UNIVERSITY OF PENNSYLVANIA

CIS 520: Machine Learning
Final, Fall 2013

Exam policy: This exam allows two one-page, two-sided cheat sheets; 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 scantron 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 10 pages long and has 93 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?* Ridge regression finds the global optimum for minimizing its loss function (squared error plus an appropriate penalty).

3. [2 points] Ridge regression minimizes which of the following? (Assume, as usual, n observations).

   (a) $\sum_i (y_i - w^\top x_i)^2 + \lambda \|w\|_2^2$
   (b) $\sum_i (y_i - w^\top x_i)^2 + \lambda \|w\|_2$
   (c) $(1/n) \sum_i (y_i - w^\top x_i)^2 + \lambda \|w\|_2^2$
   (d) $(1/n) \sum_i (y_i - w^\top x_i)^2 + \lambda \|w\|_2$
   (e) $\sum_i (y_i - w^\top x_i)^2 - \lambda \|w\|_2$

4. [2 points] When doing linear regression with $n = 1,000,000$ observations and $p = 10,000$ 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) no penalty

5. [1 points] *True or False?* The conjugate prior to the Gaussian distribution is a Gaussian distribution.

6. [1 points] *True or False?* Lasso finds the global optimum for minimizing its loss function (squared error plus an appropriate penalty).

7. [1 points] *True or False?* Stepwise regression with a BIC penalty term finds the global optimum for minimizing its loss function (squared error plus an appropriate penalty).

8. [1 points] *True or False?* The EM algorithm finds the global optimum for clustering data from a mixture of Gaussians, assuming the number of clusters K is set to the correct value.

9. [1 points] *True or False?* When using k-nearest neighbors, the implementation choice that most impacts predction accuracy is usually the choice of k.

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

    (a) Hinge loss
    (b) Squared loss
    (c) Exponential loss
    (d) 0-1 loss

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

12. [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.

13. [1 points] *True or False?* Naive Bayes can be formulated either as minimizing an L_2 loss function or as maximizing a likelihood function.

14. [1 points] *True or False?* Naive Bayes, as used in practice, in generally an MLE algorithm.

15. [1 points] *True or False?* Ridge regression can be formulated as an MLE algorithm.

16. [1 points] *True or False?* SVMs are generally formulated as MLE algorithms.

17. [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.

18. [1 points] *True or False?* L_2 penalized linear regression is, in general, more sensitive to outliers than L_1 penalized linear regression. (I.e. one point that is far from the predicted regression line will have more effect on the regression coefficients.)

19. [1 points] *True or False?* Large margin methods like SVMs tend to be slightly less accurate in predictions (when measured with an L_0 loss function) than logistic regression.

20. [1 points] *True or False?* One can do kernelized logistic regression to get many of the same benefits one would get using kernels in SVMs.

21. [1 points] *True or False?* It is *not* possible to both reduce bias and reduce variance by changing model forms (e.g. by adding a kernel to an SVM).

22. [1 points] *True or False?* It is sometimes useful to do linear regression in the dual space.

23. [1 points] *True or False?* Linear SVMs tend to overfit more than standard logistic regression.

24. [2 points] You are doing ridge regression. You first estimate a regression model with some data. You then estimate a second model with four times as many observations (but the same ridge penalty). Roughly how do you expect the regression coefficients to change when more data is used?

    (a) The coefficients should on average shrink towards zero (become smaller in absolute value).

    (b) The coefficients should on average move away from zero (become larger in absolute value).

    (c) The coefficients will not change.

25. [1 points] *True or False?* Suppose we have two datasets X_1 and X_2 which each represent the same observations (and the same Y labels) except that X_1's features are a subset of the features of X_2. (I.e., X_2 is X_1 with extra columns added.) Stepwise linear regression with an L_0 penalty will always add at least as many features when trained on the bigger feature set, X_2.

26. [1 points] *True or False?* RIC (Risk Inflation Criterion) can be viewed as an MDL method.

27. [1 points] *True or False?* If you expect a tiny fraction of the features to enter a model, BIC is a better penalty to use than RIC.

28. [1 points] *True or False?* If you expect a tiny fraction of the features to enter a model, an appropriate L_0 penalty will probably work better than than an L_1 penalty.

29. [1 points] *True or False?* The elastic net tends to select fewer features than well-optimized L_0 penalty methods.

30. [1 points] *True or False?* Estimating the elastic net is a convex optimization problem, and hence relatively fast.

31. [1 points] *True or False?* The appropriate penalty in L_1-penalized linear regression can be determined by theory, e.g. using an MDL approach.

32. [1 points] *True or False?* The larger (in magnitude) coefficients in a linear regression correspond to (multiply) the more 'important' features (those that when changed have a larger effect on y).

33. [1 points] *True or False?* KL divergence can be used to measure (dis)similarity when doing k-means clustering.

34. [1 points] *True or False?* KL divergence is a valid distance metric.

35. [1 points] *True or False?* KL divergence is generally used to see how well one distribution approximates another one.

36. [1 points] *True or False?* Given a label drawn from the following probability distribution
P(Apple) = 0.5, P(Banana) = 0.25, P(Carrot) = 0.25,
its entropy (in bits) is
$0.5log_2(0.5)+0.5log_2(0.5)+0.25log_2(0.25)+0.75log_2(0.75)+0.25log_2(0.25)+0.75log_2(0.75)$.

37. [1 points] *True or False?* The theory behind boosting is based on there being a weak classifier h_t whose error on the weighted dataset is always less than (or equal to) 0.5.

38. [1 points] *True or False?* At each round t of boosting, the minimum error of the classifier on the weighted dataset ϵ_t is a monotone non-decreasing function, i.e. $\epsilon_i \leq \epsilon_j$ for all $i < j$.

39. [1 points] *True or False?* Hinge loss is upper bounded by the loss function of boosting.

40. [1 points] *True or False?* Boosting usually averages over many decision trees, each of which is learned without pruning.

41. [1 points] *True or False?* In general, if you have multiple methods for making predictions, it is better to pick the best one rather than using the majority vote of the methods as the prediction.

42. [1 points] *True or False?* Perceptrons are an online alternative to SVMs that generally converge to the same solution as SVMs.

43. [1 points] *True or False?* Averaged and voted perceptrons result in models that contain similar numbers of parameters and hence are of similar accuracy and similar computational cost to use in making preditions.

44. [2 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) Standard SVM with optimal regularization

(b) Logistic regression with optimal regularization

(c) Naive Bayes

(d) Bayesian classifier (one that classifies using an estimated model $p(y|x;\theta)$ using the true distributional form (the correct equation for $p(y|x)$.

45. [2 points] An SVM using a Gaussian kernel gives the same separating hyperplane as a linear SVM when σ in the denominator of the exponential in the Gaussian approaches:

    (a) 0

    (b) ∞

    (c) the Gaussian kernel SVM does not give the same hyperplane as the linear SVM in any limit

46. [1 points] *True or False?* SVMs are, in general, *more* sensitive to the actual distribution of the data than logistic regression is.

47. [2 points] Which model generally has the highest bias?

    (a) linear SVM

    (b) Gaussian SVM

    (c) perceptron with constant update

48. [2 points] It is best to use the primal SVM when ... (Choose the *best* answer.)

    (a) $p >> n$

    (b) $n >> p$

    (c) we don't have a kernel

    (d) the dual doesn't exist

49. [2 points] What is the objective for L_2 regularized L_1-loss SVM?

    (a) min $\|w\|_1 + C\sum_i \xi_i$
    subject to $y_i(w^Tx + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$ for $i = 1...n$

    (b) min $(1/2)\|w\|_2^2 + C\sum_i \xi_i$
    subject to $y_i(w^Tx + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$ for $i = 1...n$

    (c) min $\|w\|_1 + C\sum_i \xi_i^2$
    subject to $y_i(w^Tx + b) = 1 - \xi_i$ for $i = 1...n$

    (d) min $(1/2)\|w\|_2^2 + C\sum_i \xi_i^2$
    subject to $y_i(w^Tx + b) = 1 - \xi_i$ for $i = 1...n$

    (e) None of the above

50. [2 points] Which type(s) of SVM will tend to have the largest number of support vectors for any given dataset?

    (a) L_1 regularized L_1 loss SVM

    (b) L_1 regularized L_2 loss SVM

(c) L_2 regularized L_1 loss SVM

(d) L_2 regularized L_2 loss SVM

(e) Two of the above

51. [1 points] *True or False?* In the limit of infinite data ($n \to \infty$), L_1 regularized L_1 loss SVM and L_2 regularized L_1 loss SVM will give the same hyperplane.

52. [1 points] *True or False?* In the limit of infinite data ($n \to \infty$), L_2 regularized L_1 loss SVM and L_2 regularized L_2 loss SVM will give the same hyperplane.

53. [2 points] The effect of increasing the loss penalty, C, on L_2 regularized L_1 loss SVM is:

    (a) Increased bias, decreased variance

    (b) Increased bias, increased variance

    (c) Decreased bias, decreased variance

    (d) Decreased bias, increased variance

    (e) Not enough information to tell

54. [2 points] Suppose we have two datasets X_1, X_2 representing the same observations (with the same Y labels) except that X_1's features are a subset of the features of X_2. If we apply L_2 regularized L_1 loss SVM to this dataset then which is true?

    (a) The number of support vectors for $X_1 \leq$ the number of support vectors for X_2.

    (b) The number of support vectors for $X_1 \geq$ the number of support vectors for X_2.

    (c) Not enough information to tell.

55. [2 points] Kernels work naturally with which of the following:
    I. PCA II. linear regression
    III. SVMs IV. decision trees

    (a) I and III

    (b) II and III

    (c) I, II and III

    (d) all of them

56. [1 points] *True or False?* $k(x, y) = exp(-\|x - y\|)$ is a valid kernel for any norm $\|z\|_p$.

57. [1 points] *True or False?* $k(x, y) = \sqrt{\|x - y\|_2^2 + c^2}$ is a valid kernel.

58. [1 points] *True or False?* $k(x, y) = \sum_{i=1}^{n} \max(x_i, y_i)$ is a valid kernel.

59. [2 points] The number of parameters needed to specify a Gaussian Mixture Model with 4 clusters, data of dimension 5, and diagonal covariances is:

    (a) Between 5 and 15

    (b) Between 16 and 26

    (c) Between 27 and 37

    (d) Between 38 and 48

(e) More than 49

60. [2 points] The number of parameters needed to specify a Gaussian Mixture Model with 3 clusters, data of dimension 4, and spherical covariances is:

    (a) Between 5 and 15
    (b) Between 16 and 26
    (c) Between 27 and 37
    (d) Between 38 and 48
    (e) More than 49

61. [2 points] Suppose you are given an EM algorithm that finds maximum likelihood estimates for a model with latent variables. You are asked to modify the algorithm so that it finds MAP estimates instead. Which step or steps do you need to modify?

    (a) Expectation
    (b) Maximization
    (c) No modification is necessary
    (d) Both

62. [1 points] *True or False?* EM is a search algorithm for finding maximum likelihood (or sometimes MAP) estimates. Thus, it can be replaced by other search algorithm that also maximizes the same likelihood function.

63. [1 points] *True or False?* PCA can be formulated as an optimization problem that finds the set of k basis functions that best represent a set of observations X, in the sense of minimizing the L_2 reconstruction error.

64. [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.

65. [1 points] *True or False?* A positive definite symmetric real square matrix has only positive entries.

66. [1 points] *True or False?* The Frobenius norm of a matrix is equal to the sum of the squares of its eigenvalues.

67. [1 points] *True or False?* The Frobenius norm of a matrix is equal to the sum of the squares of all of its elements.

68. [1 points] *True or False?* PCA is often used for preprocessing images.

69. [1 points] *True or False?* PCA (Principal Component Analysis) is a supervised learning method.

70. [1 points] *True or False?* It is often possible to get a higher testing accuracy by training the same model on the PCA'ed dataset than on the original dataset.

71. [1 points] *True or False?* Using principal components (PCs) as the features when training Naive Bayes assures (since the principal components are orthogonal) that the Naive Bayes assumption is met.

72. [1 points] *True or False?* Thin SVD of a rectangular matrix $X = U_1 D_1 V_1^\top$ (where X is n*p with $n > p$ and D_1 is k*k) and that of its transpose $X^\top = U_2 D_2 V_2^\top$ (again with D_2 also k*k) always yields $D_1 = D_2$ as long as there are no repeated singular values.

73. [1 points] *True or False?* Thin SVD of a rectangular matrix $X = U_1 D_1 V_1^\top$ (where X is n*p with $n > p$ and D_1 is k*k) and that of its transpose $X^\top = U_2 D_2 V_2^\top$ (again with D_2 also k*k) always yields $U_1 = V_2$ as long as there are no repeated singular values.

74. [1 points] *True or False?* A k*k*k tensor Γ can be viewed as a mapping from three length k vectors (x_1, x_2, and x_3) to a scalar (y) such that the mapping is linear in each of the input vectors. Thus, given a bunch of observations of the form (y, x_1, x_2, x_3) the elements of Γ can be estimated using linear regression.

75. [1 points] *True or False?* Deep Neural Networks are usually semi-parametric models.

76. [1 points] *True or False?* Deep neural networks are almost always supervised learning methods.

77. [1 points] *True or False?* Random Forests is a generative model.

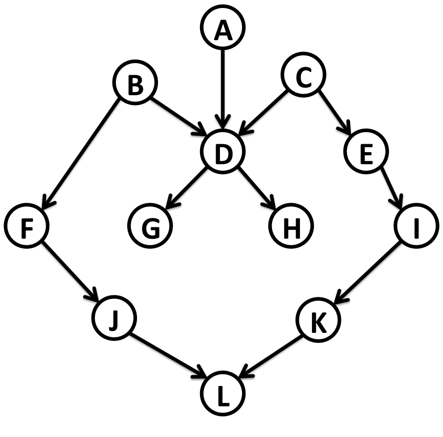
78. [1 points] *True or False?* Random Forests tend to work well for problems like image classification.

79. [2 points] Radial basis functions ____ the dimensionality of the feature space

    (a) only increase
    (b) only decrease
    (c) can either increase, decrease, or not change
    (d) don't change

80. [1 points] *True or False?* Latent Dirichlet Allocation is a latent variable model that can be estimated using EM-style optimization.

    The following questions refer to the following figure.



81. [1 points] *True or False?* $\neg(A \perp B | H)$

82. [1 points] *True or False?* $(C \perp I | E, L, G)$

83. [1 points] *True or False?* $(A \perp L | J, K, L)$

84. [1 points] *True or False?* $(B \perp K | H)$

85. [1 points] *True or False?* $(G \perp H | A, B, C)$

86. [1 points] *True or False?* J d-separates F and L

87. [2 points] What is the minimum number of parameters needed to represent the full joint distribution $P(A, B, C, D, E, F, G, H, I, J, K, L)$ 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) 12
   (b) 4095
   (c) 24
   (d) 29

88. [2 points] How many edges need to be added to "moralize" the network?

   (a) 3
   (b) 4
   (c) 5
   (d) 6

89. [1 points] *True or False?* Randomized search procedures, although though slow, can be used to reliably learn causal belief networks.

90. [1 points] *True or False?* The Markov transition matrix in an HMM gives the probability $p(h_t | h_{t+1})$.

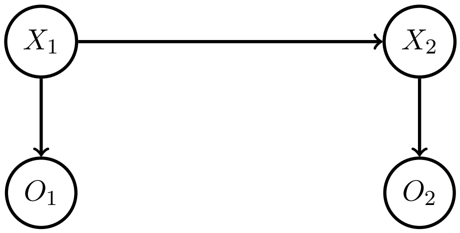
91. [1 points] *True or False?* Hidden Markov Models are generative models and hence are usually learned using EM algorithms.

92. [2 points] In EM models, which of the following equations do you use to derive the parameters θ?

   (a) $\log P(D|\theta) = \sum_i \log \sum_z q(Z_i = z|x_i) \frac{p_\theta(Z_i=z,x_i)}{q(Z_i=z|x_i)}$
   (b) $F(q, \theta) = \sum_i (H(q(Z_i|x_i)) + \sum_z q(Z_i = z|x_i) \log p_\theta(Z_i = z, x_i))$
   (c) $F(q, \theta) = \sum_i (-KL(q(Z_i|x_i)||p_\theta(Z_i|x_i)) + \log p_\theta(x_i))$
   (d) $\log P(D|\theta) = \sum_i \sum_z q(Z_i = z|x_i) \log \frac{p_\theta(Z_i=z,x_i)}{q(Z_i=z|x_i)}$

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

$$P(X_2 = a|X_1 = a) = 0.7$$
$$P(X_1 = a) = 0.4 \qquad P(X_2 = a|X_1 = b) = 0.2$$



$$P(O_i = 0|X_i = a) = 0.6$$
$$P(O_i = 0|X_i = b) = 0.3$$

93. [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) $X_1 = b, X_2 = b$