UNIVERSITY OF PENNSYLVANIA

CIS 520: Machine Learning
Final, Fall 2014

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

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 exam number.

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 scantron 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.

The exam is 9 pages long and has 78 questions.

Name: ______________________________

1. [0 points] This is version **A** of the exam. Please fill in the "bubble" for that letter.

2. [1 points] *True or False?* Under the usual assumptions ridge regression is consistent.

   ★ **SOLUTION:** True

3. [1 points] *True or False?* Under the usual assumptions ridge regression is unbiased.

   ★ **SOLUTION:** False

4. [1 points] *True or False?* Stepwise regression finds the global optimum minimizing its loss function (squared error plus the usual L_0 penalty).

   ★ **SOLUTION:** False

5. [1 points] *True or False?* k-means clustering finds the global optimum minimizing its loss function.

   ★ **SOLUTION:** False

6. [2 points] When doing linear regression with $n = 10,000$ observations and $p = 1,000,000$ features, if one expects around 500 or 1,000 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

7. [2 points] When doing linear regression, if we expect a very small fraction of the features to enter the model, we should use an

   (a) L_0 penalty
   (b) L_1 penalty
   (c) L_2 penalty

   ★ **SOLUTION:** A

8. [1 points] *True or False?* In general, in machine learning, we prefer to use unbiased algorithms due to their better accuracy.

★ **SOLUTION:** False

9. [1 points] *True or False?* L_1 penalized regression (Lasso) solves a convex optimization problem.

★ **SOLUTION:** True

10. [2 points] Which of the following loss functions is **least** sensitive to outliers?

    (a) Hinge loss

    (b) L_1 loss

    (c) Squared (L_2) loss

    (d) Exponential loss

★ **SOLUTION:** A

11. [1 points] *True or False?* For small training sets, Naive Bayes generally is more accurate than logistic regression.

★ **SOLUTION:** True

12. [1 points] *True or False?* Naive Bayes, as used in practice, in generally an MAP algorithm.

★ **SOLUTION:** True

13. [1 points] *True or False?* One can make a good argument that minimizing an L_1 loss penalty in regression gives "better" results than the more traditional L_2 loss function minimized by ordinary least squares.

★ **SOLUTION:** True

14. [1 points] *True or False?* Linear SVMs tend to be slower, but more accurate than logistic regression.

★ **SOLUTION:** False

15. [2 points] You estimate a ridge regression model with some data taken from your robot, and find (using cross validation) and optimal ridge penalty λ_1. You then buy a new sensor which has noise with 1/4 the variance (half the standard deviation) as before. Using the same number of observations as before you collect new data, and find a new optimal ridge penalty λ_2.

    Which of the following will be closest to true?

    (a) $\lambda_1/\lambda_2 = 1/4$
    (b) $\lambda_1/\lambda_2 = 1/2$
    (c) $\lambda_1/\lambda_2 = 1$
    (d) $\lambda_1/\lambda_2 = 2$
    (e) $\lambda_1/\lambda_2 = 4$

    ★ **SOLUTION:** E

16. [1 points] *True or False?* BIC can be viewed as an MDL method.

    ★ **SOLUTION:** True

17. [1 points] *True or False?* If you expect half of the features to enter a model, and have $n \gg p$, BIC is a better penalty to use than RIC.

    ★ **SOLUTION:** True

18. [1 points] *True or False?* The elastic net tends to select fewer features than well-optimized L_0 penalty methods.

    ★ **SOLUTION:** False

19. [1 points] *True or False?* The elastic net generally gives at least as good a model (in terms of test error) as Lasso.

    ★ **SOLUTION:** True

20. [1 points] *True or False?* The appropriate penalty in L_0-penalized linear regression can be determined by theory, e.g. using an MDL approach.

★ **SOLUTION:** True

21. [1 points] *True or False?* Ridge regression is 'scale invariant' in the sense the that test set prediction accuracy is unchanged if one rescales the features, x.

★ **SOLUTION:** False

22. [1 points] *True or False?* The the fact that a coefficient in a penalized linear regression is kept or killed (removed from the model) is generally a good indicator of the importance of the corresponding feature; features that are highly correlated with y will be kept and those with low correlation will be dropped.

★ **SOLUTION:** False

Consider an L_0 penalized linear regression with three different penalties. The resulting models have the following properties:

	bits to	bits to
	code residual	code the model
model 1	400	300
model 2	300	400
model 3	320	350

23. [1 points] *True or False?* Method 1 is overfitting.

★ **SOLUTION:** False

24. [1 points] *True or False?* Method 2 is overfitting.

★ **SOLUTION:** True

25. [1 points] *True or False?* The KL divergence between a "true" distribution ($p(A) = 0.5, p(B) = 0.5, p(C) = 0$) and an approximating distribution ($q(A) = 0.3, q(B) = 0.3q(C) = 0.4$) will be infinite.

★ **SOLUTION:** False

26. [1 points] *True or False?* Decision trees select features to add based on their expected "information gain". This has effect that features which take on many possible values tend to be preferentially added, since they tend to have higher information gain.

★ **SOLUTION:** True

27. [1 points] *True or False?* Boosting can be shown to optimize weights for an exponential loss function.

    ★ **SOLUTION:** True

28. [1 points] *True or False?* Key to the boosting algorithm is the fact that at each iteration more weight is given points that were misclassified. This often enables test set accuracy to continue to improve even after training set error goes to zero.

    ★ **SOLUTION:** True

29. [1 points] *True or False?* Perceptrons are a form of stagewise regression.

    ★ **SOLUTION:** False

30. [1 points] *True or False?* Voted perceptrons are generally more accurate than regular ("simple") perceptrons.

    ★ **SOLUTION:** True

31. [1 points] *True or False?* Voted perceptrons are generally faster (at test time) than averaged perceptrons.

    ★ **SOLUTION:** False

32. [1 points] *True or False?* Perceptrons (approximately) optimize a hinge loss.

    ★ **SOLUTION:** True

33. [1 points] *True or False?* For a linearly separable problem, standard perceptrons are guaranteed to find a lienarly separating hyperplane if there are no repeated x's with inconsistent labels.

    ★ **SOLUTION:** True

34. [1 points] Which of the following classifiers has the lowest 0-1 error (L_0 loss) given a training set with an infinite number of observations.

    (a) Logistic regression

    (b) Naive Bayes

★ **SOLUTION:** A

35. [1 points] *True or False?* If we consider a linear SVM as a kernel SVM, then the kernel function is the inner product between the x's.

    ★ **SOLUTION:** True

36. [1 points] *True or False?* Radial Basis Functions (RBFs) can be used either to reduce or to increase the effective dimensionality p of a regression problem.

    ★ **SOLUTION:** True

37. [1 points] *True or False?* Any SVM problem can be made linearly separable with the right selection of a kernel function.

    ★ **SOLUTION:** True

38. [1 points] *True or False?* The number of support vectors found by an SVM depends upon the size of the penalty on the slack variables.

    ★ **SOLUTION:** True

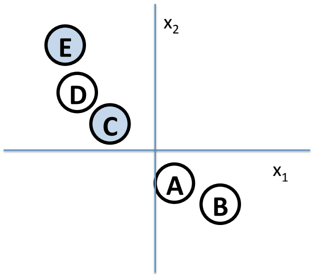
39. [0 points] *True or False?* Because SVMs already seek large margin solutions, they do not, in general, require inclusion of a separate regularization penalty. *This is badly phrased and was thrown out.*

    ★ **SOLUTION:** False

40. [1 points] *True or False?* An SVM with a Gaussian kernel, $exp(-\frac{\|x-y\|}{C})$, will have a lower expected variance when $C = 1$ than when $C = 10$,

    ★ **SOLUTION:** False

41. [2 points] In the figure below which points are support vectors? A B and D are in class 1, C and E are in class 2



(a) A, C
(b) A, C, D
(c) Not enough information was provided to tell.

★ **SOLUTION:** C

42. [2 points] For the primal problem for non-negative weighted regression is:
$min_w \sum_i (y_i - w^T x)^2$
s.t. $-w_j \leq 0$ for $j = 1...p$
The dual problem is to solve

(a) $max_\lambda \sum_i (y_i - w^T x)^2 + \sum_j \lambda_j w_j$ s.t. $\lambda_j \geq 0$
(b) $max_\lambda \sum_i (y_i - w^T x)^2 + \lambda \sum_j w_j$ s.t. $\lambda \geq 0$
(c) $max_\lambda \sum_i (y_i - w^T x)^2 - \sum_j \lambda_j w_j$ s.t. $\lambda_j \geq 0$
(d) $min_\lambda \sum_i (y_i - w^T x)^2 + \sum_j \lambda_j w_j$ s.t. $\lambda_j \geq 0$
(e) $min_\lambda \sum_i (y_i - w^T x)^2 - \sum_j \lambda_j w_j$ s.t. $\lambda_j \leq 0$

★ **SOLUTION:** C

43. [1 points] *True or False?* For the above optimization problem, the constraint corresponding to each weight is *binding* if and only if the weight is zero.

★ **SOLUTION:** True

44. [2 points] Which of the following methods **cannot** be kernelized?

(a) k-NN
(b) linear regression
(c) perceptrons
(d) PCA
(e) All of the above methods can be kernelized.

★ **SOLUTION:** E

45. [1 points] *True of False?* Any function $\phi(x)$ can be used to generate a kernel using $k(x, y) = \phi(x)^T\phi(y)$.

    ★ **SOLUTION:** True

46. [1 points] *True or False?* If there exists a pair of points x and y such that $k(x, y) < 0$, then $k()$ can not be a kernel.

    ★ **SOLUTION:** False

47. [1 points] *True or False?* All entries in a kernel matrix must be non-negative.

    ★ **SOLUTION:** False

48. [1 points] *True or False?* A kernel matrix must be symmetric.

    ★ **SOLUTION:** True

49. [1 points] *True or False?* Any norm $||x||$ can be used to define a distance by defining $d(x, y) = ||x - y||$

    ★ **SOLUTION:** True

50. [1 points] *True or False?* $k(x, y) = e^{(||x-y||_2^2)}$ is a legitimate kernel function

    ★ **SOLUTION:** False

51. [2 points] The number of parameters needed to specify a Gaussian Mixture Model with 4 clusters, data of dimension 3, and a single (full) covariance matrix shared across all 4 clusters is:

    (a) fewer than 16

    (b) between 16 and 20 (inclusive)

    (c) 21 or 22

    (d) 23

    (e) 24 or more

★ **SOLUTION:** C

52. [1 points] *True or False?* EM is a search algorithm for finding maximum likelihood (or sometimes MAP) estimates. Thus, it can, in theory, be replaced by other search algorithm that also maximizes the same likelihood function.

    ★ **SOLUTION:** True

53. [1 points] *True or False?* The L_2 reconstruction error from using k-component PCA to approximate a set of observations X, can be characterizaed in in terms of the k largest eigenvalues of $X'X$.

    ★ **SOLUTION:** False

54. [1 points] *True or False?* A positive definite symmetric real square matrix has only positive eigenvalues.

    ★ **SOLUTION:** True

55. [1 points] *True or False?* For real world data sets, the principle components of a matrix X are preferably found using SVD of X rather than actually finding the eigenvectors of $X'X$.

    ★ **SOLUTION:** True

56. [1 points] *True or False?* The singular values of X are equal to the eigenvalues of $X'X$.

    ★ **SOLUTION:** False

57. [1 points] *True or False?* For an NxP matrix, X. $(N > P)$ the k "largest" right singular vectors will be the same as the loadings of X.

    ★ **SOLUTION:** True

58. [1 points] *True or False?* Principle component regression (PCR) in effect does regularization, and thus offers a partial, if not exact replacement for Ridge regression.

    ★ **SOLUTION:** True

59. [1 points] *True or False?* Principle component regression (PCR), like linear regression, is scale invariant.

★ **SOLUTION:** False

60. [2 points] The dominant cost of linear regression, when $n >> p$ scales as

    (a) np
    (b) np^2
    (c) n^2p
    (d) p^3

★ **SOLUTION:** B

61. [1 points] *True or False?* Deep neural networks are almost always supervised learning methods.

★ **SOLUTION:** False

62. [1 points] *True or False?* Deep neural networks most often use an architecture in which all "nodes" on each layer are connected to all "nodes" on the following layer, but have no connections to other "deeper" layers.

★ **SOLUTION:** False

63. [1 points] *True or False?* Deep neural networks currently hold the records for best machine learning performance in problems ranging from speeh and vision to natural language processing and brain image modeling (MRI).

★ **SOLUTION:** False

Consider the following confusion matrix

		corrent answer	
		True	False
predicted	True	8	2
answer	False	12	11

64. [1 points] For the above "confusion matrix" the precision is

    (a) 2/10
    (b) 8/20
    (c) 19/33
    (d) none of the above

★ **SOLUTION:** D

65. [1 points] For the above "confusion matrix" the recall is

    (a) 2/10
    (b) 8/20
    (c) 19/33
    (d) none of the above

★ **SOLUTION:** B

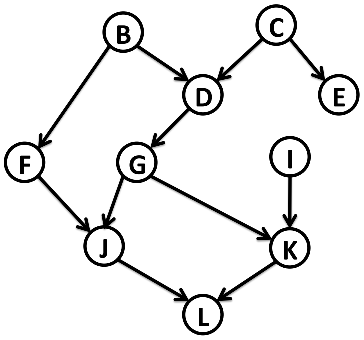
66. [1 points] *True or False?* L_2 loss (sometime with a regularization penalty) is widely used because it usually reflects the actual loss function for applications in business and science.

★ **SOLUTION:** False

67. [1 points] *True or False?* One can compute a description length (for the model plus residual) for a belief net and the data it represents, and a causal belief net is likely to have a shorter description length than one that just captures the conditional independence structure of the same data.

★ **SOLUTION:** True

The following questions refer to the following figure;
$\perp$ means "is conditionally independent of."



68. [1 points] *True or False?* $(B \perp C | D)$

★ **SOLUTION:** False

69. [1 points] *True or False?* $(J \perp K | G)$

★ **SOLUTION:** True

70. [1 points] *True or False?* $(J \perp K | L)$

★ **SOLUTION:** False

71. [1 points] *True or False?* $(B \perp J | F, G)$

★ **SOLUTION:** True

72. [1 points] *True or False?* $(F \perp I | G, L)$

★ **SOLUTION:** False

73. [1 points] *True or False?* G d-separates D and J

★ **SOLUTION:** False

74. [2 points] What is the minimum number of parameters needed to represent the full joint distribution $P(B, C, D, E, F, G, I, J, K, L)$ in the network, given that all variables are binary? *hint:* Parameters here refer to the value of each probability. For example, we need 1 parameter for $P(X)$ and 3 parameters for $P(X, Y)$ if X and Y are binary.

    (a) < 20
    (b) 20-30
    (c) 31-99
    (d) more than 100

★ **SOLUTION:** B

75. [1 points] *True or False?* When building a belief net from a set of observations where A, B are Boolean variables, if $P(A = True | B = True) = P(A = True)$, then we know that there should **not** be a link from A to B.

★ **SOLUTION:** True

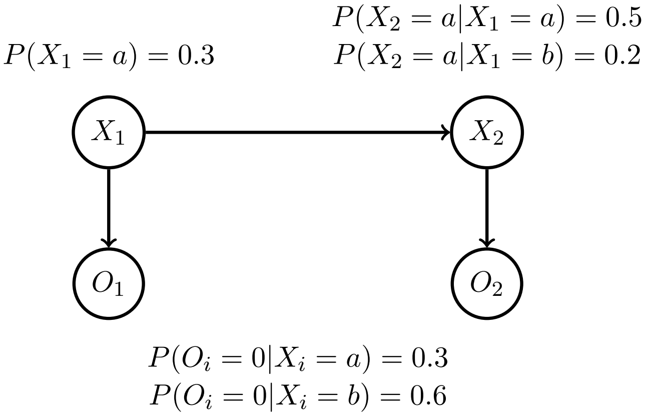
76. [1 points] *True or False?* When building a belief net from a set of observations where A, B are Boolean variables, if $P(A = True | B = True) = P(A = True)$, then we know that there should **not** be a link from B to A.

★ **SOLUTION:** True

77. [1 points] *True or False?* The emission matrix in an HMM represents the probability of the state, given an observation.

★ **SOLUTION:** False

For the next question, consider the Bayes Net below with parameter values labeled. This is an instance of an HMM. (Similar to homework 8)


$P(X_2 = a|X_1 = a) = 0.5$
$P(X_2 = a|X_1 = b) = 0.2$
$P(X_1 = a) = 0.3$
X_1
X_2
O_1
O_2
$P(O_i = 0|X_i = a) = 0.3$
$P(O_i = 0|X_i = b) = 0.6$

78. [3 points] Suppose you have the observation sequence $O_1 = 1, O_2 = 0$. What is the prediction of Viterbi Decoding? (Maximize $P(X_1, X_2 \mid O_1 = 1, O_2 = 0)$)

(a) $X_1 = a, X_2 = a$

(b) $X_1 = a, X_2 = b$

(c) $X_1 = b, X_2 = a$

(d) $X_1 = b, X_2 = b$

★ **SOLUTION:** D