UNIVERSITY OF PENNSYLVANIA

CIS 520: Machine Learning
Final Exam, 2019

Exam policy: This exam allows one one-page, two-sided cheat sheet; No other materials.

Time: 120 minutes. Be sure to write your name and Penn student ID (the 8 bigger digits on your ID card) on the answer form and fill in the associated bubbles *in pencil*.

If you think a question is ambiguous, mark what you think is the best answer. As always, we will consider written regrade requests if your interpretation of a question differed from what we intended. *We will only grade the answer forms.*

For the "TRUE or FALSE" questions, note that "TRUE" is (a) *and "FALSE" is* (b)*. For the multiple choice questions, select exactly one answer.*

Questions follow the course convention that n or N represents the number of observations and p the number of features in the model.

There are **68** questions with a total of **90** points.

1. [2 points] For very large training data sets, which of the following will usually have the lowest training time?

   (a) logistic regression
   (b) neural nets
   (c) K-nearest neighbors
   (d) random forests
   (e) a linear SVM

   ★ **SOLUTION:** C

2. [2 points] For very large training data sets, which of the following will usually produce the smallest models (requiring the fewest parameters)?

   (a) logistic regression
   (b) neural nets
   (c) K-nearest neighbors
   (d) random forests
   (e) a linear SVM

   ★ **SOLUTION:** * A, C or E Logistic regression or SVM produces the smallest models, but k-NN has the fewest parameters (none).

3. [1 points] *True or False?* We train a decision tree model on a dataset that is not feature scaled. We standardize features on the test set (set all to mean zero, standard deviation 1). We expect the model to accurately predict on the feature-scaled test set because decision trees are scale-invariant.

   ★ **SOLUTION:** False

4. [2 points] A K-NN classifier will give accuracies most similar to

   (a) logistic regression
   (b) decision trees

(c) k-means (with each cluster labeled with the majority class of the items in it)

(d) neural networks

(e) Naive Bayes

★ **SOLUTION:** C

5. [1 points] *True or False?* The KL-divergence $D_{KL}(p||q)$ measures how different an approximate distribution p is from a "true" (or at least more accurate) distribution q.

★ **SOLUTION:** False

6. [1 points] *True or False?* The distribution $P(A) = 1/2, P(B) = 1/2$ has higher entropy than the distribution $P(A) = 1/3, P(B) = 1/3, P(C) = 1/3$

★ **SOLUTION:** False

7. [1 points] *True or False?* L2 loss is more robust to outliers than L1 loss.

★ **SOLUTION:** False

8. [2 points] Suppose you want to analyze a data set containing gene expression data for patients. For each patient, there is a feature vector containing expression levels of 20,000 genes, together with a label indicating whether the patient developed a certain disease. Your goal is to learn a model which, given a new patient (also represented by a similar 20,000-dimensional vector), can estimate the probability of this patient developing the disease. You expect the eventual model will depend on only a small number of the genes. Which of the following methods would be most suitable?

(a) Logistic regression with L2 regularization

(b) Linear regression with L2 regularization

(c) Logistic regression with L0 regularization

(d) Linear regression with L0 regularization

(e) Linear support vector machine

★ **SOLUTION:** C

9. [1 points] *True or False?* Naive Bayes is generally preferable to logistic regression if there are very little data (compared to the number of parameters), while logistic regression gives more accurate models in the limit of large training sets.

★ **SOLUTION:** True

10. [1 points] *True or False?* Generative Adversarial Networks (GANs) often converge to poor solutions (local optima) because the Generator tends to converge much more quickly than the Discriminator.

★ **SOLUTION:** False

For the next 3 questions, assume you have classification data with classes Y being +1 or -1 and features x_j also being +1 or -1 for $j \in 1, ..., p$.

In an attempt to turbocharge your classifier, you duplicate each feature, so now each example has $2p$ features, with $x_{p+j} = x_j$ for $j \in 1, ..., p$. The following questions compare the original feature set with the doubled one. You may assume that in the case of ties, class +1 is always chosen. Assume that there are equal numbers of training examples in each class.

11. [1 points] For a Naive Bayes classifier:

    (a) The test accuracy will usually be higher with the original features.

    (b) The test accuracy will usually be higher with the doubled features.

    (c) The test accuracy will be the same with either feature set.

    ★ **SOLUTION:** * C This is tricky, but since we assumed equal class

    probabilities, doubling the features squares the P(x$_j$|$class$)$terms, so does n't change the class predict$

12. [1 points] For a Naive Bayes classifier:

    (a) On a given training instance, the conditional probability $P(Y|x_1, ...)$ on a training instance will be more extreme (i.e. closer to 0 or 1) with the original features.

    (b) On a given training instance, the conditional probability $P(Y|x_1, ...)$ on a training instance will be more extreme (i.e. closer to 0 or 1) with the doubled features.

    (c) On a given training instance, the conditional probability $P(Y|x_1, ...)$ on a training instance will be the same with either feature set.

    ★ **SOLUTION:** B

13. [2 points] For a perceptron classifier:

    (a) The test accuracy will, in general, be higher with the original features.

(b) The test accuracy will, in general, be higher with the doubled features.

(c) The test accuracy will always be the same with either feature set.

★ **SOLUTION:** C

14. [2 points] In neural networks, what is the benefit of the ReLU activation function over a Sigmoid activation function?

    (a) ReLUs allow the model to learn non-linear decision boundaries
    (b) ReLUs allow for faster backpropogation gradient calculations
    (c) The ReLUs activation function can be used in output layers while a sigmoidal activation function cannot.
    (d) All of the above

    ★ **SOLUTION:** B

15. [2 points] Consider a convolutional net where the p-dimensional input data is laid out in a one-dimensional fashion (e.g. for text or speech), and where there are k filters (kernels), each of size $m \times 1$. Assume no padding, and a stride of size s. Which of the following **best** approximates the number of outputs of this layer?

    (a) ksp/m
    (b) $kspm$
    (c) ksm/p
    (d) spm
    (e) kmp/s

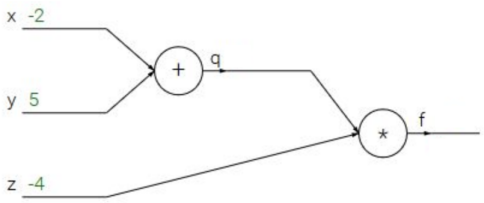
    ★ **SOLUTION:** E

16. [2 points] Suppose you have inputs as $x = -2$, $y = 5$, and $z = -4$. You have a neuron q and neuron f with functions:

$$q = x + y$$
$$f = q * z$$

    What is the gradient of f with respect to x, y, and z? See the figure below.

    (a) $(-3, 4, 4)$
    (b) $(4, 4, 3)$
    (c) $(-4, -4, 3)$
    (d) $(3, -4, -4)$



★ **SOLUTION:** C

17. [2 points] In which of the following neural net architectures do some of the weights get reused more than once in each single forward pass?

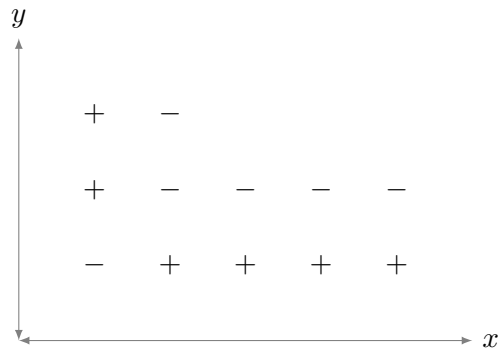
    (a) convolutional neural network
    (b) recurrent neural network
    (c) fully connected neural network
    (d) autoencoder
    (e) Both A and B

★ **SOLUTION:** E

18. [2 points] In AdaBoost, we choose α_t as the weight of the t-th weak learner, where

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \epsilon_t}{\epsilon_t} \right)$$

    $\varepsilon_t = P_{x \sim D_t} [h_t(x) \neq y]$ is the weighted fraction of examples misclassified by the t-th weak learner. Here our weak learners are depth 1 decision trees that yield vertical or horizontal half-plane decision boundaries. If we conduct two iterations of boosting on the following dataset, Which is larger, α_1 or α_2?



(a) $\alpha_1 > \alpha_2$

(b) $\alpha_2 > \alpha_1$

(c) $\alpha_1 = \alpha_2$

(d) Not enough information

★ **SOLUTION:** A

19. [2 points] Suppose you have a classification problem where you want to penalize misclassifications more the farther they are from the decision boundary. How many of the following loss functions would be appropriate?

- $0-1$ loss
- hinge loss
- logistic loss
- exponential loss

(a) none

(b) 1

(c) 2

(d) 3

(e) all 4

★ **SOLUTION:** D All but 0/1

20. [2 points] Suppose you encountered a classification problem where you **do not** want to additionally reward highly confident correct classifications. What choice(s) of loss would be appropriate?

    (a) $0-1$ loss only
    (b) hinge loss only
    (c) exponential loss only
    (d) $0-1$ or hinge loss
    (e) hinge or exponential loss

★ **SOLUTION:** D

21. [2 points] Which of the following losses are **not** convex?

    (a) $0-1$ loss
    (b) hinge loss
    (c) exponential loss
    (d) (a) and (b)
    (e) (a), (b) and (c)

★ **SOLUTION:** A

22. [1 points] *True or False?* Removal of a support vector will always change the SVM decision boundary.

★ **SOLUTION:** False

23. [1 points] *True or False?* Radial Basis Functions (RBFs) use a Gaussian kernel to transform a p-dimensional feature space (x) to a (k-dimensional) transformed feature space where $k \leq p$.

★ **SOLUTION:** False

24. [1 points] *True or False?* Principle Component Regression (PCR) uses the right singular vectors of the feature matrix X to transform a p-dimensional feature space (x) to a (k-dimensional) transformed feature space where $k \leq p$.

★ **SOLUTION:** True

25. [1 points] *True or False?* A perceptron is guaranteed to learn a perfect decision boundary within a finite number of iterations for linearly separable data.

★ **SOLUTION:** True

26. [1 points] *True or False?* Least Mean Squares (LMS) is an online approximation to linear regression and perceptrons are online approximations to SVMs.

★ **SOLUTION:** True

27. [2 points] After performing SVD on a dataset with 5 features, you retrieve eigenvalues 6, 5, 4, 3, 2. How many components should we include to explain at least 75% of the variance of the dataset?

    (a) 1
    (b) 2
    (c) 3
    (d) 4

★ **SOLUTION:** C

28. [1 points] *True or False?* After performing SVD on a dataset, you notice the eigenvalues returned are all approximately equal. You expect variance explained to be approximately linear to the number of components used for PCA.

★ **SOLUTION:** True

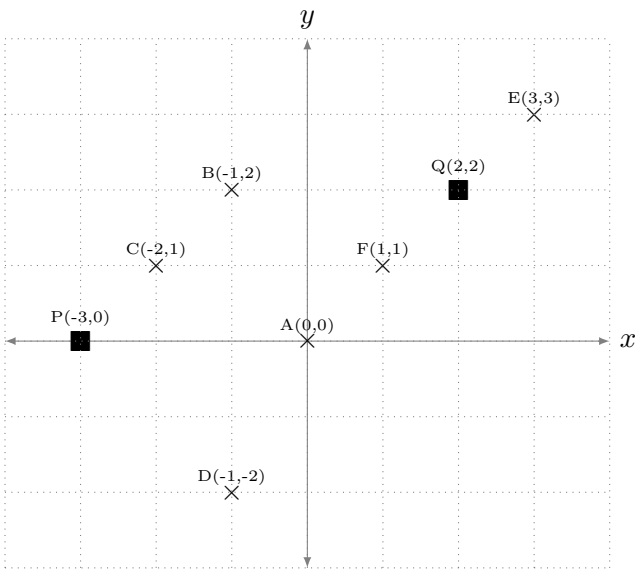
29. [1 points] *True or False?* Given a dataset with n features, if we know $(n-1)$ principal components of the dataset, then we can determine the missing principal component.

★ **SOLUTION:** True

30. [1 points] *True or False?* Changing one feature of a dataset with multiple features from centimeters to inches will not affect the outcome of PCA.

★ **SOLUTION:** False

For the next 4 questions, refer the the graph presented.



31. [2 points] In the first step of K-means with the standard Euclidean distance metric, which points will be assigned to the cluster centered at P?

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

★ SOLUTION: C

32. [2 points] Continue running K-means with the standard Euclidean distance metric. What does the cluster center P get updated to? (Do not include P as a point).

    (a) $(-\frac{3}{2}, \frac{1}{2})$
    (b) $(-\frac{4}{3}, \frac{1}{3})$
    (c) $(-\frac{4}{3}, 0)$
    (d) $(-1, \frac{1}{3})$
    (e) $(-\frac{3}{2}, \frac{1}{3})$

★ **SOLUTION:** B

While K-means used Euclidean distance in class, we can extend it to other distance functions, where the assignment and update phases still iteratively minimize the total (non-Euclidian) distance. Here, consider the Manhattan distance:

$$d'((A_1, A_2), (B_1, B_2)) = |A_1 - B_1| + |A_2 - B_2|$$

Again start from the original locations for P and Q as shown in the figure, and perform the update assignment step and the update cluster center step using Manhattan distance as the distance function:

33. [2 points] Starting from the same initial configuration, select all points that get assigned to the cluster with center at P, under this new distance function $d'(A, B)$.

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

★ **SOLUTION:** B

34. [2 points] What does cluster center P now get updated to, under this new distance function $d'(A, B)$? (Do not include P as a point).

    (a) $(-\frac{3}{2}, \frac{1}{2})$
    (b) $(-\frac{4}{3}, \frac{1}{3})$
    (c) $(-\frac{4}{3}, 0)$
    (d) $(-1, -\frac{1}{3})$
    (e) $(-\frac{3}{2}, \frac{1}{3})$

★ **SOLUTION:** * The exam as given did not have the correct solution option; it is now D

35. [1 points] *True or False?* The F1 score is generally a better performance measure than accuracy when there is extreme class imbalance in the labels.

★ **SOLUTION:** True

36. [1 points] *True or False?* An AUC (Area under the ROC curve) of 0.4 on test data suggests overfitting.

★ **SOLUTION:** True

37. [1 points] *True or False?* LIME (Local Interpretable Model-Agnostic Explanations) fits a linear model to observations close a point of interest and determines which features in that linear model are most influential in making the prediction.

★ **SOLUTION:** * The use of "observations" here is ambiguous; it fits points that are created as perturbations of the original point.

38. [1 points] *True or False?* When doing boosting to compute ensembles of trees, using complex (high depth) trees generally helps to improve test set accuracy.

★ **SOLUTION:** False

39. [1 points] *True or False?* When running linear regressions, it is a good idea to look at the largest (in absolute value) regression weights to see which features are most influential in determining the predictions.

★ **SOLUTION:** False

40. [2 points] For Naive Bayes, what happens to our document posteriors as we increase our pseudo-count parameter?

    (a) They approach 0
    (b) They approach 1
    (c) They approach the document priors
    (d) None of the above

★ **SOLUTION:** C

41. [2 points] We are trying to use Naive Bayes to classify a Facebook meme as either funny or sad. Suppose out of 100 training memes, we see that 60 of these memes are funny while 40 are sad. Also, assume that our dictionary consists of only three words, and the counts of the words for funny and sad memes are listed below.

Word	Funny Count	Sad Count
Yum	40	5
Friends	40	15
Cry	20	30

    If we see a meme post with one occurrence of the word friends and one occurrence of the word cry (with no other words), which class has higher posterior probability?

    (a) Funny
    (b) Sad

★ **SOLUTION:** A There are 100 memes, 60 funny and 40 sad. (as stated in the question text)

Of the 60 funny memes, 40 contain "yum". 40 contain "friends" and 20 contain "cry". (The memes have more than one word in them, so a meme like "froyo with my friends, yum!" has both "friends" and "yum" in it.)

When we compute p(friends—funny), that is short for the probability a meme with class label "funny" contains the word "friends", which is 40/60, not 40/(40+40+20).

42. [1 points] *True or False?* Because LDA has "hidden variables" representing the mixture of topics within each document and the topic that each word in each document come from, it is often solved using the EM algorithm.

★ **SOLUTION:** True

43. [1 points] *True or False?* EM algorithms are attractive because for problems such as estimating Gaussian Mixture Models, they are guaranteed to find a global optimum in likelihood.

★ **SOLUTION:** False

44. [1 points] *True or False?* Power methods for estimating eigenvectors are attractive because they are guaranteed to find a global optimum in reconstruction error when used in PCA.

★ **SOLUTION:** True

45. [2 points] Which of the following statement about Hidden Markov Model (HMM) is **not** true?

    (a) An HMM is a deterministic model because it explicitly describes the conditional distribution of the output given the current state.

    (b) The Markov process assumes that future is independent of the past given the present.

    (c) HMMs allow us to compute the joint probability of a set of hidden states given a set of observed states.

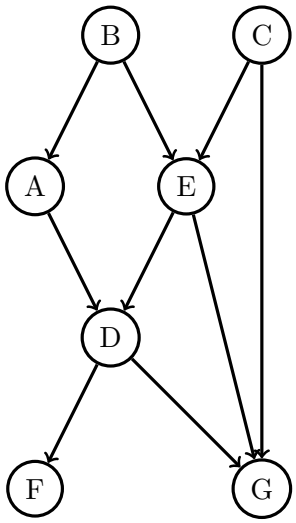
    (d) In HMM, given initial hidden states, constant transition matrix, emission matrix, and no other information, the observation will converge to only one state after a long sequence of prediction.

★ **SOLUTION:** * A or D

46. [1 points] *True or False?* The EM algorithm does a kind of "gradient descent" in likelihood, since both steps are guaranteed to decrease the negative log-likelihood.

★ **SOLUTION:** True

The following 5 questions are related to this graph:



47. [1 points] *True or False?* The joint probability of this graph can be represented as:

$P(A \mid B)P(B)P(C)P(D \mid A, E)P(E \mid B, C)P(F \mid D)P(G \mid C, E, D)$

★ **SOLUTION:** True

48. [1 points] *True or False?* The class of joint probability distributions that can be represented by the resulting Bayesian network:

$P(A \mid B)P(B)P(C)P(D \mid A, E)P(E \mid B, C)P(F \mid A, B, C, D, E)P(G \mid A, B, C, E, D)$

is smaller than the original network shown above.

★ **SOLUTION:** False

49. [1 points] *True or False?*

$$B \perp C \mid F$$

★ **SOLUTION:** False

50. [1 points] *True or False?*

$$A \perp C \mid G$$

★ **SOLUTION:** False

51. [1 points] *True or False?*

$$A \perp G \mid C, E$$

★ **SOLUTION:** False

52. [1 points] *True or False?* Although speech-to-text and text-to-speech is usually modeled using LSTMs or other RNNs, one could also use CNNs with one-dimensional filters.

★ **SOLUTION:** True

53. [1 points] *True or False?* CNNs work well on high dimensional problems like medical diagnosis from health records (which contain varied features like age, weight, temperature, lab results, disease history, etc.)

★ **SOLUTION:** False

54. [1 points] *True or False?* Vanilla RNNs, unlike HMMs, do not forget things exponentially quickly.

★ **SOLUTION:** False

55. [1 points] *True or False?* Q-learning is guaranteed to converge (for discrete states and actions) so long as all (state, action) pairs are visited infinitely often.

★ **SOLUTION:** * It is, but only given specific constraints on the learning rate.

56. [1 points] *True or False?* In Q-learning, $Q(s, a)$ represents the expected discounted reward of taking action a in state s and subsequently following an optimal policy.

★ **SOLUTION:** False

57. [1 points] *True or False?* Epsilon-greedy Reinforcement Learning methods "exploit" by using an optimal policy a (small) fraction, given by ϵ, and "explore" a large fraction $(1 - \epsilon)$ of the time.

★ **SOLUTION:** False

58. [1 points] *True or False?* Current RL methods for game play, such as alphaZero, unlike earlier methods that trained a "new" Q-function by playing against an older one, now just play the "new" network against itself.

★ **SOLUTION:** True

59. [1 points] *True or False?* Value Iteration iteratively updates V using Bellman's equation, and is guaranteed to converge to the unique optimum represented by the solution to Bellman's equation (if all states are visited infinite numbers of times).

★ **SOLUTION:** True

60. [1 points] *True or False?* Autoencoders always take an input and pass it through an "encoder" which produces a lower dimensional representation which is then passed through a "decoder" to reconstruct the input as accurately as possible.

★ **SOLUTION:** False

61. [1 points] *True or False?* When picking which additional points, x, to label for linear regression, it is desirable to pick points that are "as spread out as possible" (i.e. as far away as possible from the existing points)

★ **SOLUTION:** True

62. [1 points] *True or False?* When picking which additional points, x, to label for SVMs, it is desirable to pick points that are "as spread out as possible" (i.e. far away from the existing points)

★ **SOLUTION:** False

63. [1 points] *True or False?* The most widely used experimental design methods pick new points to label such that they maximize a norm $||X^T X||_p$ for some p.

★ **SOLUTION:** * Too complex; everyone was given credit: They minimize the norm of the inverse of that matrix. This is often, but not always the same as maximizing that matrix.

64. [1 points] *True or False?* When doing active learning for SVMs, labeling the x's for which one is "most uncertain" will tend to select points that are closer to the separating hyperplane.

★ **SOLUTION:** True

65. [1 points] *True or False?* The "Query by Committee" active learning method makes more sense to use with linear regression than with random forests.

★ **SOLUTION:** False

66. [1 points] If a standard CNN model has been trained to distinguish images of men from women in a setting where the training data has 75% women, then predictions on a test set of images drawn from the same distribution is more likely to have

    (a) under 75% women

    (b) very close to 75% women

    (c) over 75% women

★ **SOLUTION:** C

67. [2 points] When training a machine learning model on a data set which is not representative of the population of interest (e.g. when using Twitter users to represent the general population) it is best to:

    (a) Use L_1 rather than L_2 loss

    (b) restratify

    (c) use imputation

    (d) none of the above

★ **SOLUTION:** B

68. [1 points] *True or False?* Least Mean Squares does stochastic gradient descent in a negative log-likelihood.

★ **SOLUTION:** True