UNIVERSITY OF PENNSYLVANIA

CIS 520: Machine Learning
Final, Fall 2018

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 single 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.*

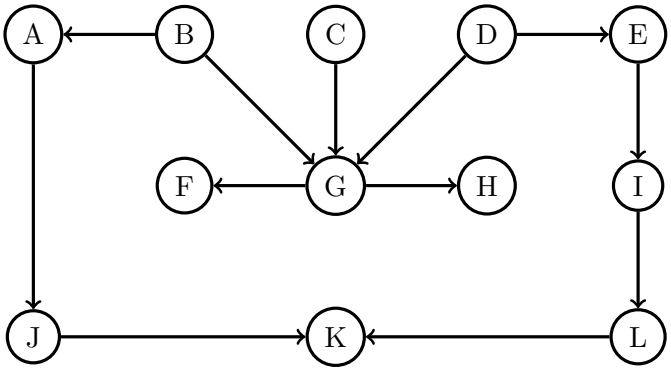
For the "TRUE or FALSE" questions, note that "TRUE" is (a) *and "FALSE" is* (b)*. For the multiple choice questions, select exactly one answer.*

The exam has 60 questions, totalling 98 points.

Name: _____

1. [1 points] *True or False?* Both PCA and linear regression can be thought of as algorithms for minimizing a sum of squared errors.
2. [1 points] *True or False?* The largest eigenvector of the covariance matrix is the direction of minimum variance in the data.
3. [1 points] *True or False?* The non-zero eigenvalues of $AA^\top$ and $A^\top A$ are the same.
4. [2 points] The left singular vectors of an arbitrary matrix A are:
   (a) Eigenvectors of A
   (b) Eigenvectors of $(A^\top A)^{-1}A^\top A$
   (c) Eigenvectors of $AA^\top$
   (d) Eigenvectors of $A^\top A$
5. [1 points] *True or False?* PCA is a type of linear autoencoder.
6. [1 points] *True or False?* A GAN may be trained via backpropogation alone.
7. [1 points] *True or False?* For $\mathbf{x} \in \mathbb{R}^d$, $f(\mathbf{x}) = \max\left(||\mathbf{x}||, \log(e^{x_1} + \cdots + e^{x_d})\right)$ is convex.
8. [1 points] *True or False?* $k(x, y) = \exp(-||x - y||)$ is a valid kernel.

The following seven questions refer to this figure:



9. [1 points] *True or False?* $C \perp\!\!\!\perp D \mid F$
10. [1 points] *True or False?* $D \perp\!\!\!\perp I \mid E, F, K$
11. [1 points] *True or False?* $C \perp\!\!\!\perp J \mid A, F, L$
12. [1 points] *True or False?* $F \perp\!\!\!\perp L \mid G$
13. [1 points] *True or False?* $\neg(G \perp\!\!\!\perp E \mid D, K)$
14. [1 points] *True or False?* I d-separates E and L

15. [2 points] What is the minimum number of parameters needed to represent the full joint probability $P(A, B, C, D, E, F, G, H, I, J, K, L)$ in the above network if all the variables are binary?

    (a) 4095
    (b) 20
    (c) 23
    (d) 24
    (e) 29

16. [2 points] Consider the following objective function for a GAN, where $G(\dot{})$ represents a generator that generates a p-dimensional example given a latent variable z drawn from $p(z)$, and $D(\dot{})$ is a discriminator that outputs a prediction for the probability a p-dimensional example has been drawn from the true dataset, which has density function $p_{data}(x)$.

$$V(G, D) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p(z)}[\log(1 - D(G(z)))]$$

    Which of the following statements about $V(G, D)$ is true?

    (a) The D is chosen to maximize $V(G, D)$, and G is chosen to minimize $V(G, D)$.
    (b) The G is chosen to maximize $V(G, D)$, and D is chosen to minimize $V(G, D)$.
    (c) The objective function is incorrect since the GAN formulation assumes z is p-dimensional.
    (d) The objective function is incorrect since the GAN formulation assumes z is completely random, rather than being drawn from some distribution.
    (e) None of the above.

17. [2 points] A real estate tycoon has employed you to assist with home sale negotiations. Your employer would like you to build a model to predict the counter-offer the opposing party will make in each round of bargaining given some features about the home and the values of counter-offers at all earlier rounds of bargaining. Suppose there are at most 4 rounds of bargaining. Which deep learning architecture best matches the structure of the problem?

    (a) Feed-forward network
    (b) standard RNN
    (c) GAN
    (d) LSTM

18. [2 points] You are processing the data of a survey where people have the option to report their income. We knowt that people with extremely high or low income are less likely to report their incomes. What is the best way to deal with the missing data?

    (a) Impute (replace) the missing data with the mode of the reported income
    (b) Impute (replace) missing data with the mean of the*mean* reported income
    (c) Replace the missing income with as "0" and add an extra column indicating whether or not the data is missing
    (d) Fill in the missing data with values randomly drawn from the reported values

19. [1 points] *True or False?* Consider an MDP (Markov decision process), $\mathcal{M} = \{S, A, p, r, \gamma\}$. If there are total $|S|$ states and $|A|$ possible actions, at each iteration, policy evaluation takes $O(|S|^2)$, while value iteration takes $O(|S|^2|A|)$.

20. [2 points] Suppose you are given a (fully specified) Markov decision process with state space $S = \{1, 2, 3\}$, and action space $A = \{a, b, c, d\}$. You calculate the optimal state-action value $Q^*(s, a)$ for each state-action pair (s, a) to be as follows:

	a	b	c	d
1	3.2	4.7	2.5	4.2
2	2.8	5.0	3.6	5.0
3	6.2	5.7	5.4	5.9

If we denote π^* as the optimal deterministic policy, which of the following **cannot** be true?

(a) $\pi^*(1) = a$

(b) $\pi^*(2) = b$

(c) $\pi^*(2) = d$

(d) $\pi^*(3) = a$

(e) None or more than one of the above

21. [2 points] Which of the following statements about Q-learning and Monte Carlo methods is true?

(a) Q-learning has higher bias and lower variance than Monte Carlo methods

(b) Q-learning has lower bias and higher variance than Monte Carlo methods

(c) Both q-learning and Monte Carlo methods are on-policy algorithms

(d) Both q-learning and Monte Carlo methods are off-policy algorithms

22. [1 points] *True or False?* On a given data set X which is mean centered, you divide each feature by its standard deviation so that the variance of each feature is 1. If you do PCA on the new standardized data set and obtain scores (i.e. the transformed output of PCA), then each of the scores will also have variance equal to 1.

The next two questions are about the following piece of pseudocode:

Algorithm 1 A Reinforcement Learning Algorithm

Initialize for all $s \in S, a \in A(s)$:
    $Q(s,a) \leftarrow$ arbitrary
    $Returns(s,a) \leftarrow$ empty list
    $\pi(a|s) \leftarrow$ arbitrary ϵ-soft policy
Repeat forever:
    Generate an episode using π
    For each pair of s, a appearing in the episode:
        $G \leftarrow$ the return that follows the first occurrence of s, a
        Append G to $Returns(s,a)$
        $Q(s,a) \leftarrow$ average$(Returns(s,a))$
    For each s in episode:
        $A^* \leftarrow \text{argmax}_a Q(s,a)$
        For all $a \in A(s)$, $\pi(a|s) =$
            $1 - \epsilon + \epsilon/|A(s)|$ if $a = A^*$
            $\epsilon/|A(s)|$ if $a \neq A^*$

23. [2 points] What type of reinforcement learning is it?

    (a) Temporal difference learning
    (b) Q-learning
    (c) Dynamic programming
    (d) Monte Carlo Method

24. [2 points] Which of the following categories does the above algorithm given fall into?

    (a) Off-policy
    (b) On-policy
    (c) Multi-armed bandit
    (d) None of the above

25. [2 points] Which of the following statements about AlphaGo is FALSE?

    (a) AlphaGo uses three policy networks: a fast-rollout network, a network trained via supervised learning, and a network trained via self-play
    (b) In the final policy, AlphaGo selects actions which have been taken most often in the Monte Carlo tree search, rather than those with the highest value estimations.
    (c) In the Monte Carlo tree search, the SL (supervised learning) policy network promotes exploitation and the value network promotes exploration.
    (d) During tree search, the fast-policy network traces out a path to the end of the game at each turn.

26. [2 points] You are given two-dimensional training data for PCR. The mean of the training data is $\langle 0, 3\rangle$, and the first principal component (loadings) is $\langle 1, 1\rangle$, (after subtracting off the mean, but not standardizing the data). You learn a model $\hat{y} = f(z) = 3z$ where z are the scores w.r.t the first PC. Given a test point $x = \langle 2, 3\rangle$ What is the prediction $\hat{y}$ for this point?

    Hint: $x = \langle 0, 3\rangle + \langle 1, 1\rangle + \langle 1, -1\rangle = \langle 2, 3\rangle$

    (a) 6
    (b) 3
    (c) 9
    (d) -3
    (e) None of the above

27. [2 points] Which of these models gives a globally optimum solution to the loss function it is minimizing?

    1) Logistic Regression
    2) Neural Networks
    3) K-means clustering

    (a) 1
    (b) 1 and 2
    (c) 3
    (d) All of these methods
    (e) None of these methods

    For the next two questions:
    Suppose you have a homogeneous Hidden Markov Model (i.e. transition and emission probabilities are independent of time; as always in this class). Each hidden state Z_t has K possible values and each observed variable X_t has M possible values. Also, suppose that you are given a sequence of observed variables $x_1, \ldots x_T$.

28. [1 points] *True or False?* For a given t, we have $X_s \perp Z_t$ for all $s < t$.

29. [1 points] *True or False?* The following statement about hidden Markov models holds for all $1 \leq t \leq T$ and k

$$\begin{aligned} P(X_{t+1} = x_{t+1}, \ldots, X_T = x_T \mid X_1 = x_1, \ldots, X_t = x_t, Z_t = k) \\ = P(X_{t+1} = x_{t+1}, \ldots, X_T = x_T \mid Z_t = k) \end{aligned}$$

For the next two questions:
You have a 2-dimensional training data set X_L of 100 instances, in which each feature 8 possible values, and a binary label $y = \pm 1$. You are asked to learn a Naive Bayes binary classification model for predicting the label y. You also found another data set X_U of 100 instances that are missing binary labels y. You want to use an EM algorithm to learn a better semi-supervised model by incorporating unlabelled instances, and treating unobserved labels as latent variables Z. Answer the following questions.

30. [1 points] *True or False?* The quantity $\gamma_j = P(Z_j = 1 \mid X_j = x_j)$ for an unlabelled instance $x_j \in X_U$, is a parameter of this EM model.

31. [2 points] What is the smallest number of parameters needed to specify a *model* for this classification using EM algorithm?

    (a) 15
    (b) 63
    (c) 115
    (d) 129
    (e) None of the above

32. [2 points] You are hired by Cambridge Analytica as a Machine Learning consultant. Your task is to use Facebook data of 100 million (10^8) people as training data to learn a classification model to predict the binary election vote for each person, represented by $y = \pm 1$. You decide to use regularized Logistic regression, which has the following penalized loss:

$$\min_{\mathrm{w}} \frac{1}{10^8} \sum_{i=1}^{10^8} \log(1 + \exp(-y_i \mathrm{w}^T \mathrm{X}_i)) + \lambda ||\mathrm{w}||_2^2$$

    Using cross-validation you find the best penalty hyperparameter λ_1. Later you learn that only 10 million of these people consented to this experiment, so as an ethical programmer, you decide learn a model using only 10 million people, and discard the rest. Using cross-validation again on this smaller data set you find the best penalty hyperparameter λ_2. Which of the following statements is **true**?

    (a) λ_2 is expected to be greater than λ_1
    (b) λ_2 is expected to be smaller than λ_1
    (c) $\lambda_2 \approx \lambda_1$
    (d) $10 \times \lambda_2 \approx \lambda_1$
    (e) None of the above

33. [2 points] Unfortunately, you got fired for your heroic stance, and your replacement, Mark, decides to use linear and degree 2 (quadratic) polynomial kernel SVM models trained on all of the 100 million people, instead of your Logistic regression model trained on 100 million people. Once these three models have been trained, Mark tests them by giving them a new voter to classify. Which of the following classifiers would be computationally **most** expensive to run?

    (a) Your Logistic regression model

(b) Mark's linear SVM model consisting of 1000 support vectors

(c) Mark's degree 2 polynomial kernel SVM model also consisting of 1000 support vectors

(d) Both b) and c) are equally more expensive in comparison to a)

(e) They are all equally computationally expensive

34. [2 points] You have just trained a logistic regression classifier which, given an instance x, estimates the probability of a positive label to be

$$\hat{\eta}(x) = \frac{1}{1 + e^{-\hat{w}^T x}}$$

(For simplicity, we ignore bias/threshold terms.) You are now told that the cost of a false positive (incorrectly predicting a negative example as positive) will be $\frac{3}{5}$, and that of a false negative will be $\frac{2}{5}$. In order to classify a new instance as positive or negative, what decision rule should be used?

(a) $h(x) = \text{sign}(\hat{w}^T x - \ln(3))$

(b) $h(x) = \text{sign}(\hat{w}^T x - \ln(\frac{2}{5}))$

(c) $h(x) = \text{sign}(\hat{w}^T x - \ln(\frac{3}{5}))$

(d) $h(x) = \text{sign}(\hat{w}^T x - \ln(\frac{2}{3}))$

(e) None of the above

35. [1 points] *True or False?* After $i-th$ iteration of online perceptron learning, you have a model h_i and you receive a new instance X_{i+1}. You find out that your current model misclassifies the instance as $h_i(X_{i+1}) = +1$ when you receive the actual label $Y_{i+1} = -1$. You update the model using the perceptron algorithm and get a classifier h_{i+1}. h_{i+1} is guaranteed to classify X_{i+1} correctly as -1?

36. [2 points] You have a corpus of documents on which you want to implement LDA topic modelling. Which of the following statements is **true**?

(a) LDA topic models assign a single topic to each document

(b) LDA topic models assign each word to a single topic

(c) LDA topic models contain parameters for the transition probabilities between topics

(d) Unlike Part of Speech (POS) tagging using HMMs, LDA models treat words in a document as being conditionally independent given a latent variable

(e) None of the above

37. [2 points] Which of the following statements about AdaBoost algorithm for binary classification is **true**?

1) Training error is guaranteed to approach zero as the number of iteration tends to ∞

2) AdaBoost should ideally use an underfit model as the "weak learners"

3) AdaBoost should ideally use an overfit model as as the "weak learners"

(a) 1 only

(b) 2 only

(c) 1 and 2

(d) 3 only

(e) 1 and 3

38. [2 points] For which of the following models, does the complexity increase as the given hyper-parameter increases? (Assume all other hyper-parameters stay constant).

    (a) Decision trees; minimum number of instances required in a node

    (b) Neural Networks; L_2 penalty coefficient

    (c) k-Nearest Neighbors; k (number of neighbors)

    (d) Gaussian Mixture Models; number of Gaussians

    (e) None of the above

39. [2 points] You are using an SVM with an RBF kernel defined as $e^{\frac{-||x||_2^2}{\sigma^2}}$ for a classification problem. You find that the training accuracy is 0.97 but the test accuracy is 0.65. Which of the following measures is most likely to improve the test accuracy?

    1) Increasing the kernel width σ

    2) Decreasing the kernel width σ

    3) Using a polynomial kernel instead of an RBF

    (a) 1

    (b) 2

    (c) 3

    (d) 1 and 3

    (e) 2 and 3

40. [2 points] You are training a simple neural network for a regression problem on a 2-dimensional data set. Your Neural Net architecture is as follows: 3 hidden layers with sigmoid units, trained for 1000 epochs, with L_2 penalty for each hidden layer. Using 5-fold cross-validation you learn that the 1st hidden layer should have 6 neurons, the 2nd hidden layer should have 4 neurons and the 3rd hidden layer should have 3 neurons. However, you find that the test error is 10 times the training error. Which of the following changes is most likely to bring the biggest improvement in performance?

    (a) Doing 10-fold cross-validation

    (b) Implementing early stopping

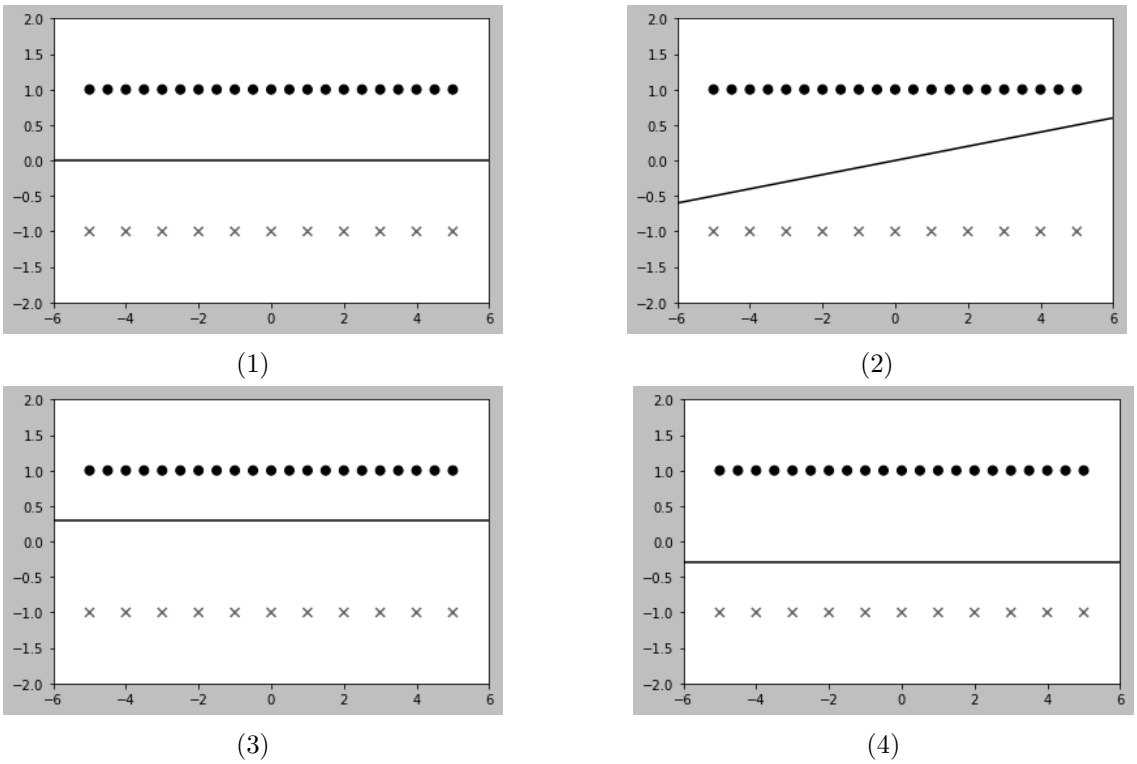
    (c) Adding a fourth hidden layer

    (d) Using ReLU activations instead of sigmoid

    (e) Using an L_1 penalty instead of L_2

For the next two questions:

A 2-dimensional training data set contains two labels, denoted by the 20 circles and 10 crosses below. The figures show possible decision boundaries for this data.



(1) (2)

(3) (4)

Figure 1: Logistic regression decision boundaries

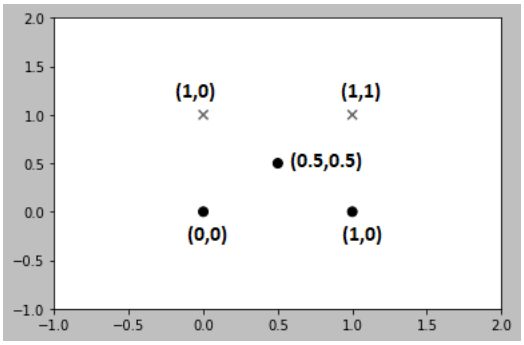
41. [2 points] You want to fit an **unregularized** Logistic regression model to determine the decision boundary, which is a line in this case. Which of the following figures shows the decision boundary line produced by the model?

    (a) Figure 1
    (b) Figure 2
    (c) Figure 3
    (d) Figure 4
    (e) All figures are valid

42. [2 points] Now you want to fit a L_2 **regularized** Logistic regression model to determine the decision boundary, which is also a line in this case. Which of the following figures cannot be a decision boundary for this model?

    (a) Figure 1
    (b) Figure 3
    (c) Figure 4
    (d) All figures are valid

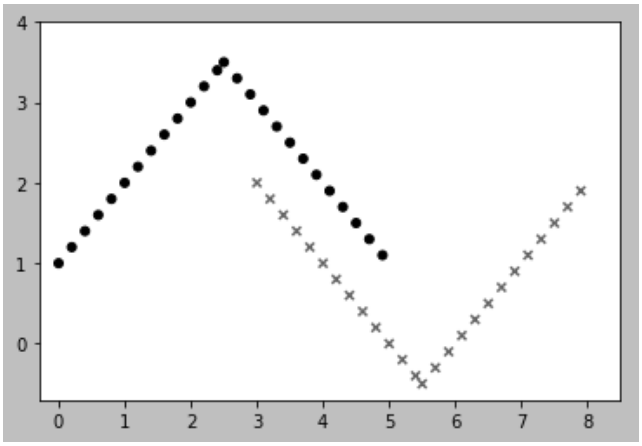
43. [1 points] The following data set consists of 5 points: each corner of a unit square and its center. Can this data set be made separable by an SVM with an RBF kernel using **only two** support vectors? (There is no restriction on the kernel width or choice of support vectors.)



(a) True

(b) False

44. [2 points] The following training set consists of binary labeled points. You want to train a Neural Net model on this data. If you use only one hidden layer with ReLU activation units, what is the smallest number of activation units required to separate this training set?



(a) 1

(b) 2

(c) 3

(d) More than 3

(e) It cannot be separated using only one hidden layer of any number of ReLU units

45. [2 points] Consider an active learning setup for a cost-sensitive binary classification with labels $\{\pm 1\}$. The loss matrix is:

	$\hat{y} = +1$	$\hat{y} = -1$
$y = +1$	0	2
$y = -1$	6	0

For any instance x, let $\eta(x) = P(Y = +1|X = x)$ denote the conditional probability that the true label is $+1$ given x. 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, $\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_2}, \mathbf{x_4}$, 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:

$$\hat{\eta}(\mathbf{x_1}) = 0.77, \hat{\eta}(\mathbf{x_2}) = 0.49, \hat{\eta}(\mathbf{x_3}) = 0.26, \hat{\eta}(\mathbf{x_4}) = 0.67$$

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

46. [1 points] Suppose you are given a binary labelled data set that is linearly separable. Using an SVM, you find a hyperplane H that separates the labels with maximum margin γ. Is it possible that there is another hyperplane, different from H, that also separates the labels with the same margin γ?

(a) Yes
(b) No

47. [2 points] When $p >> n$, which of the following methods can we **not** use to train the model? (As usual, data dimension is p and training sample size is n.)

(a) Do sparse regularization such as lasso
(b) use semi-supervised learning (if the data are available)
(c) Use dimensionality reduction
(d) All of the above can reasonably be used.

48. [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) Both

(d) None of the above

49. [2 points] Which of the following best describes the M-step of EM algorithm?

    (a) Assign values to the hidden variables
    (b) Assign probabilities to the hidden variables
    (c) Estimate the parameters of the model
    (d) Calculate the complete data log-likelihood
    (e) None of the above

50. [2 points] Which of the following is **not** best estimated using the EM algorithm.

    (a) HMMs
    (b) Gaussian Mixture Models
    (c) Belief nets where not all variables are observed.
    (d) Model-based reinforcement learning

51. [1 points] *True or False?* When deciding which points (observations) to get labels for, picking the point about which one is most uncertain will reliably lead to good coverage of the feature space.

52. [2 points] We studied a number of active learning methods. Which of the following is **not** among them?

    (a) query by majority
    (b) Monte Carlo sampling
    (c) (a) and (b)
    (d) label the most uncertain point
    (e) label the point that will most change the model

53. [2 points] Which of the following is **not** a valid method of computing the square of the Frobeneous norm of a square, symmetric matrix?

    (a) sum of the squares of the eigenvalues of the matrix
    (b) sum of the squares of the matrix entries

$$\sum_{ij} x_{ij}^2$$

    (c) square of the sum of the absolute values of the matrix entries,

$$(\sum_{ij} |x_{ij}|)^2$$

54. [2 points] Most methods of measuring variable importance (e.g. like we saw for random forests) are designed

    (a) to roughly approximate how large the effect on the output would be if that feature changes in the real world.

(b) to roughly approximate how large the effect on the output would be if the feature were removed from the model.

(c) The above two answers are the same, so both of them are correct.

55. [2 points] LIME (*pick the best single answer*):

(a) explains which features in a model are most important (across all points)

(b) explains which features in a model are most important for predicting at a particular point in the training data

(c) explains which features in a model are most important for predicting at any particular point

56. [2 points] For which of the following situations is mean centering the data before doing PCA probably a good thing to do? In each case the rows of the matrix are people.

(a) items purchased from a large company

(b) counts of words in a person's emails

(c) movie ratings

(d) medical record (age, sex, weight, BMI, blood pressure, glucose level, and five similar items)

(e) (a), (b) and (c).

57. [2 points] Which is most greedy?

(a) stagewise regression with stepwise search

(b) regular regression with stepwise search

(c) stagewise regression with streaming (in features) search

(d) regular regression with streaming (in features) search

(e) the question doesn't make sense; you can't combine stagewise regression with streamwise or stepwise search

58. [2 points] Consider the error decomposition for a least squares regression model

$$\mathbf{E}_{x,y,D}[(h(x;D)-y)^2] = \underbrace{\mathbf{E}_{x,D}[(h(x;D)-\overline{h}(x))^2]}_{\text{Variance}} + \underbrace{\mathbf{E}_{x}[(\overline{h}(x)-\overline{y}(x))^2]}_{\text{Bias}^2} + \underbrace{\mathbf{E}_{x,y}[(\overline{y}(x)-y)^2]}_{\text{Noise}}$$

where $h(x;D)$ is a model learned over a training sample D, $\overline{h}(x) = \mathbf{E}_{D\sim P^n}[(h(x;D)]$ is the average model, and $\overline{y}(x) = \mathbf{E}_{y|x}[y]$ is the optimal Bayes model. Which of the following best describes the term labeled *variance*?

(a) On average, how much your learned model differs from average model across different samples D

(b) How far is the average model from optimal Bayes model

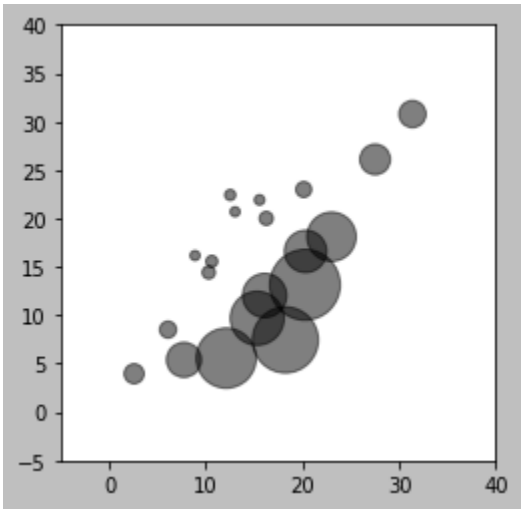
(c) The variance between predictions for a fixed sample D

(d) How accurate the model is in predicting y

59. [2 points] Suppose you are learning a CNN on greyscale images of size 105×154, so the image has only one channel. In the first convolutional layer, you use a filter of size 21×14 with stride of size 7 in both x and y dimensions without any padding or bias term. How many neurons will there be in the next layer?

    (a) 12*20
    (b) 13*21
    (c) 15*22
    (d) 16*23
    (e) None of the above

60. [2 points] Suppose you have a two dimensional training data set X with real valued labels Y. The following plot shows training data; each element X_i of the training set is the center of a circle and the radius of the circle equals its label Y_i. (Both axes have the same scale.)



    You want to reduce the data to one dimension by projecting onto a suitable direction, and then learn a linear regression model in the reduced space, i.e. do PCR *using only a single component*. In order to be able to accurately predict labels of as many of the training points as possible, which of the following projections would be best?

    (a) Projection onto the x1-axis
    (b) Projection onto the x2-axis
    (c) Projection onto the 1st principal component
    (d) Projection onto the 2nd principal component
    (e) All are equally good