UNIVERSITY OF PENNSYLVANIA

CIS 520: Machine Learning

Final Exam, Fall 2017

Exam policy: This exam allows two one-page, two-sided cheat sheets (i.e. 4 sides). No other materials are allowed.

Time: 2 hours.

Be sure to write your name and Penn student ID (the 8 bigger digits on your ID card) on the bubble form and fill in the associated bubbles *in pencil*.

If you are taking this as a WPE, then enter *only* your WPE number and fill in the associated bubbles, and do not write your name.

For all questions, select exactly one answer and fill in the corresponding circle on the bubble form. If you think a question is ambiguous, mark what you think is the best answer. The questions seek to test your general understanding; they are not intentionally "trick questions." As always, we will consider written regrade requests if your interpretation of a question differed from what we intended. *We will only grade the bubbled answer key.*

The exam has 57 questions, totalling 98 points. These are distributed as follows:

Problems 1–31 are worth 1 point each (total 31 points).
Problems 32–42 are worth 2 points each (total 22 points).
Problems 43–57 are worth 3 points each (total 45 points).

Name: _____

1. [1 points] *True or False?* Iterating between the E-step and M-step of EM algorithms always converges to a local optimum of the likelihood.

   (a) True
   (b) False

   ★ **SOLUTION:** A

2. [1 points] *True or False?* Lasso selects a subset (not necessarily strict) of the original features.

   (a) True
   (b) False

   ★ **SOLUTION:** A

3. [1 points] *True or False?* The features selected by PCA are linear combinations of the original features.

   (a) True
   (b) False

   ★ **SOLUTION:** A

4. [1 points] *True or False?* The solution to principal component analysis (PCA) can always be found using singular value decomposition (SVD).

   (a) True
   (b) False

   ★ **SOLUTION:** A

5. [1 points] *True or False?* PCA can be formulated as an optimization problem that finds the (orthogonal) directions of maximum covariance of a set of observations X.

   (a) True
   (b) False

   ★ **SOLUTION:** A

6. [1 points] *True or False?* Principal Components Regression (PCR) generally yields models in which some of the original features do not affect the prediction.

   (a) True
   (b) False

★ **SOLUTION:** B

7. [1 points] *True or False?* The eigenvectors of AA^T and A^TA are the same for any matrix A.

   (a) True

   (b) False

★ **SOLUTION:** B

8. [1 points] Suppose you learn a model for binary classification using an algorithm $\mathcal{A}$. After learning this model you observe an additional instance-label pair $(\mathbf{x}, y)$, but you suspect that one of the components x_i has been corrupted with noise. To check this, you want to infer the probability $\Pr(x_i|y)$ using your model. Which of the following algorithms $\mathcal{A}$ would work best to find this probability?

   (a) Naive Bayes

   (b) Logistic Regression

★ **SOLUTION:** A

9. [1 points] *True or False?* K-means clustering can be kernelized using the kernel trick.

   (a) True

   (b) False

   ★ **SOLUTION:** A

10. [1 points] *True or False?* PCA does nonlinear orthogonal transformation of data into a lower dimensional space.

    (a) True

    (b) False

    ★ **SOLUTION:** B

11. [1 points] Consider data with n samples $D = \{\mathbf{x}^1, \ldots, \mathbf{x}^n\}$ with $\mathbf{x}^i \in \mathbb{R}^p$ with $n \gg p \gg k$, where k is the number of components to be kept. If computation speed is **not** as issue, it is advisable to do PCA by

    (a) computing the covariance matrix of X.

    (b) computing the SVD of X.

    (c) either (a) or (b). It won't make any difference.

    ★ **SOLUTION:** C

12. [1 points] *True or False?* PCA is a type of linear autoencoder.

    (a) True

    (b) False

    ★ **SOLUTION:** A

13. [1 points] *True or False?* Consider data with n samples $\mathbf{x}^1, \ldots, \mathbf{x}^n$ with $\mathbf{x}^i \in \mathbb{R}^p$. Given the number of principal components, the covariance matrix given by

$$\Sigma = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}^i - \overline{\mathbf{x}})(\mathbf{x}^i - \overline{\mathbf{x}})^\top$$

    is **sufficient** to compute the reconstruction accuracy after applying PCA to the data.

    (a) True

    (b) False

★ **SOLUTION:** A

14. [1 points] *True or False?* Ordinary least squares (linear regression) can be formulated either as minimizing an L_2 loss function or as maximizing a likelihood function.

    (a) True

    (b) False

★ **SOLUTION:** A

15. [1 points] *True or False?* We observe data sampled from the model $y \sim \mathcal{N}(w^T x, \sigma^2)$. Since w is unknown, we would like to estimate it using linear regression. Adding priors to the parameters w, i.e. $w_j \sim \mathcal{N}(0, \lambda^2)$ is equivalent to adding an L_2 penalty on the parameters in the objective function defined by log-likelihood.

    (a) True

    (b) False

★ **SOLUTION:** A

16. [1 points] *True or False?* If some given data is not linearly separable, we can design a linear classifier that separates the data in a higher dimension as long as no point x appears twice with different y values.

    (a) True
    (b) False

    ★ **SOLUTION:** A

17. [1 points] *True or False?* L_1-penalized linear regression is unbiased.

    (a) True
    (b) False

    ★ **SOLUTION:** B

18. [1 points] *True or False?* Stepwise regression finds the global optimum, minimizing its loss function (squared error plus the usual L_0 penalty).

    (a) True
    (b) False

    ★ **SOLUTION:** B

19. [1 points] *True or False?* Inverting X^TX for data sets with many more observations than features ($n \gg p$, where X is $n \times p$) is, in general, significantly slower than computing X^TX.

    (a) True
    (b) False

    ★ **SOLUTION:** B

20. [1 points] *True or False?* It is difficult to implement 'data-parallel' (e.g. map-reduce) algorithms for linear regression, as simple data-parallel methods come at a large cost in accuracy.

    (a) True
    (b) False

    ★ **SOLUTION:** B

21. [1 points] *True or False?* The complexity of contemporary deep learning systems for vision (as measured, for example, by the number of bits of information to specify them) is coming close to that of the human visual cortex.

    (a) True
    (b) False

★ **SOLUTION:** B

22. [1 points] *True or False?* Progress in machine learning has continued at a rapid pace over the past decade in spite of the fact that the cost of computing (as, for example, measured by the number of multiplies per second that can be done for a dollar) is no longer decreasing by a factor of two roughly every 18 months.

    (a) True

    (b) False

★ **SOLUTION:** B

23. [1 points] *True or False?* When doing machine learning on large data sets, it is good practice to test which algorithms work best on a small subset of the data before running the best model on the whole data set, since the same algorithms that work best on small data sets almost always also work best on big sets of the same data.

    (a) True
    (b) False

    ★ **SOLUTION:** B

24. [1 points] *True or False?* In the video assigned in class, the speaker (Killian Weinberger) argued that deep learning systems for image recognition approximately map images to a manifold in which certain directions correspond to meaningful directions such as faces appearing younger/older or more male/female.

    (a) True
    (b) False

    ★ **SOLUTION:** A

25. [1 points] *True or False?* The elastic net tends to select more features than well-optimized L_0 penalty methods.

    (a) True
    (b) False

    ★ **SOLUTION:** A

26. [1 points] *True or False?* PCA, when the data are mean centered, but not standardized, is scale invariant.

    (a) True
    (b) False

    ★ **SOLUTION:** B

27. [1 points] *True or False?* Stepwise regression (linear regression with an L_0 penalty) is scale invariant.

    (a) True
    (b) False

★ **SOLUTION:** A

28. [1 points] *True or False?* K-means clustering (using standard Euclidean distance) is scale invariant.

    (a) True
    (b) False

★ **SOLUTION:** B

29. [1 points] *True or False?* MLE is more likely to overfit than MAP since MAP tends to shrink parameters.

    (a) True
    (b) False

★ **SOLUTION:** A

30. [1 points] *True or False?* Consider a "true" distribution p given by

$$p(A) = 0.5,\ p(B) = 0.25,\ p(C) = 0.25$$

and an "approximating" distribution q given by

$$q(A) = 0.5,\ q(B) = 0.5,\ q(C) = 0\,.$$

The KL divergence $\mathrm{KL}(p\|q)$ is $\frac{1}{2}\log(2)$.

(a) True
(b) False

★ **SOLUTION:** B

31. [1 points] *True or False?* When you do principal components analysis on an $n * p$ observation matrix and keep k components, you get dimensions as follows: loadings: $n * k$, scores: $k * p$

(a) True
(b) False

★ **SOLUTION:** B

32. [2 points] Duplicating a feature in linear regression

(a) Does not reduce the L2-Penalized Residual Sum of Squares.
(b) Does not reduce the Residual Sum of Squares (RSS).
(c) Can reduce the L1-Penalized Residual Sum of Squares (RSS).
(d) None of the above

★ **SOLUTION:** B

33. [2 points] *True or False?* Given a set of data points $x_1, \cdots, x_n$, the K-means objective for finding cluster centers $\mu_1, \ldots, \mu_K$, and cluster assignments r_{ik}'s is as follows:

$$J(\mu, r) = \sum_{i=1}^{n}\sum_{k=1}^{K} r_{ik}\|\mu_k - x_i\|_2^2\,.$$

Minimizing this objective is a convex optimization problem.

(a) True
(b) False

★ **SOLUTION:** B

34. [2 points] We apply the Expectation Maximization algorithm to $f(D, Z, \theta)$ where D denotes the data, Z denotes the hidden variables and θ the variables we seek to optimize. Which of the following are correct?

    (a) EM will always return the same solution which may not be optimal

    (b) EM will always return the same solution which must be optimal

    (c) The solution depends on the initialization

★ **SOLUTION:** C

35. [2 points] In a convolutional neural net with an image of size 5x5x3 (where 3 is red/green/blue), we pad with a single zero all around the image and then use 4 local receptive fields ('filters') of size 3x3x3 and a stride of size 2. The outputs of these local receptive fields are sent to a single output. Assuming that there are no bias terms in this model, the total number of parameters (degrees of freedom) in the network is:

    (a) 3*3*4 + 4
    (b) 3*3*3*4 + 4
    (c) 3*3*3*4 + 4*9
    (d) 7*7*7*4 + 4
    (e) none of the above

    ★ **SOLUTION:** C

36. [2 points] When doing linear regression with $n = 1,000$ observations and $p = 100,000$ features, if one expects around 5 or 10 features to enter the model, the best penalty to use is

    (a) AIC penalty
    (b) BIC penalty
    (c) RIC penalty
    (d) This problem is hopeless – you couldn't possibly find a model that reliably beats just using a constant.

    ★ **SOLUTION:** C

37. [2 points] If you know the noise in measuring each observation y_i is $N(0, \sigma_i^2)$, then to obtain an optimal model using linear regression, during training you should weight each observation

    (a) by its variance, σ_i^2
    (b) by its standard deviation σ_i
    (c) equally
    (d) by its inverse standard deviation, σ_i^{-1}
    (e) by its inverse variance, σ_i^{-2}

    ★ **SOLUTION:** E

38. [2 points] The AdaBoost algorithm can be viewed as minimizing the:

    (a) exponential loss
    (b) logistic loss
    (c) hinge loss
    (d) squared loss
    (e) none of the above

    ★ **SOLUTION:** A

39. [2 points] In a supervised learning problem, the true quantity we really want to minimize is:

    (a) the training error
    (b) the test error
    (c) the generalization error
    (d) the cross-validation error
    (e) none of the above

    ★ **SOLUTION:** C

40. [2 points] Suppose you train two binary classifiers, h_1 and h_2, on the same training data, from two function classes $\mathcal{H}_1$ and $\mathcal{H}_2$ with $\text{VCdim}(\mathcal{H}_1) < \text{VCdim}(\mathcal{H}_2)$. Suppose h_1 and h_2 have the same training error. Then the VC-dimension based generalization error bound for h_1 is:

    (a) smaller than that for h_2
    (b) larger than that for h_2
    (c) equal to that for h_2
    (d) We can't say anything about the relationship between the two

    ★ **SOLUTION:** A

41. [2 points] Which of the following provides a discriminative learning algorithm/model for structured prediction?

    (a) Sum-product algorithm
    (b) Conditional random fields
    (c) Viterbi algorithm
    (d) Baum-Welch algorithm
    (e) None of the above

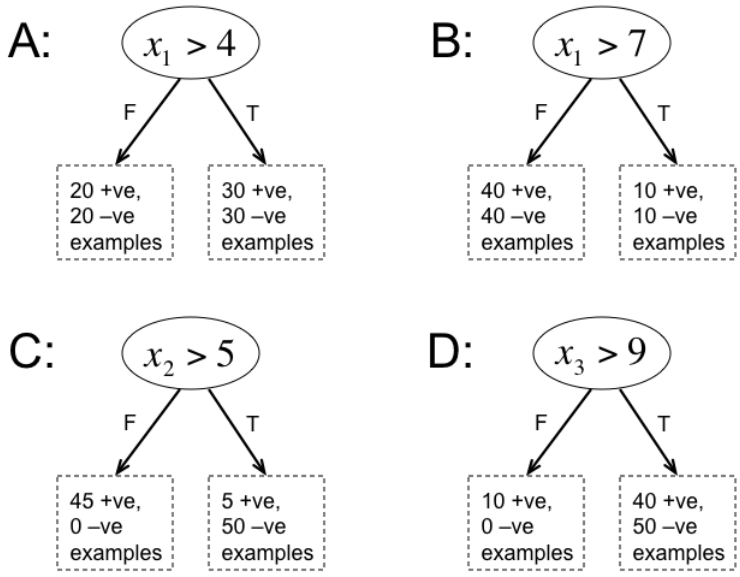
    ★ **SOLUTION:** B

42. [2 points] In reinforcement learning, a deterministic policy is

    (a) a mapping from states to states
    (b) a mapping from state-action pairs to states
    (c) a mapping from actions to states
    (d) a mapping from states to actions
    (e) none of the above

    ★ **SOLUTION:** D

43. [3 points] Suppose you have a binary classification problem with 3-dimensional feature vectors $\mathbf{x} \in \mathbb{R}^3$. You are given 50 positive and 50 negative training examples, and want to build a decision tree classifier. Consider 4 possible splits at the root node:



Which of the above splits gives the highest information gain?

(a) A
(b) B
(c) C
(d) D

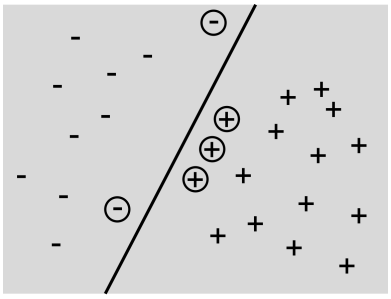
★ **SOLUTION:** C

44. [3 points] Consider modeling observations with 2 features using a Gaussian mixture model with 3 mixture components, where each component can have a different (full) covariance matrix. How many parameters are needed?

(a) 12
(b) 14
(c) 15
(d) 17
(e) none of the above

★ **SOLUTION:** D (17 = 3*2 (means) + 3*3 (covariance matrices) + 2 (mixing coefficients))

45. [3 points] Consider a binary classification problem in a 2-dimensional instance space $\mathcal{X} = \mathbb{R}^2$. You are given a linearly separable training set. You run the hard-margin SVM algorithm and obtain the separating hyperplane below (support vectors are circled):



What is the smallest number of data points that would have to be removed from the training set in order for the SVM solution to change?

(a) 1
(b) 2
(c) 3
(d) 4
(e) None of the above

★ **SOLUTION:** A

46. [3 points] Consider a binary classification problem in a d-dimensional instance space $\mathcal{X} = \mathbb{R}^d$. Your friend has a training set containing m labeled examples. She computes two $m \times m$ kernel matrices, K_1 and K_2, and gives them to you. The first matrix K_1 is obtained by applying an RBF kernel $K(\mathbf{x}, \mathbf{x}') = e^{-\|\mathbf{x}-\mathbf{x}'\|^2}$ to the original d-dimensional data points. The second matrix K_2 is obtained by selecting $r < d$ features and applying the RBF kernel (with the same width parameter as before) to the reduced r-dimensional data points. You are given these two kernel matrices and are asked to train an SVM classifier in each case. The training time for K_2 will be:

(a) smaller than that for K_1
(b) greater than that for K_1
(c) roughly similar to that for K_1

★ **SOLUTION:** C

47. [3 points] Consider running the perceptron algorithm for an online binary classification task. Recall that on each round t, the algorithm receives an instance $\mathbf{x}_t$ and uses the current weight vector $\mathbf{w}_t$ to predict $\widehat{y}_t = \text{sign}(\mathbf{w}_t^\top \mathbf{x}_t)$; it then receives the true label y_t, and if it made a mistake, it updates the weight vector as

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_t \, \mathbf{x}_t \, .$$

Suppose that on a particular round t, the algorithm predicts $\widehat{y}_t = -1$ and receives the true label $y_t = +1$. Assume $\mathbf{x}_t \neq \mathbf{0}$. In this case, after updating the weight vector, it is guaranteed that:

(a) $\mathbf{w}_{t+1}^\top \mathbf{x}_t > \mathbf{w}_t^\top \mathbf{x}_t$
(b) $\mathbf{w}_{t+1}^\top \mathbf{x}_t < \mathbf{w}_t^\top \mathbf{x}_t$
(c) $\mathbf{w}_{t+1}^\top \mathbf{x}_t > 0$
(d) $\mathbf{w}_{t+1}^\top \mathbf{x}_t < 0$
(e) none or more than one of the above

★ **SOLUTION:** A

48. [3 points] Consider running the AdaBoost algorithm for a binary classification problem in which you are given a small training set of 5 examples, $\{(x_i, y_i)\}_{i=1}^5$. In the first round, all 5 examples have equal weight, $D_1(i) = 1/5$. Suppose that the true labels and the predictions made by the weak classifier h_1 learned in the first round are as follows:

i	y_i	$h_1(x_i)$
1	-1	-1
2	-1	$+1$
3	$+1$	-1
4	$+1$	$+1$
5	$+1$	$+1$

In the second round, how many of the 5 examples will receive a higher weight than they had in the first round?

(a) 1
(b) 2
(c) 3
(d) 4
(e) none of the above

★ **SOLUTION:** B

49. [3 points] Consider a binary classification problem with the following loss function:

		$\widehat{y}$	
		-1	$+1$
y	-1	0	0.8
	$+1$	0.2	0

For a particular instance x, your class probability estimation (CPE) model $\widehat{\eta}$ predicts the probability of a positive label to be $\widehat{\eta}(x) = 0.75$. To minimize expected loss, the predicted label $\widehat{y}$ for this instance should be:

(a) $+1$
(b) -1
(c) Both are equally good

★ **SOLUTION:** B

50. [3 points] Consider a binary classification problem in which the label $+1$ is rare. You have learned a binary classifier $h : \mathcal{X} \to \{\pm 1\}$ from some training data. On a test set of 100 data points, the classifier's predictions, as well as the true labels, are as follows:

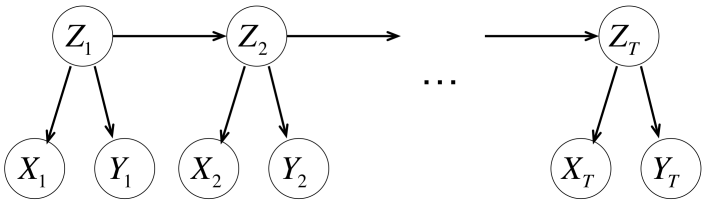
		$h(x)$	
		-1	$+1$
y	-1	75	15
	$+1$	3	7

What is the true positive rate (TPR) of the classifier h on the above test set?

(a) 3/10
(b) 7/10
(c) 3/78
(d) 7/22
(e) None of the above

★ **SOLUTION:** B

51. [3 points] Consider a variant of a hidden Markov model in which each hidden state Z_t generates two conditionally independent observations X_t, Y_t:



Suppose each hidden state Z_t takes one of K possible values, each observation X_t takes one of M_1 possible values, and each observation Y_t takes one of M_2 possible values. Assume the model is homogeneous, so that transition and emission probabilities are the same for all t. What is the total number of parameters in this model? Choose the tightest expression below.

(a) $O(K^2 M_1 M_2)$
(b) $O(K^2(M_1 + M_2))$
(c) $O(K^2 + KM_1M_2)$
(d) $O(K^2 + K(M_1 + M_2))$
(e) None of the above

★ **SOLUTION:** D

52. [3 points] Consider learning a (homogeneous) hidden Markov model for a part-of-speech tagging task, where observations X_t are words and hidden states Z_t are parts of speech such as 'noun' (N), 'verb' (V), 'determiner' (D), etc. You are given labeled training data consisting of the following two sentences with corresponding parts of speech:

N	V	N
Mary	likes	mountains

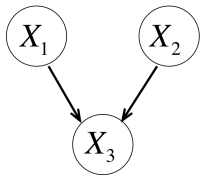
D	N	V	D	N
The	dog	ate	the	candy

What is the maximum likelihood estimate of the transition probability $A_{\mathsf{V},\mathsf{N}} = P(Z_{t+1} = \mathsf{N} \mid Z_t = \mathsf{V})$?

(a) 1
(b) 1/3
(c) 1/4
(d) 0
(e) None of the above

★ **SOLUTION:** E (Answer: 1/2)

53. [3 points] Consider three random variables X_1, X_2, X_3, each of which takes one of 2 possible values. Suppose their joint probability distribution is known to factor according to the Bayesian network structure below:

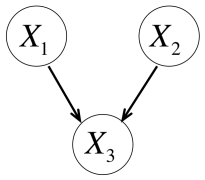


Given this information, how many parameters are needed to specify the joint probability distribution?

(a) 7
(b) 6
(c) 5
(d) 4
(e) None of the above

★ **SOLUTION:** B

54. [3 points] Consider a probability distribution that factors according to the Bayesian network structure below:

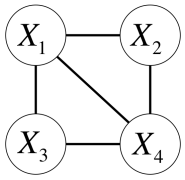


Which of the following (conditional) independence statements *must* be true?

(a) $X_1 \perp\!\!\!\perp X_3 \,|\, X_2$
(b) $X_1 \perp\!\!\!\perp X_2 \,|\, X_3$
(c) $X_1 \perp\!\!\!\perp X_3$
(d) $X_1 \perp\!\!\!\perp X_2$
(e) None of the above

★ **SOLUTION:** D

55. [3 points] Consider a probability distribution that factors according to the Markov network structure below:

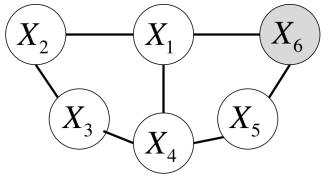


Which of the following (conditional) independence statements *must* be true?

(a) $X_2 \perp\!\!\!\perp X_3 \,|\, X_1$
(b) $X_1 \perp\!\!\!\perp X_4 \,|\, \{X_2, X_3\}$
(c) $X_2 \perp\!\!\!\perp X_3 \,|\, \{X_1, X_4\}$
(d) $X_2 \perp\!\!\!\perp X_3$
(e) None of the above

★ **SOLUTION:** C

56. [3 points] Consider 6 random variables $X_1, \ldots, X_6$, each of which takes one of K possible values. You are given their joint probability distribution, which factors according to the Markov network structure below; you are also told that X_6 takes the value $\overline{x}_6$.



You are asked to find the posterior probability distribution $p(x_1 \,|\, \overline{x}_6)$. You decide to use the variable elimination algorithm and eliminate variables in the order $(5, 4, 3, 2)$. As a function of K, how many computations will you need? Choose the tightest expression below.

(a) $O(K^2)$
(b) $O(K^3)$
(c) $O(K^4)$
(d) $O(K^5)$
(e) None of the above

★ **SOLUTION:** B

57. [3 points] Consider an active learning setup for binary classification with labels $\{\pm 1\}$ and 0-1 loss. You are given a small labeled training set, from which you learn a logistic regression model. You are also given four more unlabeled data points, $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, and $\mathbf{x}_4$, and are allowed to query the label of one of these. Your logistic regression model predicts the probabilities of each of these instances having label $+1$ as follows:

$$\widehat{\eta}(\mathbf{x}_1) = 0.30\,; \quad \widehat{\eta}(\mathbf{x}_2) = 0.40\,; \quad \widehat{\eta}(\mathbf{x}_3) = 0.55\,; \quad \widehat{\eta}(\mathbf{x}_4) = 0.75\,.$$

If you use an uncertainty sampling approach, which of the above instances would be chosen to query a label for?

(a) $\mathbf{x}_1$
(b) $\mathbf{x}_2$
(c) $\mathbf{x}_3$
(d) $\mathbf{x}_4$
(e) None of the above

★ **SOLUTION:** C