False.

2. (2 pts) You are given a dataset $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ with binary labels ($y_n \in \{-1, +1\}$) that is not linearly separable. You transform the features using a mapping function $\phi(\boldsymbol{x}) = \mathbf{W}\boldsymbol{x}$ for some matrix $\mathbf{W}$. Then a perceptron may be able to achieve zero training error on $\{(\phi(\boldsymbol{x}_n), y_n)\}_{n=1}^N$.

   True False

   Solution: False.

3. (2 pts) For any kernel function, you can always find the corresponding mapping function with finite output dimensions.

   True False

   Solution: False

4. (2 pts) As you increase the value of the regularization parameter λ, the optimal value of the weights for ridge regression approaches zero.

   True False

   Solution: True

5. (2 pts) For any constrained optimization problem, the optimal value of the dual problem is always less than or equal to the optimal value of the primal problem.

   True False

   Solution: True

6. (2 pts) For linearly separable data, the optimal solution of the hard-margin linear SVM is always a linear combination of support vectors.

True False

Solution: True

7. (2 pts) The log-likelihood of the data will never decrease through successive iterations of the expectation maximation (EM) algorithm.

True False

Solution: True

8. (2 pts) You and your friend want to detect clusters in the same dataset. Each of you runs the K-means algorithm on this dataset with the same value of K. It is possible that the clusters that each of you finds is different.

True False

Solution: True

9. (2 pts) Regardless of the size of the neural network, the backpropagation algorithm can always find the weights for the neural network that is a global minimizer of the empirical risk.

True False

Solution: False. It neednt since for a general multi-layer NN with non-linear threshold units, the function optimized by the backpropagation algorithm is not convex and has lots of local minimal points

10. (2 pts) Stochastic gradient descent performs less computation per parameter update than batch gradient descent.

True False

Solution: True

2 Multiple choice (40 pts)

MARK ALL CORRECT CHOICES (in some cases, there may be more than one)

1. (4 pts) Which of the following algorithms can achieve zero training error on XOR problem (note: "linear" implies no use of a non-linear feature transformation)?

   (a) AdaBoost

   (b) Logistic regression (no non-linear feature transformation)

   (c) Soft-margin linear SVM

   (d) Hard-margin linear SVM

   Solution: a

2. (4 pts) Assume you are fitting a linear regression model $h_{\boldsymbol{\theta}} = \boldsymbol{\theta}^T \boldsymbol{x}$ by minimizing the residual sum of squares cost function and you notice that the model is overfitting. What strategies can help reduce overfitting ?

   (a) Adding more features
   (b) Adding a regularization term to the cost function
   (c) Increasing the number of samples in the training data
   (d) Normalizing the features

   Solution: b,c

3. (4 pts) With N training samples, the computational complexity of nearest neighbors classifier to classify a new data point is (assuming the cost of computing the distance between a pair of points is constant):

   (a) O(1)
   (b) O(N)
   (c) O(N^2)
   (d) O($N \log N$)

   Solution: b

4. (4 pts) Given two training data points $\boldsymbol{x}_1 = (0, 1)^\top$ and $\boldsymbol{x}_2 = (1, 0)^\top$ with label $y_1 = 1$ and $y_2 = -1$. Consider a linear classifier $h(\boldsymbol{x}) = SIGN(\boldsymbol{w}^\top \boldsymbol{x} + b)$ with $\boldsymbol{w}^\top = (-1, 2)^\top$ and $b = -1$. Which of the following are true?

   (a) This classifier can achieve zero training error.

   (b) The distance from $\boldsymbol{x}_1$ to the hyper-plane $\boldsymbol{w}^\top \boldsymbol{x} + b = 0$ is $1/\sqrt{5}$.

   (c) The distance from $\boldsymbol{x}_2$ to the hyper-plane $\boldsymbol{w}^\top \boldsymbol{x} + b = 0$ is $1/\sqrt{5}$

   (d) There exists a linear classifier that can achieve margin $\gamma = 1$ on these two examples.

   Solution: a,b

5. (4 pts) For which of the following does normalizing your input features not change the predictions (normalizing here means that for each feature, we subtract out the mean and divide by the standard deviation) ?

   (a) Linear regression with l_2 regularization

   (b) Decision tree (learned with the ID3 algorithm)

   (c) Neural network

   (d) Soft-margin support vector machine

   Solution: b

6. (4 pts) In a soft-margin support vector machine, decreasing the slack penalty term C causes:

   (a) more overfitting.

   (b) a smaller margin.

   (c) less overfitting.

   (d) a larger margin.

   Solution: c,d

7. (4 pts) You perform an eigendecomposition of the covariance matrix of your data. The eigenvalues of the covariance matrix are $(10, 5, 0, 0)$. What is the minimum number of principal components we need to reconstruct the original data with no error?

   (a) 1

   (b) 2

   (c) 3

   (d) 4

   Solution: b

8. (4 pts) Consider an ensemble learning algorithm for binary classification that uses simple majority voting among 3 hypotheses. Suppose each of the hypotheses has error 0.1 in classifying a new instance and their errors are mutually independent. The expected error of the ensemble on the same instance is:

   (a) 0.1
   (b) $(0.1)^3$
   (c) $(0.1)^3 + 3 \times 0.1(0.9)^2$
   (d) $(0.1)^3 + 3 \times 0.9(0.1)^2$

   Solution: d

9. (4 pts) Suppose you are running a learning experiment on a new algorithm for binary classification. You have a data set consisting of 100 positive and 100 negative examples. You plan to use leave-one-out cross-validation and compare your algorithm to a baseline function: a simple majority function. What is the average cross-validation accuracy of the baseline?

   (a) 0.50
   (b) 1.00
   (c) 0.00
   (d) Not enough information

   Solution: c

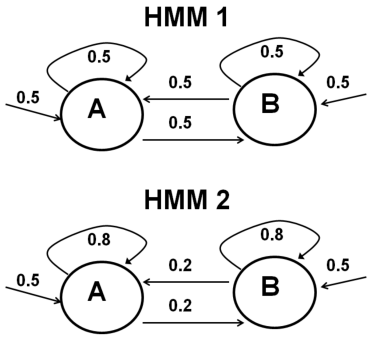
10. (4 pts) Which of the following holds for any valid kernel function $k(\cdot, \cdot)$ and for all $\boldsymbol{x}, \boldsymbol{y}$?

   (a) $k(\boldsymbol{x}, \boldsymbol{x}) \geq 0$
   (b) $k(\boldsymbol{x}, \boldsymbol{y}) \geq 0$
   (c) $k(\boldsymbol{x}, \boldsymbol{x}) + k(\boldsymbol{y}, \boldsymbol{y}) \geq 2k(\boldsymbol{x}, \boldsymbol{y})$
   (d) $k(\boldsymbol{x}, \boldsymbol{x}) + k(\boldsymbol{y}, \boldsymbol{y}) \geq -2k(\boldsymbol{x}, \boldsymbol{y})$

   Solution: a,c,d

3 HMM [5 points]

The figure below presents two HMMs. Hidden states are represented by circles (so each HMM has two hidden states) and transitions by edges. Transition probabilities and initial probabilities are listed next to the relevant edges. The initial probability is represented by leftmost and rightmost edges into each circle. In both HMMs, emissions are deterministic and listed inside the states, *e.g.*, the state represented by the circle on the left always emits symbol A. For example, in HMM 1 we have a probability of 0.5 to start with the state that emits A and a probability of 0.5 to transition to the state that emits B if we are now in the state that emits A. In the questions below, $O_3 = A$ means that the 3^{rd} symbol emitted by the HMM is A.



1. (2 pts) What is $P(O_1 = A, O_2 = A, O_3 = A)$ for HMM1?

   (a) 0.5^{303}

   (b) $3 * 0.5^3$

   (c) 0.5^3

   (d) $6 * 0.5^3$

   Solution: 0.5^3

2. (2 pts) What is $P(O_1 = A, O_2 = A, O_3 = A)$ for HMM2?

   (a) $0.5 * 0.8^2$

   (b) $(0.5 * 0.8 + 0.5 * 0.2)^3$

   (c) 0.8^3

   (d) $(0.5 * 0.8)^3$

   Solution: $0.5 * 0.8^2$

3. (1 pts) Let P_1 be: $P_1 = P(O_1 = A, O_2 = B, O_3 = A, O_4 = B)$ for HMM1 and let P_2 be: $P_2 = P(O_1 = A, O_2 = B, O_3 = A, O_4 = B)$ for HMM2. Choose the correct answer from the choices below and briefly explain.

   (a) $P_2 > P_1$

   (b) $P_1 > P_2$

(c) $P_1 = P_2$

(d) Impossible to tell the relationship between the two probabilities.

Solution: Here P_1 evaluates to 0.5^4 while P_2 is $0.50.2^4$. So clearly $P_1 > P_2$ or option B

4 Clustering [8 points]

Recall that in $K-$means clustering we attempt to find K cluster centroids $\boldsymbol{\mu}_k \in \mathbb{R}^d, k \in \{1, \ldots, K\}$ such that the total distance between each datapoint and the nearest cluster centroid is minimized. In other words, we attempt to solve:

$$\min_{\{\boldsymbol{\mu}_k\},\{r_{nk}\}} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2, \tag{1}$$

where N is the number of data points, r_{nk} is a binary variable that is 1 if sample n is assigned to cluster k and zero otherwise.

1. (3 pts) Instead of holding the number of clusters K fixed, one can think of minimizing (1) over all of K and $\boldsymbol{\mu}_k$ and r_{nk}. Show that this is a bad idea. Specifically, what is the minimum possible value of (1) ? What values of K and $\boldsymbol{\mu}_k$ result in this value ?

   Solution: The minimum possible value of the objective is then zero when we set $K = N$ and $\boldsymbol{\mu}_k$ to each data point.

   [1pt each for the minimum, and for the value of K and μ]

2. (2 pts) Recall that in one of the steps of the K-means algorithm, for a fixed assignment of each sample to one of the K clusters (r_{nk}), we compute the new cluster centroids $\boldsymbol{\mu}_k$ by minimizing $\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|_2^2$.

   Compute the gradient of the above objective function with respect to $\boldsymbol{\mu}_k$. For reference, here is a useful identity:

$$f(\boldsymbol{x}) = \|\boldsymbol{x}\|_2^2 \qquad \nabla f(\boldsymbol{x}) = 2\boldsymbol{x}$$

   Solution:

$$\begin{aligned}
\nabla_{\boldsymbol{\mu}_k}\left(\sum_{n=1}^{N}\sum_{l=1}^{K} r_{nl}\|\boldsymbol{x}_n - \boldsymbol{\mu}_l\|^2\right) &= \sum_{n=1}^{N}\nabla_{\boldsymbol{\mu}_k}\left(\sum_{l=1}^{K} r_{nl}\|\boldsymbol{x}_n - \boldsymbol{\mu}_l\|^2\right) \\
&= \sum_{n=1}^{N}\nabla_{\boldsymbol{\mu}_k} r_{nk}\|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2 \\
&= \sum_{n=1}^{N} r_{nk}\nabla_{\boldsymbol{\mu}_k}\|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2 \\
&= \sum_{n=1}^{N} r_{nk}\left(-2(\boldsymbol{x}_n - \boldsymbol{\mu}_k)\right) \\
&= -2\left(\sum_{n=1}^{N} r_{nk}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)\right) \\
&= -2\left(\sum_{n=1}^{N} r_{nk}\boldsymbol{x}_n - \sum_{n=1}^{N} r_{nk}\boldsymbol{\mu}_k\right)
\end{aligned}$$

   [2pt for gradient]

3. (1 pts) Set the gradient to zero and solve for $\boldsymbol{\mu}_k$. Show that the optimal ${\boldsymbol{\mu}_k}^*$ corresponds to the mean of the samples assigned to cluster k.

   Solution: Setting the gradient to zero and solving for $\boldsymbol{\mu}_k$:

$${\boldsymbol{\mu}_k}^* = \frac{\sum_{n=1}^{N} r_{nk}\boldsymbol{x}_n}{\sum_{n=1}^{N} r_{nk}}$$

   r_nk is one for all samples assigned to cluster k and zero otherwise. Thus, the numerator is the sum of the featueres for all samples assigned to cluster k and the denominator is the number of samples assigned to cluster k.

   [1pt for answer]

4. (2 pts) We now consider clustering 1D data using K-means. We assume the number of clusters $K = 2$. You are given four instances: $(x_1, x_2, x_3, x_4) = (1, 10, 20, 9)$ where each $x_n \in \mathbb{R}, n \in \{1, 2, 3, 4\}$. The current estimates of r_{nk} is represented by the following matrix:

$$\boldsymbol{R} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Here entry (n, k) of this matrix is r_{nk}.

Show the update for the cluster centroids μ_1, μ_2 (you do not need to simplify your answer).
Solution:

$$\mu_1 = \frac{1+9}{2}$$
$$\mu_2 = \frac{10+20}{2}$$

[1pt for each of centroids]