Name and CU email address: ______________________________

Midterm Exam | CSCI 4622: Machine Learning
Fall 2019 | Instructor: C. Monteleoni

This exam has 7 questions, for a total of 33 points and 2 bonus points. Question 7 is extra credit; its bonus points can be used to increase your score. Note: a total score of 33 will be considered 100%.

Definitions:

- $|a|$ denotes the absolute value of scalar a.

For a vector $x \in \mathbb{R}^d$, the following norms are defined as follows:

- L2 norm: Note: when there is no subscript to the norm, we will assume it is L2.

$$\|\vec{x}\| = \|\vec{x}\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2}$$

- L1 norm:

$$\|\vec{x}\|_1 = \sum_{i=1}^{d} |x_i|$$

For grading. Please do not write here:

Question	Points	Bonus Points	Score
1	6	0	
2	4	0	
3	4	0	
4	6	0	
5	7	0	
6	6	0	
7	0	2	
Total:	33	2	

1. Total for Question 1: 6

For each of the following statements, check the box indicating whether the statement is True or False.

(a) (1 point) Consider a neural network with a single hidden layer, so that $h_i = f(\vec{w}_i \cdot \vec{x})$, and $\hat{y} = \sum_i u_i h_i$, where each u_i is a real number. Suppose that the activation function f is the *identity function* (i.e., $f(a) = a$). This neural network can represent nonlinear decision boundaries.
☐ True $\surd$ **False**

(b) (1 point) The Backpropagation algorithm will always find the global minimum of the loss function, regardless of how the neural network's weights were initialized.
☐ True $\surd$ **False**

(c) (1 point) There exists a neural network, with a single hidden layer, that can correctly compute the XOR function on binary inputs.
$\surd$ **True** ☐ False

(d) (1 point) Adding a regularization penalty of squared L2 norm, on the parameter vector $\vec{w}$, when estimating a logistic regression model, guarantees that some of the parameters (weights, w_i, associated with the components of the input vectors) will be zero.
☐ True $\surd$ **False**

(e) (1 point) $K(x, z) = \exp\{-\frac{1}{17}\|x - z\|^2\}$ is a valid kernel.
$\surd$ **True** ☐ False

(f) (1 point) If K is a valid kernel and $\alpha > 0$ is a real value, then $G(x, z) = \alpha K(x, z)$ is also a valid kernel.
$\surd$ **True** ☐ False

2. (4 points) Total for Question 2: 4

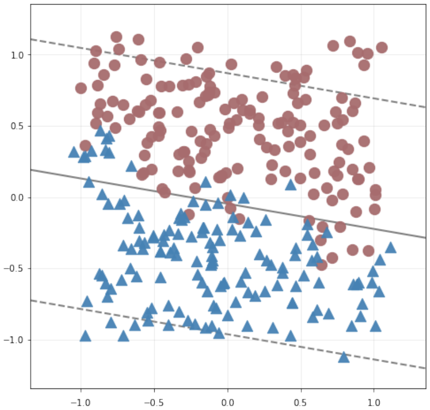
Circle **all** choices that are likely to *reduce overfitting*.

A. Adding a term proportional to $\|\vec{w}\|_2$ to a minimization, over parameter vectors $w \in \mathbb{R}^d$, of empirical losses.

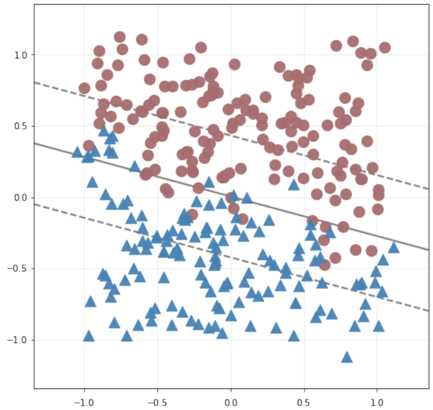
B. Adding additional edges to a Bayes Net.

C. Learning a forest (an ensemble) of decision trees, instead of a single decision tree.

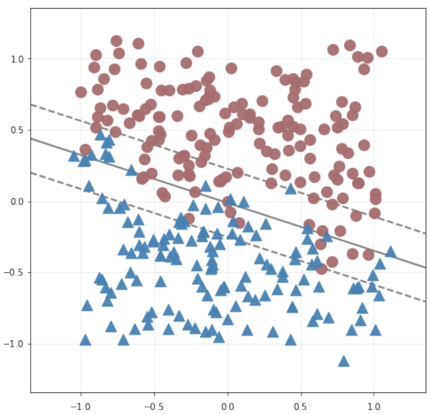
D. Adding a term proportional to $\|\vec{w}\|_1$ to a minimization, over parameter vectors $w \in \mathbb{R}^d$, of empirical losses.



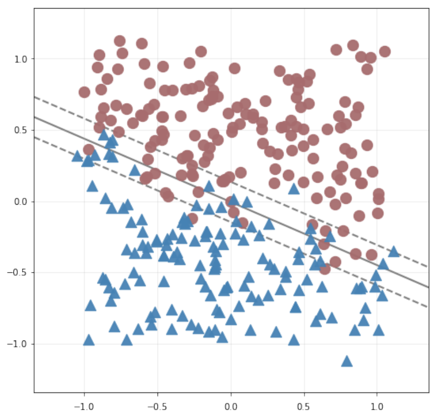
(a)



(b)



(c)



(d)

3. Total for Question 3: 4

(a) (2 points) Which (one) of the above figures is associated with the classifier that uses the highest value for C in soft-margin SVM?

A. (a)
B. (b)
C. (c)
D. (d)

(b) (2 points) Which (one) of the above figures is associated with the classifer with the most support vectors?

A. (a)
B. (b)
C. (c)
D. (d)

4. (6 points) Total for Question 4: 6

For each of these statements, fill in the blank with a one-to-three word phrase or a short math expression. Be as specific as possible. For instance, given **"A depth k decision tree queries at most $\boxed{\textbf{all}}$ features"**, an answer of "all" is incorrect (it's insufficiently specific). Each box is worth 1 point (some items have more than one box).

(a) A soft-margin SVM (objective: $\frac{1}{2}\|\vec{w}\|^2 + C\sum_n \xi_n$) is likely to underfit as C tends toward $\boxed{0}$.

(b) The L1 norm regularization penalty on the parameter vector, $\vec{w}$, encourages the resulting parameter vector to be $\boxed{\text{sparse}}$

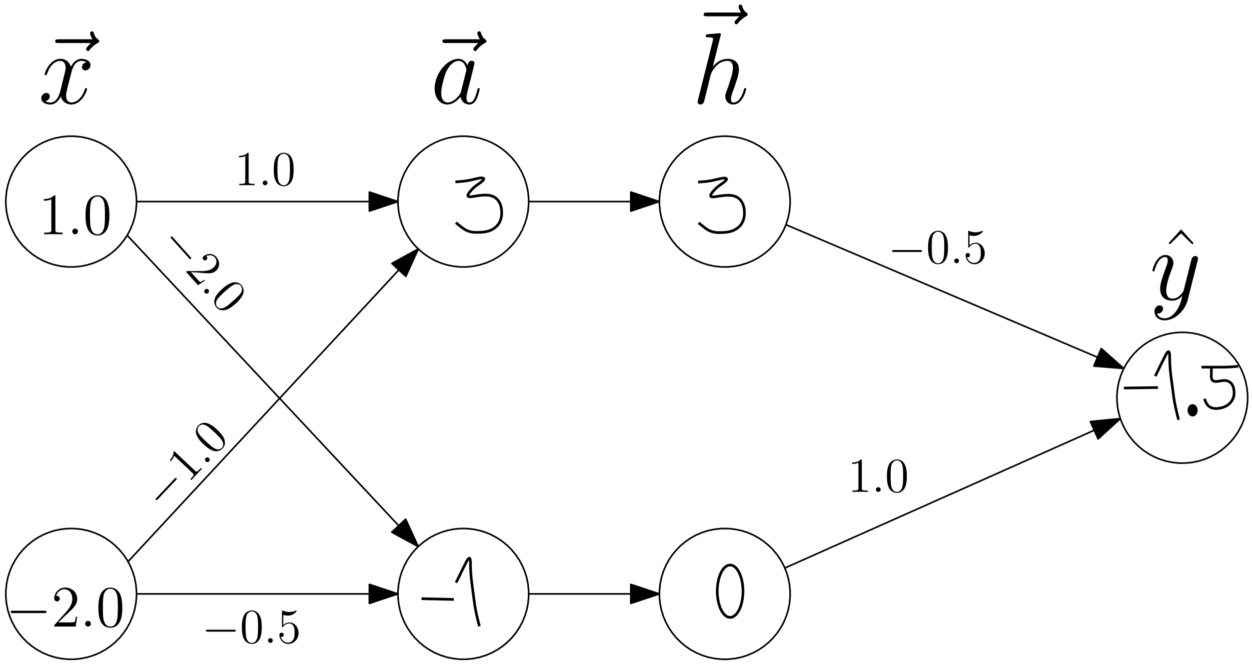
(c) Consider a polynomial degree three feature expansion, $\phi(\vec{x})$. If the original data, $\vec{x}$, is in D dimensions, using a kernel function rather than explicit dot products between $\vec{\phi}$ vectors reduces the computational complexity of the dot product from $\mathcal{O}(\boxed{D^3})$ to
$\mathcal{O}(\boxed{D})$.

(d) We replace 0/1 loss with a surrogate loss function (hinge, logistic, etc.) because these surrogates are $\boxed{\text{convex}}$ and 0/1 loss is not.

(e) For linear models, having a large margin is equivalent to having small $\boxed{\|\vec{w}\|^2}$.

5. Total for Question 5: 7

Consider the neural network shown below:



There are two input features, which get fed into two activation functions, from which two hidden-layer values are computed. For each hidden node, a is its input, and h is its output after applying the activation function to the value a. (In this network there are no offset parameters).

(a) (1 point) How many parameters (weights) does this network have?

(a) **6**

(b) (5 points) Execute forward propagation on this network, writing the appropriate values in the nodes in the graph above. Assume that the activation functions (applied between $\vec{a}$ and $\vec{h}$) are rectified linear units, i.e., ReLU: zero if input is negative, otherwise the identity function.

(c) (1 point) Is this network being used for a regression task? (Hint: look at the value of $\hat{y}$, the output of the network, that you computed above.)

A. Yes

B. No

6. (6 points) Total for Question 6: 6

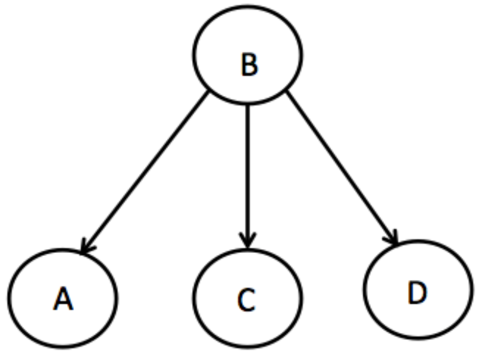
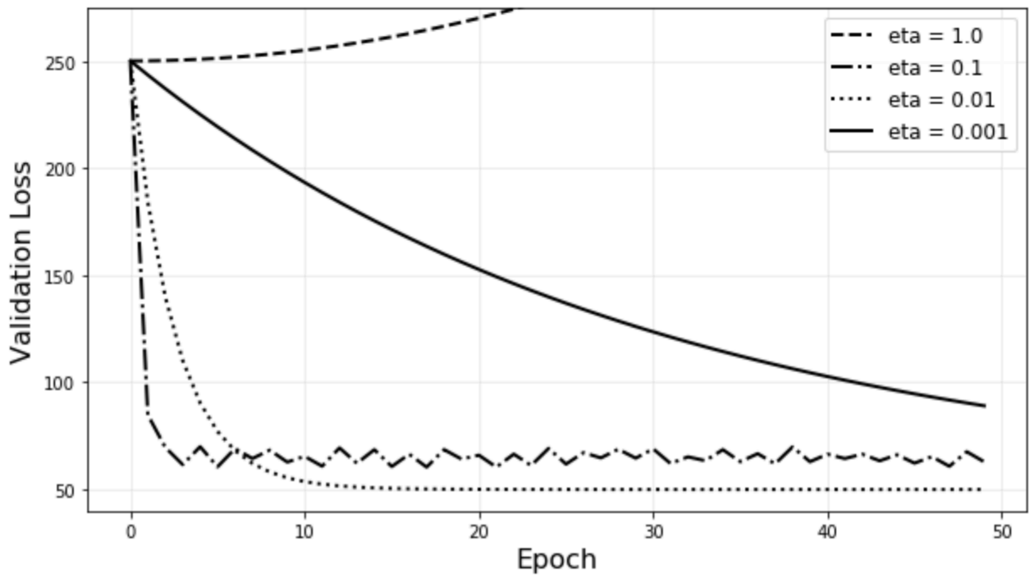


Figure 2: A Bayes Net.

Figure 2 shows a Bayes Net; nodes indicate random variables and edges indicate conditional probability distributions. Circle **all** of the following choices corresponding to True statements: [NOTE: correct answers are in **bold**.]

1. **The following is implied by Figure 2:** $P(A, C \mid B) = P(A \mid B)P(C \mid B)$**.**
2. The following is implied by Figure 2: $P(A, D \mid C) = P(A \mid C)P(D \mid C)$.
3. The following is implied by Figure 2: $P(A \mid C) = P(A)$.
4. **If B is a discrete-valued random variable whose values are interpreted as classification labels, and A, C, D are discrete-valued random variables whose values are interpreted as the features of a data point, then Figure 2 can be interpreted as a Naive Bayes model.**
5. If A and B are discrete-valued random variables, one possible distribution for $P(A \mid B = b_1) = \mathcal{N}(\mu_1, \sigma_1^2 I)$, a spherical Gaussian with parameters μ_1, σ_1^2.
6. **If B and C are binary-valued random variables, one possible conditional distribution can be specified as follows:** $P(C = c_1 \mid B = b_1) = 0.6$**,** $P(C = c_1 \mid B = b_2) = 0.1$**.**

7. Total for Question 7: 2 (bonus)



(a) (2 points (bonus)) This figure plots the loss function of a model evaluated on the validation data set, using Stochastic Gradient Descent with various choices of learning rate (step-size), "`eta`." Which value of the learning rate performs the best in this plot?

`A. eta = 1.0`

`B. eta = 0.1`

`C. eta = 0.01`

`D. eta = 0.001`