UNIVERSITY OF PENNSYLVANIA

CIS 520: Machine Learning
Final, Fall 2016

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 number and fill in the associated bubbles, and do not write your name.

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.*

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 has 88 questions, totalling 113 points.

Name: ________________________________

1. [0 points] This is version **B** 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, L1-penalized 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. [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

6. [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
   (d) L_0, L_1 and elastic net will all work approximately equally well.

★ **SOLUTION:** A

7. [1 points] *True or False?* Elastic net (L1/L2 regularization) solves a convex optimization problem.

★ **SOLUTION:** True

8. [2 points] Which of the following loss functions is **most** sensitive to outliers?

   (a) Hinge loss

   (b) L_1 loss

   (c) Squared (L_2) loss

★ **SOLUTION:** C

9. [1 points] *True or False?* Naive Bayes, as used in practice, is generally an MLE algorithm.

★ **SOLUTION:** False

10. [2 points] The hinge loss function used to train a support vector machine will give a nonzero penalty for a training observation x *only* when that observation is...

    (a) within the SVM margin

    (b) on the wrong side of the SVM separating hyperplane and outside the margin

    (c) on the correct side of the SVM separating hyperplane and outside the margin

    (d) A and B

    (e) A, B, and C

★ **SOLUTION:** D

11. [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

12. [1 points] *True or False?* Any method which predicts $\hat{y}(x)$ using a function of the form $\hat{y}(x_i) = w^T x_i$ where $w = f(X, y; \theta)$ is a function of the training data X, y with parameters θ is a *linear smoother*.

★ **SOLUTION:** False

13. [1 points] *True or False?* The elastic net tends to select fewer features than well-optimized L_0 penalty methods.

★ **SOLUTION:** False

14. [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

15. [1 points] *True or False?* L1-penalized regression ("LASSO") is 'scale invariant' in the sense that the test set prediction accuracy is unchanged if one rescales the features, x.

★ **SOLUTION:** False

16. [1 points] *True or False?* PCA is 'scale invariant' in the sense that the principle components (but not the loadings) are unchanged if one rescales the features, x.

★ **SOLUTION:** False

17. [1 points] *True or False?* Standard Support Vector Machines are 'scale invariant' in the sense that the test set prediction accuracy is unchanged if one rescales the features, x.

★ **SOLUTION:** False

18. [1 points] *True or False?* SVMs are in general preferable to logistic regression (with a ridge penalty) for problems with $n \leq p$

★ **SOLUTION:** False

19. [1 points] *True or False?* For linearly separable problems, perceptrons are guaranteed to converge to an optimal solution (one that classifies the training data perfectly).

★ **SOLUTION:** True

20. [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 ($q(A) = 0.5, q(B) = 0.25q(C) = 0.25$) is $0.5log(2)$

★ **SOLUTION:** True

21. [2 points] In the ID3 algorithm for training decision trees on data with labels y, a variable x on which to split is chosen which maximizes the ...

    (a) entropy of the labels y minus the entropy of the labels y given X

    (b) entropy of the labels y given X minus entropy of the labels y

    (c) entropy of the labels y given X

    (d) entropy of X

    (e) entropy of X minus the entropy of the labels y given X

★ **SOLUTION:** A

22. [1 points] *True or False?* For any two variables x and y having joint distribution $p(x, y)$, it is always true that $H[x, y] = H[x] + H[y]$ where H is the entropy function.

★ **SOLUTION:** False

23. [2 points] The largest possible entropy for a probability distribution defined over a space of N possible outcomes is

    (a) $\log N$

    (b) $\frac{1}{N} \log N$

    (c) $N \log N$

    (d) 1

    (e) N

★ **SOLUTION:** A

24. [2 points] In each round of AdaBoost, the misclassification penalty for a particular training observation is increased going from round t to round $t + 1$ if the observation was...

    (a) classified incorrectly by the weak learner trained in round t

    (b) classified incorrectly by the full ensemble trained up to round t

    (c) classified incorrectly by a majority of the weak learners trained up to round t

    (d) B and C

    (e) A, B, and C

★ **SOLUTION:** A

25. [1 points] *True or False?* Boosting minimizes an exponential loss function (subject to the model constraints).

★ **SOLUTION:** True

26. [1 points] *True or False?* Voted perceptrons are generally more accurate than regular ("simple") perceptrons.

    ★ **SOLUTION:** True

27. [1 points] *True or False?* The 'passive-aggressive' Perceptron updates the weight vector after seeing an observation if and only if the prediction it makes for that observation is on the wrong side of the separating hyperplane.

    ★ **SOLUTION:** False

28. [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

29. [1 points] *True or False?* Naive Bayes is widely used because it is fast and tends to give accurate estimates of the probabilities of category labels given observed features such as words in a document.

    ★ **SOLUTION:** False

30. [1 points] *True or False?* MLE is less likely to overfit than MAP since MLE tends to shrink parameters.

    ★ **SOLUTION:** False

31. [2 points] Suppose we wish to compute an MAP estimate of the mean μ of a Gaussian distribution. If our sample data has mean μ_D and we impose a Gaussian prior on μ with mean μ_0, then the MAP estimate for μ is...

    (a) a nonlinear function of μ_D and μ_0

    (b) a non-convex linear combination of μ_D and μ_0

    (c) a convex combination of μ_D and μ_0

    (d) proportional to the square root of the sum of squares of μ_D and μ_0

    (e) The functional form of the estimate depends on the data.

★ **SOLUTION:** C

32. [1 points] *True or False?* Any SVM problem can be made linearly separable with the right selection of a kernel function if there are not any identical x's with different labels.

★ **SOLUTION:** True

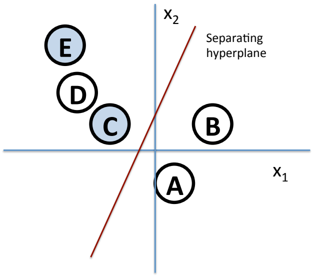
33. [1 points] *True or False?* The number of support vectors found by an SVM depends upon the size of the penalty on the slack variables.

★ **SOLUTION:** True

34. [1 points] *True or False?* An SVM with a Gaussian kernel, $exp(-\frac{\|x-y\|_2^2}{C})$, will have a lower expected variance when $C = 1$ than when $C = 10$,

★ **SOLUTION:** False

35. [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, B, C
(b) A, B, C, D
(c) something else (but it can be determined)
(d) Not enough information was provided to tell.

★ **SOLUTION:** B

36. [2 points] The primal problem for regression regularized by requiring all weights to be less than C is $min_w \sum_i (y_i - w^T x_i)^2$
s.t. $w_j \leq C$ for $j = 1...p$
The dual problem is to solve

(a) $max_\lambda \sum_i (y_i - w^T x_i)^2 + \sum_j \lambda_j (C - w_j)$ s.t. $\lambda_j \geq 0$

(b) $max_\lambda \sum_i (y_i - w^T x_i)^2 + \lambda \sum_j w_j$ s.t. $\lambda \geq 0$

(c) $max_\lambda \sum_i (y_i - w^T x_i)^2 - \sum_j \lambda_j (C - w_j)$ s.t. $\lambda_j \geq 0$

(d) $min_\lambda \sum_i (y_i - w^T x_i)^2 + \sum_j \lambda_j (C - w_j)$ s.t. $\lambda_j \geq 0$

(e) $min_\lambda \sum_i (y_i - w^T x_i)^2 - \sum_j \lambda_j (C - w_j)$ s.t. $\lambda_j \leq 0$

★ **SOLUTION:** C

37. [1 points] *True of False?* Radial Basis Functions use a Gaussian kernel; one that looks like a Gaussian probability density, but (generally) without the normalization constant (so it is no longer a probability density).

★ **SOLUTION:** True

38. [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

39. [1 points] *True or False?* For any valid kernel $k(x, y)$, it is always true that $k(x, y) \geq 0$,

★ **SOLUTION:** False

40. [1 points] *True or False?* For any valid kernel $k(x, y)$, $k(x, x) \geq k(x, y)$,

★ **SOLUTION:** false

41. [1 points] *True or False?* Any norm $||x||$ can be used to define a distance by defining $d(x, y) = ||x - y||$

★ **SOLUTION:** True

42. [1 points] *True or False?* $k(x, y) = 5x^\top y + 10(x^\top y)^2$ is a legitimate kernel function

★ **SOLUTION:** True

43. [1 points] *True or False?* The following matrix is positive semi-definite.

```
-1   1
 1  -1
```

★ **SOLUTION:** False

44. [2 points] Which of the following is **not** a legitimate matrix norm for a positive semi-definite matrix?

(a) The square root of the sum of the eigenvalues
(b) The sum of the eigenvalues
(c) The largest eigenvalue
(d) The square root of the sum of the squares of the matrix entries
(e) All of the above are legitimate norms

★ **SOLUTION:** A

45. [2 points] The number of parameters needed to specify a Gaussian Mixture Model with 3 clusters, data of dimension 2, where each of the 3 clusters can have it's own *diagonal* covariance matrix is:

(a) fewer than 12
(b) 12
(c) 13
(d) 14
(e) 15 or more

★ **SOLUTION:** D

46. [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 any other search algorithm that also maximizes the same likelihood function.

★ **SOLUTION:** True

47. [1 points]*True or False?* Iterating between the E-step and M-step will always converge to a local optimum of the likelihood (which may or may not also be a global optimum).

★ **SOLUTION:** True

48. [1 points] *True or False?* Gaussian Mixture Models have closed form solutions for both the E- and the M-steps in EM; LDA does not have a closed form solution for both.

★ **SOLUTION:** True

49. [1 points] *True or False?* In LDA, the words in each document are assumed to be drawn from a Dirichlet distribution. These distributions can vary across documents.

★ **SOLUTION:** False

50. [2 points] In expectation maximization (EM),

(a) the E-step updates model parameters, and the M-step updates hidden variable probabilities for each observation

(b) the E-step updates hidden variable probabilities for each observation, and the M-step updates model parameters

(c) the E-step evaluates the model loss function, and the M-step selects a feature which maximizes information gain

(d) the E-step selects a feature which maximizes information gain, and the M-step evaluates the model loss function

(e) the E-step initializes the model, and the M-step begins a two-step iterative process to optimize the model

★ **SOLUTION:** B

51. [1 points] *True or False?* The L_2 reconstruction error from using k-component PCA to approximate a matrix X of n observations of p features can be characterized in terms of the $p - k$ smallest eigenvalues of $X'X$.

★ **SOLUTION:** True

52. [1 points] *True or False?* For real world data sets, it is often more expensive to compute $X'X$ than to find the k largest eigenvalues of it.

★ **SOLUTION:** True

53. [1 points] *True or False?* The most efficient method of computing the eigenvectors with the largest eigenvalues is to use the power method to find the 'largest' eigenvector (the one with the largest eigenvalue), project it off, and then repeat the process to find the second largest one, etc.

★ **SOLUTION:** False

54. [1 points] *True or False?* The most efficient method to compute the dominant principle components of a matrix X is to use a power method to find the eigenvectors of $X'X$ with the largest eigenvalues.

★ **SOLUTION:** False

55. [1 points] *True or False?* The right singular vectors of a square matrix, X, are equal to the eigenvectors of $X'X$.

★ **SOLUTION:** True

56. [1 points] *True or False?* Principle component regression (PCR) sets small eigenvalues of $X'X$ to zero, while Ridge regression shrinks them, but does not zero them out.

★ **SOLUTION:** True

57. [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

58. [2 points] Given data X with observations as rows and features as columns, if all features in X are uncorrelated, then "whitening" X has the following effect:

(a) Subtracting the mean observation (row) from each row of X.

(b) Subtracting the mean feature (column) from each column of X

(c) Scaling each observation by its norm

(d) Scaling each feature by its variance

(e) Scaling each feature by its standard deviation

★ **SOLUTION:** E

59. [1 points] *True or False?* Deep neural networks currently hold the records for best machine learning performance in problems ranging from speech and vision to natural language processing such as machine translation.

★ **SOLUTION:** True

60. [2 points] For a neural network, which of the following function types is NOT used to model neurons?

    (a) logistic
    (b) hyperbolic tangent
    (c) rectified linear unit
    (d) all of the above are used in neural networks

★ **SOLUTION:** D

61. [2 points] Autoencoders can be thought of as a non-linear generalization of...

    (a) PCA
    (b) ICA
    (c) K-means
    (d) A and B
    (e) A, B, and C

★ **SOLUTION:** D

62. [2 points] Convolutional neural networks for image analysis usually contain ...

    (a) Local receptive fields
    (b) Recurrent structures
    (c) Fully-connected layers
    (d) A and B
    (e) A and C

★ **SOLUTION:** E

63. [1 points] *True or False?* LSTMs are more complex models than Gated Recurrent Neural Nets (GRNNs).

★ **SOLUTION:** True

64. [1 points] *True or False?* LSTMs usually give better performance than Gated Recurrent Neural Nets (GRxNNs).

★ **SOLUTION:** False

65. [1 points] *True or False?* 'Deep' neural networks are usually optimized using a standard stochastic gradient method in which model parameters are updated after seeing each individual observation.

★ **SOLUTION:** False

66. [2 points] Which of the following is **not** true of drop-out.

(a) is a form of regularization
(b) is an ensemble method
(c) reduces the probability of being stuck in local minima
(d) all of the above are true.

★ **SOLUTION:** B

Consider the following confusion matrix

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

67. [1 points] For the above "confusion matrix" the precision is

(a) 2/10
(b) 8/20
(c) 12/33
(d) none of the above

★ **SOLUTION:** D

68. [1 points] For the above "confusion matrix" the recall is

(a) 2/10
(b) 8/20
(c) 12/33
(d) none of the above

★ **SOLUTION:** D

69. [1 points] *True or False?* L_1 loss (as a norm on the difference between y and $\hat{y}$) should in general not be used because solving such non-convex problems can be prohibitively slow.

★ **SOLUTION:** False

70. [1 points] *True or False?* L_2 loss (sometimes with a regularization penalty) is widely used because it usually reflects the actual loss function for applications in business and science.

★ **SOLUTION:** False

71. [2 points] In LDA on a set of documents written using a set of English words (the ”lexicon”), a ”topic” is...

    (a) a probability distribution over a proper subset of words in the lexicon (i.e. some of them)

    (b) a subset of the words in the lexicon

    (c) a probability distribution over all words in the lexicon

    (d) a set of subsets of the words in the lexicon

    (e) a set of rules for combining words in the lexicon

★ **SOLUTION:** C

72. [1 points] *True or False?* Belief nets constructed by interviewing experts tend to be simpler than those learned from data.

★ **SOLUTION:** True

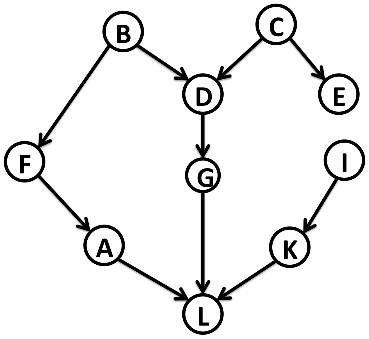
73. [1 points] *True or False?* When belief net structures are learned using the algorithm we used on the homework, different orderings of the variables can lead to belief nets with different numbers of parameters.

★ **SOLUTION:** True

74. [1 points] *True or False?* The EM algorithm is useful for **all** of the following: estimating Gaussian Mixture Models, estimating LDA models, and estimating Naive Bayes models with missing features (x) or class labels (y).

★ **SOLUTION:** True

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



75. [1 points] *True or False?* $(B \perp C | D)$

    ★ **SOLUTION:** False

76. [1 points] *True or False?* $(A \perp K | G)$

    ★ **SOLUTION:** True

77. [1 points] *True or False?* $(A \perp K | L)$

    ★ **SOLUTION:** False

78. [1 points] *True or False?* $(B \perp A | F, G)$

    ★ **SOLUTION:** True

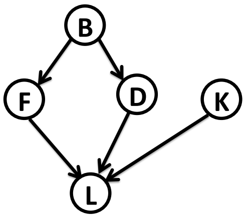
79. [1 points] *True or False?* $(F \perp I | G, L)$

    ★ **SOLUTION:** False

80. [1 points] *True or False?* G d-separates D and A

★ **SOLUTION:** False

The following question refers to this figure:



81. [2 points] What is the minimum number of parameters needed to represent the full joint distribution $P(B, D, F, K, L)$ in the above network if B takes on three possible values and all other variables are binary?
hint: Parameters here refer to the value of each probability; Do not include redundant probabilities like $P(X)$ and $P(\sim X)$.
   (a) < 10
   (b) 10-14
   (c) 15-19
   (d) 20-25
   (e) > 25

★ **SOLUTION:** C

82. [1 points] *True or False?* The emission matrix in an HMM represents the probability of the state, given an observation.

★ **SOLUTION:** False

83. [2 points] The local Markov assumption encoded by directed graphical models such as Bayes nets states...
   (a) A variable is conditionally independent of its non-descendants
   (b) A variable is conditionally independent of its non-descendants given only its parents
   (c) A variable is conditionally independent of its non-descendants given its parents and any other set of variables
   (d) A variable is conditionally independent of its descendants given its parents
   (e) A variable is never assumed to be conditionally independent of another variable no matter what is given

★ **SOLUTION:** B

84. [1 points] *True or False?* When deciding which points to label for a linear regression model, if one believes that the labels really were generated using the model (linear in x with Gaussian noise), it is best to label points that are at the extremes of the range of x, as opposed to being spread more evenly across the range of x.

★ **SOLUTION:** True

85. [1 points] *True or False?* When deciding which points to label for a binary classification problem, it is always best to label those points about which the classifier is *least* certain.

★ **SOLUTION:** False

86. [2 points] 'Big data' standardly refers to

(a) $n \gg p$

(b) $p \gg n$

(c) data that will not fit into memory

(d) none of the above; there is no standard definition

★ **SOLUTION:** D

87. [2 points] Over the past 50 years, computer power, as measured in megabytes storage or FLOPS (floating point operations) per $1,000 has doubled roughly every

(a) 12 months

(b) 18 months

(c) 24 months

(d) 36 months

(e) 48 months

★ **SOLUTION:** B

88. [1 points] Computers are more similar to human brains in terms of

(a) memory capacity (e.g. gigabytes)

(b) computational speed (e.g. instructions per second)

★ **SOLUTION:** B