Learning From Data
Yaser Abu-Mostafa, *Caltech*
http://work.caltech.edu/telecourse
Self-paced version

Homework # 1

*All questions have multiple-choice answers (***[a], [b], [c], ...***). You can collaborate with others, but do not discuss the selected or excluded choices in the answers. You can consult books and notes, but not other people's solutions. Your solutions should be based on your own work. Definitions and notation follow the lectures.*

Note about the homework

- The goal of the homework is to facilitate a deeper understanding of the course material. The questions are not designed to be puzzles with catchy answers. They are meant to make you roll up your sleeves, face uncertainties, and approach the problem from different angles.

- The problems range from easy to difficult, and from practical to theoretical. Some problems require running a full experiment to arrive at the answer.

- The answer may not be obvious or numerically close to one of the choices, but one (and only one) choice will be correct if you follow the instructions precisely in each problem. You are encouraged to explore the problem further by experimenting with variations on these instructions, for the learning benefit.

- You are also encouraged to take part in the forum

  http://book.caltech.edu/bookforum

  where there are many threads about each homework set. We hope that you will contribute to the discussion as well. Please follow the forum guidelines for posting answers (see the "BEFORE posting answers" announcement at the top there).

• The Learning Problem

1. What types of Machine Learning, if any, best describe the following three scenarios:

(i) A coin classification system is created for a vending machine. The developers obtain exact coin specifications from the U.S. Mint and derive a statistical model of the size, weight, and denomination, which the vending machine then uses to classify coins.

(ii) Instead of calling the U.S. Mint to obtain coin information, an algorithm is presented with a large set of labeled coins. The algorithm uses this data to infer decision boundaries which the vending machine then uses to classify its coins.

(iii) A computer develops a strategy for playing Tic-Tac-Toe by playing repeatedly and adjusting its strategy by penalizing moves that eventually lead to losing.

[**a**] (i) Supervised Learning, (ii) Unsupervised Learning, (iii) Reinforcement Learning

[**b**] (i) Supervised Learning, (ii) Not learning, (iii) Unsupervised Learning

[**c**] (i) Not learning, (ii) Reinforcement Learning, (iii) Supervised Learning

[**d**] (i) Not learning, (ii) Supervised Learning, (iii) Reinforcement Learning

[**e**] (i) Supervised Learning, (ii) Reinforcement Learning, (iii) Unsupervised Learning

2. Which of the following problems are best suited for Machine Learning?

(i) Classifying numbers into primes and non-primes.

(ii) Detecting potential fraud in credit card charges.

(iii) Determining the time it would take a falling object to hit the ground.

(iv) Determining the optimal cycle for traffic lights in a busy intersection.

[**a**] (ii) and (iv)

[**b**] (i) and (ii)

[**c**] (i), (ii), and (iii)

[**d**] (iii)

[**e**] (i) and (iii)

- **Bins and Marbles**

3. We have 2 opaque bags, each containing 2 balls. One bag has 2 black balls and the other has a black ball and a white ball. You pick a bag at random and then pick one of the balls in that bag at random. When you look at the ball, it is black. You now pick the second ball from that same bag. What is the probability that this ball is also black?

[**a**] 1/4

[**b**] 1/3

[**c**] 1/2

[**d**] 2/3

[**e**] 3/4

Consider a sample of 10 marbles drawn from a bin containing red and green marbles. The probability that any marble we draw is red is $\mu = 0.55$ (independently, with replacement). We address the probability of getting no red marbles ($\nu = 0$) in the following cases:

4. We draw only one such sample. Compute the probability that $\nu = 0$. The closest answer is ('closest answer' means: $|$your answer $-$ given option$|$ is closest to 0):

[**a**] 7.331×10^{-6}

[**b**] 3.405×10^{-4}

[**c**] 0.289

[**d**] 0.450

[**e**] 0.550

5. We draw 1,000 independent samples. Compute the probability that (at least) one of the samples has $\nu = 0$. The closest answer is:

[**a**] 7.331×10^{-6}

[**b**] 3.405×10^{-4}

[**c**] 0.289

[**d**] 0.450

[**e**] 0.550

● Feasibility of Learning

Consider a Boolean target function over a 3-dimensional input space $\mathcal{X} = \{0, 1\}^3$ (instead of our ±1 binary convention, we use 0,1 here since it is standard for Boolean functions). We are given a data set $\mathcal{D}$ of five examples represented in the table below, where $y_n = f(\mathbf{x}_n)$ for $n = 1, 2, 3, 4, 5$.

	$\mathbf{x}_n$		y_n
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	1

Note that in this simple Boolean case, we can enumerate the entire input space (since there are only $2^3 = 8$ distinct input vectors), and we can enumerate the set of all possible target functions (there are only $2^{2^3} = 256$ distinct Boolean function on 3 Boolean inputs).

Let us look at the problem of learning f. Since f is unknown except inside $\mathcal{D}$, any function that agrees with $\mathcal{D}$ could conceivably be f. Since there are only 3 points in $\mathcal{X}$ outside $\mathcal{D}$, there are only $2^3 = 8$ such functions.

The remaining points in $\mathcal{X}$ which are not in $\mathcal{D}$ are: 101, 110, and 111. We want to determine the hypothesis that agrees the most with the possible target functions. In order to quantify this, count how many of the 8 possible target functions agree with each hypothesis on all 3 points, how many agree on just 2 of the points, on just 1 point, and how many do not agree on any points. The final score for each hypothesis is computed as follows:

Score = *(# of target functions agreeing with hypothesis on all* 3 *points)*$\times 3$ + *(# of target functions agreeing with hypothesis on exactly* 2 *points)*$\times 2$ + *(# of target functions agreeing with hypothesis on exactly* 1 *point)*$\times 1$ + *(# of target functions agreeing with hypothesis on* 0 *points)*$\times 0$.

6. Which hypothesis g agrees the most with the possible target functions in terms of the above score?

[**a**] g returns 1 for all three points.

[**b**] g returns 0 for all three points.

[**c**] g is the XOR function applied to $\mathbf{x}$, i.e., if the number of 1s in $\mathbf{x}$ is odd, g returns 1; if it is even, g returns 0.

[**d**] g returns the opposite of the XOR function: if the number of 1s is odd, it returns 0, otherwise returns 1.

[**e**] They are all equivalent (equal scores for g in [**a**] through [**d**]).

- **The Perceptron Learning Algorithm**

In this problem, you will create your own target function f and data set $\mathcal{D}$ to see how the Perceptron Learning Algorithm works. Take $d = 2$ so you can visualize the problem, and assume $\mathcal{X} = [-1, 1] \times [-1, 1]$ with uniform probability of picking each $\mathbf{x} \in \mathcal{X}$.

In each run, choose a random line in the plane as your target function f (do this by taking two random, uniformly distributed points in $[-1, 1] \times [-1, 1]$ and taking the line passing through them), where one side of the line maps to $+1$ and the other maps to -1. Choose the inputs $\mathbf{x}_n$ of the data set as random points (uniformly in $\mathcal{X}$), and evaluate the target function on each $\mathbf{x}_n$ to get the corresponding output y_n.

Now, in each run, use the Perceptron Learning Algorithm to find g. Start the PLA with the weight vector $\mathbf{w}$ being all zeros (consider $\text{sign}(0) = 0$, so all points are initially misclassified), and at each iteration have the algorithm choose a point randomly from the set of misclassified points. We are interested in two quantities: the number of iterations that PLA takes to converge to g, and the disagreement between f and g which is $\mathbb{P}[f(\mathbf{x}) \neq g(\mathbf{x})]$ (the probability that f and g will disagree on their classification of a random point). You can either calculate this probability exactly, or approximate it by generating a sufficiently large, separate set of points to estimate it.

In order to get a reliable estimate for these two quantities, you should repeat the experiment for 1000 runs (each run as specified above) and take the average over these runs.

7. Take $N = 10$. How many iterations does it take on average for the PLA to converge for $N = 10$ training points? Pick the value closest to your results (again, 'closest' means: |your answer – given option| is closest to 0).

[**a**] 1

[**b**] 15

[**c**] 300

[**d**] 5000

[**e**] 10000

8. Which of the following is closest to $\mathbb{P}[f(\mathbf{x}) \neq g(\mathbf{x})]$ for $N = 10$?

[**a**] 0.001

[**b**] 0.01

[**c**] 0.1

[**d**] 0.5

[**e**] 0.8

9. Now, try $N = 100$. How many iterations does it take on average for the PLA to converge for $N = 100$ training points? Pick the value closest to your results.

[**a**] 50

[**b**] 100

[**c**] 500

[**d**] 1000

[**e**] 5000

10. Which of the following is closest to $\mathbb{P}[f(\mathbf{x}) \neq g(\mathbf{x})]$ for $N = 100$?

[**a**] 0.001

[**b**] 0.01

[**c**] 0.1

[**d**] 0.5

[**e**] 0.8

Learning From Data
Yaser Abu-Mostafa, *Caltech*
http://work.caltech.edu/telecourse
Self-paced version

Answer Key To Homework # 1

1. **[d]**
2. **[a]**
3. **[d]**
4. **[b]**
5. **[c]**
6. **[e]**
7. **[b]**
8. **[c]**
9. **[b]**
10. **[b]**

Visit the forum (http://book.caltech.edu/bookforum) for discussion. Please follow the forum guidelines for posting answers (see the "BEFORE posting answers" announcement at the top thread there).

Learning From Data
Yaser Abu-Mostafa, *Caltech*
http://work.caltech.edu/telecourse
Self-paced version

Homework # 2

*All questions have multiple-choice answers (***[a], [b], [c], ...***). You can collaborate with others, but do not discuss the selected or excluded choices in the answers. You can consult books and notes, but not other people's solutions. Your solutions should be based on your own work. Definitions and notation follow the lectures.*

Note about the homework

- The goal of the homework is to facilitate a deeper understanding of the course material. The questions are not designed to be puzzles with catchy answers. They are meant to make you roll up your sleeves, face uncertainties, and approach the problem from different angles.

- The problems range from easy to difficult, and from practical to theoretical. Some problems require running a full experiment to arrive at the answer.

- The answer may not be obvious or numerically close to one of the choices, but one (and only one) choice will be correct if you follow the instructions precisely in each problem. You are encouraged to explore the problem further by experimenting with variations on these instructions, for the learning benefit.

- You are also encouraged to take part in the forum

  http://book.caltech.edu/bookforum

  where there are many threads about each homework set. We hope that you will contribute to the discussion as well. Please follow the forum guidelines for posting answers (see the "BEFORE posting answers" announcement at the top there).

• Hoeffding Inequality

Run a computer simulation for flipping 1,000 virtual fair coins. Flip each coin independently 10 times. Focus on 3 coins as follows: c_1 is the first coin flipped, c_{rand} is a coin chosen randomly from the 1,000, and c_{min} is the coin which had the minimum frequency of heads (pick the earlier one in case of a tie). Let ν_1, ν_{rand}, and ν_{min} be the *fraction* of heads obtained for the 3 respective coins out of the 10 tosses.

Run the experiment 100,000 times in order to get a full distribution of ν_1, ν_{rand}, and ν_{min} (note that c_{rand} and c_{min} will change from run to run).

1. The average value of ν_{min} is closest to:

[**a**] 0

[**b**] 0.01

[**c**] 0.1

[**d**] 0.5

[**e**] 0.67

2. Which coin(s) has a distribution of ν that satisfies the (single-bin) Hoeffding Inequality?

[**a**] c_1 only

[**b**] c_{rand} only

[**c**] c_{min} only

[**d**] c_1 and c_{rand}

[**e**] c_{min} and c_{rand}

• Error and Noise

Consider the bin model for a hypothesis h that makes an error with probability μ in approximating a deterministic target function f (both h and f are binary functions). If we use the same h to approximate a noisy version of f given by:

$$P(y \mid \mathbf{x}) = \begin{cases} \lambda & y = f(x) \\ 1 - \lambda & y \neq f(x) \end{cases}$$

3. What is the probability of error that h makes in approximating y? *Hint: Two wrongs can make a right!*

[a] μ

[b] λ

[c] 1-μ

[d] $(1-\lambda) * \mu + \lambda * (1-\mu)$

[e] $(1-\lambda) * (1-\mu) + \lambda * \mu$

4. At what value of λ will the performance of h be independent of μ?

[a] 0

[b] 0.5

[c] $1/\sqrt{2}$

[d] 1

[e] No values of λ

• Linear Regression

In these problems, we will explore how Linear Regression for classification works. As with the Perceptron Learning Algorithm in Homework # 1, you will create your own target function f and data set $\mathcal{D}$. Take $d = 2$ so you can visualize the problem, and assume $\mathcal{X} = [-1,1] \times [-1,1]$ with uniform probability of picking each $\mathbf{x} \in \mathcal{X}$. In each run, choose a random line in the plane as your target function f (do this by taking two random, uniformly distributed points in $[-1,1] \times [-1,1]$ and taking the line passing through them), where one side of the line maps to $+1$ and the other maps to -1. Choose the inputs $\mathbf{x}_n$ of the data set as random points (uniformly in $\mathcal{X}$), and evaluate the target function on each $\mathbf{x}_n$ to get the corresponding output y_n.

5. Take $N = 100$. Use Linear Regression to find g and evaluate E_{in}, the fraction of in-sample points which got classified incorrectly. Repeat the experiment 1000 times and take the average (keep the g's as they will be used again in Problem 6). Which of the following values is closest to the average E_{in}? (*Closest* is the option that makes the expression |your answer − given option| closest to 0. Use this definition of *closest* here and throughout.)

[a] 0

[b] 0.001

[c] 0.01

[d] 0.1

[e] 0.5

6. Now, generate 1000 fresh points and use them to estimate the out-of-sample error E_{out} of g that you got in Problem 5 (number of misclassified out-of-sample points / total number of out-of-sample points). Again, run the experiment 1000 times and take the average. Which value is closest to the average E_{out}?

[a] 0

[b] 0.001

[c] 0.01

[d] 0.1

[e] 0.5

7. Now, take $N = 10$. After finding the weights using Linear Regression, use them as a vector of initial weights for the Perceptron Learning Algorithm. Run PLA until it converges to a final vector of weights that completely separates all the in-sample points. Among the choices below, what is the closest value to the average number of iterations (over 1000 runs) that PLA takes to converge? (When implementing PLA, have the algorithm choose a point randomly from the set of misclassified points at each iteration)

[a] 1

[b] 15

[c] 300

[d] 5000

[e] 10000

• Nonlinear Transformation

In these problems, we again apply Linear Regression for classification. Consider the target function:

$$f(x_1, x_2) = \text{sign}(x_1^2 + x_2^2 - 0.6)$$

Generate a training set of $N = 1000$ points on $\mathcal{X} = [-1, 1] \times [-1, 1]$ with a uniform probability of picking each $\mathbf{x} \in \mathcal{X}$. Generate simulated noise by flipping the sign of the output in a randomly selected 10% subset of the generated training set.

8. Carry out Linear Regression without transformation, i.e., with feature vector:

$$(1, x_1, x_2),$$

to find the weight $\mathbf{w}$. What is the closest value to the classification in-sample error E_{in}? (Run the experiment 1000 times and take the average E_{in} to reduce variation in your results.)

[a] 0

[b] 0.1

[c] 0.3

[d] 0.5

[e] 0.8

9. Now, transform the $N = 1000$ training data into the following nonlinear feature vector:

$$(1, x_1, x_2, x_1x_2, x_1^2, x_2^2)$$

Find the vector $\tilde{\mathbf{w}}$ that corresponds to the solution of Linear Regression. Which of the following hypotheses is closest to the one you find? Closest here means agrees the most with your hypothesis (has the highest probability of agreeing on a randomly selected point). Average over a few runs to make sure your answer is stable.

[a] $g(x_1, x_2) = \text{sign}(-1 - 0.05x_1 + 0.08x_2 + 0.13x_1x_2 + 1.5x_1^2 + 1.5x_2^2)$

[b] $g(x_1, x_2) = \text{sign}(-1 - 0.05x_1 + 0.08x_2 + 0.13x_1x_2 + 1.5x_1^2 + 15x_2^2)$

[c] $g(x_1, x_2) = \text{sign}(-1 - 0.05x_1 + 0.08x_2 + 0.13x_1x_2 + 15x_1^2 + 1.5x_2^2)$

[d] $g(x_1, x_2) = \text{sign}(-1 - 1.5x_1 + 0.08x_2 + 0.13x_1x_2 + 0.05x_1^2 + 0.05x_2^2)$

[e] $g(x_1, x_2) = \text{sign}(-1 - 0.05x_1 + 0.08x_2 + 1.5x_1x_2 + 0.15x_1^2 + 0.15x_2^2)$

10. What is the closest value to the classification out-of-sample error E_{out} of your hypothesis from Problem 9? (Estimate it by generating a new set of 1000 points and adding noise, as before. Average over 1000 runs to reduce the variation in your results.)

[a] 0

[b] 0.1

[c] 0.3

[d] 0.5

[e] 0.8

Learning From Data
Yaser Abu-Mostafa, *Caltech*
http://work.caltech.edu/telecourse
Self-paced version

Answer Key To Homework # 2

1. **[b]**
2. **[d]**
3. **[e]**
4. **[b]**
5. **[c]**
6. **[c]**
7. **[a]**
8. **[d]**
9. **[a]**
10. **[b]**

Visit the forum (http://book.caltech.edu/bookforum) for discussion. Please follow the forum guidelines for posting answers (see the “BEFORE posting answers” announcement at the top thread there).

Learning From Data
Yaser Abu-Mostafa, *Caltech*
http://work.caltech.edu/telecourse
Self-paced version

Homework # 3

*All questions have multiple-choice answers (***[a], [b], [c], ...***). You can collaborate with others, but do not discuss the selected or excluded choices in the answers. You can consult books and notes, but not other people's solutions. Your solutions should be based on your own work. Definitions and notation follow the lectures.*

Note about the homework

- The goal of the homework is to facilitate a deeper understanding of the course material. The questions are not designed to be puzzles with catchy answers. They are meant to make you roll up your sleeves, face uncertainties, and approach the problem from different angles.

- The problems range from easy to difficult, and from practical to theoretical. Some problems require running a full experiment to arrive at the answer.

- The answer may not be obvious or numerically close to one of the choices, but one (and only one) choice will be correct if you follow the instructions precisely in each problem. You are encouraged to explore the problem further by experimenting with variations on these instructions, for the learning benefit.

- You are also encouraged to take part in the forum

  http://book.caltech.edu/bookforum

  where there are many threads about each homework set. We hope that you will contribute to the discussion as well. Please follow the forum guidelines for posting answers (see the "BEFORE posting answers" announcement at the top there).

- **Generalization Error**

1. The modified Hoeffding Inequality provides a way to characterize the generalization error with a probabilistic bound

$$\mathbb{P}\left[|E_{in}(g) - E_{out}(g)| > \epsilon\right] \leq 2Me^{-2\epsilon^2 N}$$

for any $\epsilon > 0$. If we set $\epsilon = 0.05$ and want the probability bound $2Me^{-2\epsilon^2 N}$ to be at most 0.03, what is the least number of examples N (among the given choices) needed for the case $M = 1$?

[**a**] 500

[**b**] 1000

[**c**] 1500

[**d**] 2000

[**e**] More examples are needed.

2. Repeat for the case $M = 10$.

[**a**] 500

[**b**] 1000

[**c**] 1500

[**d**] 2000

[**e**] More examples are needed.

3. Repeat for the case $M = 100$.

[**a**] 500

[**b**] 1000

[**c**] 1500

[**d**] 2000

[**e**] More examples are needed.

- **Break Point**

4. As shown in class, the (smallest) break point for the Perceptron Model in the two-dimensional case ($\mathbb{R}^2$) is 4 points. What is the smallest break point for the Perceptron Model in $\mathbb{R}^3$? (i.e., instead of the hypothesis set consisting of separating lines, it consists of separating planes.)

[**a**] 4

[**b**] 5

[**c**] 6

[**d**] 7

[**e**] 8

- **Growth Function**

5. Which of the following are possible formulas for a growth function $m_{\mathcal{H}}(N)$:

$$\begin{array}{rlrl} \text{i)} & 1+N & \text{iv)} & 2^{\lfloor N/2 \rfloor} \\ \text{ii)} & 1+N+\binom{N}{2} & \text{v)} & 2^{N} \\ \text{iii)} & \sum_{i=1}^{\lfloor \sqrt{N} \rfloor} \binom{N}{i} & & \end{array}$$

where $\lfloor u \rfloor$ is the biggest integer $\leq u$, and $\binom{M}{m} = 0$ when $m > M$.

[**a**] i, v

[**b**] i, ii, v

[**c**] i, iv, v

[**d**] i, ii, iii, v

[**e**] i, ii, iii, iv, v

- **Fun with Intervals**

6. Consider the "2-intervals" learning model, where h: $\mathbb{R} \rightarrow \{-1, +1\}$ and $h(x) = +1$ if the point is within either of two arbitrarily chosen intervals and -1 otherwise. What is the (smallest) break point for this hypothesis set?

[**a**] 3

[**b**] 4

[**c**] 5

[**d**] 6

[**e**] 7

7. Which of the following is the growth function $m_H(N)$ for the "2-intervals" hypothesis set?

[a] $\binom{N+1}{4}$

[b] $\binom{N+1}{2} + 1$

[c] $\binom{N+1}{4} + \binom{N+1}{2} + 1$

[d] $\binom{N+1}{4} + \binom{N+1}{3} + \binom{N+1}{2} + \binom{N+1}{1} + 1$

[e] None of the above

8. Now, consider the general case: the "M-intervals" learning model. Again $h : \mathbb{R} \to \{-1, +1\}$, where $h(x) = +1$ if the point falls inside any of M arbitrarily chosen intervals, otherwise $h(x) = -1$. What is the (smallest) break point of this hypothesis set?

[a] M

[b] $M + 1$

[c] M^2

[d] $2M + 1$

[e] $2M - 1$

- **Convex Sets: The Triangle**

9. Consider the "triangle" learning model, where $h : \mathbb{R}^2 \to \{-1, +1\}$ and $h(\mathbf{x}) = +1$ if $\mathbf{x}$ lies within an arbitrarily chosen triangle in the plane and -1 otherwise. Which is the largest number of points in $\mathbb{R}^2$ (among the given choices) that can be shattered by this hypothesis set?

[a] 1

[b] 3

[c] 5

[d] 7

[e] 9

- **Non-Convex Sets: Concentric Circles**

10. Compute the growth function $m_{\mathcal{H}}(N)$ for the learning model made up of two concentric circles in $\mathbb{R}^2$. Specifically, $\mathcal{H}$ contains the functions which are $+1$ for

$$a^2 \leq x_1^2 + x_2^2 \leq b^2$$

and -1 otherwise. The growth function is [**a**] $N+1$

[**b**] $\binom{N+1}{2}+1$

[**c**] $\binom{N+1}{3}+1$

[**d**] $2N^2+1$

[**e**] None of the above

Learning From Data

Yaser Abu-Mostafa, *Caltech*
http://work.caltech.edu/telecourse

Self-paced version

Answer Key To Homework # 3

1. **[b]**
2. **[c]**
3. **[d]**
4. **[b]**
5. **[b]**
6. **[c]**
7. **[c]**
8. **[d]**
9. **[d]**
10. **[b]**

Visit the forum (http://book.caltech.edu/bookforum) for discussion. Please follow the forum guidelines for posting answers (see the “BEFORE posting answers” announcement at the top thread there).

Learning From Data
Yaser Abu-Mostafa, *Caltech*
http://work.caltech.edu/telecourse
Self-paced version

Homework # 4

*All questions have multiple-choice answers (***[a]***,* **[b]***,* **[c]***,* **...***). You can collaborate with others, but do not discuss the selected or excluded choices in the answers. You can consult books and notes, but not other people's solutions. Your solutions should be based on your own work. Definitions and notation follow the lectures.*

Note about the homework

- The goal of the homework is to facilitate a deeper understanding of the course material. The questions are not designed to be puzzles with catchy answers. They are meant to make you roll up your sleeves, face uncertainties, and approach the problem from different angles.

- The problems range from easy to difficult, and from practical to theoretical. Some problems require running a full experiment to arrive at the answer.

- The answer may not be obvious or numerically close to one of the choices, but one (and only one) choice will be correct if you follow the instructions precisely in each problem. You are encouraged to explore the problem further by experimenting with variations on these instructions, for the learning benefit.

- You are also encouraged to take part in the forum

  http://book.caltech.edu/bookforum

  where there are many threads about each homework set. We hope that you will contribute to the discussion as well. Please follow the forum guidelines for posting answers (see the "BEFORE posting answers" announcement at the top there).

- **Generalization Error**

In Problems 1-3, we look at generalization bounds numerically. For $N > d_{\text{VC}}$, use the simple approximate bound $N^{d_{\text{VC}}}$ for the growth function $m_{\mathcal{H}}(N)$.

1. For an $\mathcal{H}$ with $d_{\text{VC}} = 10$, if you want 95% confidence that your generalization error is at most 0.05, what is the closest numerical approximation of the sample size that the VC generalization bound predicts?

[**a**] 400,000

[**b**] 420,000

[**c**] 440,000

[**d**] 460,000

[**e**] 480,000

2. There are a number of bounds on the generalization error ϵ, all holding with probability at least $1-\delta$. Fix $d_{\text{VC}} = 50$ and $\delta = 0.05$ and plot these bounds as a function of N. Which bound is the smallest for very large N, say $N = 10,000$? Note that [**c**] and [**d**] are implicit bounds in ϵ.

[**a**] Original VC bound: $\epsilon \leq \sqrt{\frac{8}{N}\ln\frac{4m_{\mathcal{H}}(2N)}{\delta}}$

[**b**] Rademacher Penalty Bound: $\epsilon \leq \sqrt{\frac{2\ln(2Nm_{\mathcal{H}}(N))}{N}} + \sqrt{\frac{2}{N}\ln\frac{1}{\delta}} + \frac{1}{N}$

[**c**] Parrondo and Van den Broek: $\epsilon \leq \sqrt{\frac{1}{N}(2\epsilon + \ln\frac{6m_{\mathcal{H}}(2N)}{\delta})}$

[**d**] Devroye: $\epsilon \leq \sqrt{\frac{1}{2N}(4\epsilon(1+\epsilon) + \ln\frac{4m_{\mathcal{H}}(N^2)}{\delta})}$

[**e**] They are all equal.

3. For the same values of d_{VC} and δ of Problem 2, but for small N, say $N = 5$, which bound is the smallest?

[**a**] Original VC bound: $\epsilon \leq \sqrt{\frac{8}{N}\ln\frac{4m_{\mathcal{H}}(2N)}{\delta}}$

[**b**] Rademacher Penalty Bound: $\epsilon \leq \sqrt{\frac{2\ln(2Nm_{\mathcal{H}}(N))}{N}} + \sqrt{\frac{2}{N}\ln\frac{1}{\delta}} + \frac{1}{N}$

[**c**] Parrondo and Van den Broek: $\epsilon \leq \sqrt{\frac{1}{N}(2\epsilon + \ln\frac{6m_{\mathcal{H}}(2N)}{\delta})}$

[**d**] Devroye: $\epsilon \leq \sqrt{\frac{1}{2N}(4\epsilon(1+\epsilon) + \ln\frac{4m_{\mathcal{H}}(N^2)}{\delta})}$

[**e**] They are all equal.

• Bias and Variance

Consider the case where the target function $f : [-1, 1] \rightarrow \mathbb{R}$ is given by $f(x) = \sin(\pi x)$ and the input probability distribution is uniform on $[-1, 1]$. Assume that the training set has only two examples (picked independently), and that the learning algorithm produces the hypothesis that minimizes the mean squared error on the examples.

4. Assume the learning model consists of all hypotheses of the form $h(x) = ax$. What is the expected value, $\bar{g}(x)$, of the hypothesis produced by the learning algorithm (expected value with respect to the data set)? Express your $\bar{g}(x)$ as $\hat{a}x$, and round $\hat{a}$ to two decimal digits only, then match *exactly* to one of the following answers.

[**a**] $\bar{g}(x) = 0$

[**b**] $\bar{g}(x) = 0.79x$

[**c**] $\bar{g}(x) = 1.07x$

[**d**] $\bar{g}(x) = 1.58x$

[**e**] None of the above

5. What is the closest value to the bias in this case?

[**a**] 0.1

[**b**] 0.3

[**c**] 0.5

[**d**] 0.7

[**e**] 1.0

6. What is the closest value to the variance in this case?

[**a**] 0.2

[**b**] 0.4

[**c**] 0.6

[**d**] 0.8

[**e**] 1.0

7. Now, let's change $\mathcal{H}$. Which of the following learning models has the least expected value of out-of-sample error?

[**a**] Hypotheses of the form $h(x) = b$

[**b**] Hypotheses of the form $h(x) = ax$

[c] Hypotheses of the form $h(x) = ax + b$

[d] Hypotheses of the form $h(x) = ax^2$

[e] Hypotheses of the form $h(x) = ax^2 + b$

• VC Dimension

8. Assume $q \geq 1$ is an integer and let $m_{\mathcal{H}}(1) = 2$. What is the VC dimension of a hypothesis set whose growth function satisfies: $m_{\mathcal{H}}(N+1) = 2m_{\mathcal{H}}(N) - \binom{N}{q}$? Recall that $\binom{M}{m} = 0$ when $m > M$.

[a] $q - 2$

[b] $q - 1$

[c] q

[d] $q + 1$

[e] None of the above

9. For hypothesis sets $\mathcal{H}_1, \mathcal{H}_2, ..., \mathcal{H}_K$ with finite, positive VC dimensions $d_{\text{VC}}(\mathcal{H}_k)$, some of the following bounds are correct and some are not. Which among the correct ones is the tightest bound (the smallest range of values) on the VC dimension of the **intersection** of the sets: $d_{\text{VC}}(\bigcap_{k=1}^{K}\mathcal{H}_k)$? (The VC dimension of an empty set or a singleton set is taken as zero)

[a] $0 \leq d_{\text{VC}}(\bigcap_{k=1}^{K}\mathcal{H}_k) \leq \sum_{k=1}^{K} d_{\text{VC}}(\mathcal{H}_k)$

[b] $0 \leq d_{\text{VC}}(\bigcap_{k=1}^{K}\mathcal{H}_k) \leq \min\{d_{\text{VC}}(\mathcal{H}_k)\}_{k=1}^{K}$

[c] $0 \leq d_{\text{VC}}(\bigcap_{k=1}^{K}\mathcal{H}_k) \leq \max\{d_{\text{VC}}(\mathcal{H}_k)\}_{k=1}^{K}$

[d] $\min\{d_{\text{VC}}(\mathcal{H}_k)\}_{k=1}^{K} \leq d_{\text{VC}}(\bigcap_{k=1}^{K}\mathcal{H}_k) \leq \max\{d_{\text{VC}}(\mathcal{H}_k)\}_{k=1}^{K}$

[e] $\min\{d_{\text{VC}}(\mathcal{H}_k)\}_{k=1}^{K} \leq d_{\text{VC}}(\bigcap_{k=1}^{K}\mathcal{H}_k) \leq \sum_{k=1}^{K} d_{\text{VC}}(\mathcal{H}_k)$

10. For hypothesis sets $\mathcal{H}_1, \mathcal{H}_2, ..., \mathcal{H}_K$ with finite, positive VC dimensions $d_{\text{VC}}(\mathcal{H}_k)$, some of the following bounds are correct and some are not. Which among the correct ones is the tightest bound (the smallest range of values) on the VC dimension of the **union** of the sets: $d_{\text{VC}}(\bigcup_{k=1}^{K}\mathcal{H}_k)$?

[a] $0 \leq d_{\text{VC}}(\bigcup_{k=1}^{K}\mathcal{H}_k) \leq \sum_{k=1}^{K} d_{\text{VC}}(\mathcal{H}_k)$

[b] $0 \leq d_{\text{VC}}(\bigcup_{k=1}^{K}\mathcal{H}_k) \leq K - 1 + \sum_{k=1}^{K} d_{\text{VC}}(\mathcal{H}_k)$

[c] $\min\{d_{\text{VC}}(\mathcal{H}_k)\}_{k=1}^K \;\le\; d_{\text{VC}}(\bigcup_{k=1}^K \mathcal{H}_k) \;\le\; \sum_{k=1}^K d_{\text{VC}}(\mathcal{H}_k)$

[d] $\max\{d_{\text{VC}}(\mathcal{H}_k)\}_{k=1}^K \;\le\; d_{\text{VC}}(\bigcup_{k=1}^K \mathcal{H}_k) \;\le\; \sum_{k=1}^K d_{\text{VC}}(\mathcal{H}_k)$

[e] $\max\{d_{\text{VC}}(\mathcal{H}_k)\}_{k=1}^K \;\le\; d_{\text{VC}}(\bigcup_{k=1}^K \mathcal{H}_k) \;\le\; K - 1 + \sum_{k=1}^K d_{\text{VC}}(\mathcal{H}_k)$

Learning From Data
Yaser Abu-Mostafa, *Caltech*
http://work.caltech.edu/telecourse
Self-paced version

Answer Key To Homework # 4

1. **[d]**
2. **[d]**
3. **[c]**
4. **[e]**
5. **[b]**
6. **[a]**
7. **[b]**
8. **[c]**
9. **[b]**
10. **[e]**

Visit the forum (http://book.caltech.edu/bookforum) for discussion. Please follow the forum guidelines for posting answers (see the "BEFORE posting answers" announcement at the top thread there).

Learning From Data
Yaser Abu-Mostafa, *Caltech*
http://work.caltech.edu/telecourse
Self-paced version

Homework # 5

*All questions have multiple-choice answers (***[a]***,* **[b]***,* **[c]***,* **...***). You can collaborate with others, but do not discuss the selected or excluded choices in the answers. You can consult books and notes, but not other people's solutions. Your solutions should be based on your own work. Definitions and notation follow the lectures.*

Note about the homework

- The goal of the homework is to facilitate a deeper understanding of the course material. The questions are not designed to be puzzles with catchy answers. They are meant to make you roll up your sleeves, face uncertainties, and approach the problem from different angles.
- The problems range from easy to difficult, and from practical to theoretical. Some problems require running a full experiment to arrive at the answer.
- The answer may not be obvious or numerically close to one of the choices, but one (and only one) choice will be correct if you follow the instructions precisely in each problem. You are encouraged to explore the problem further by experimenting with variations on these instructions, for the learning benefit.
- You are also encouraged to take part in the forum

  http://book.caltech.edu/bookforum

  where there are many threads about each homework set. We hope that you will contribute to the discussion as well. Please follow the forum guidelines for posting answers (see the "BEFORE posting answers" announcement at the top there).

• Linear Regression Error

Consider a noisy target $y = \mathbf{w}^{*T}\mathbf{x} + \epsilon$, where $\mathbf{x} \in \mathbb{R}^d$ (with the added coordinate $x_0 = 1$), $y \in \mathbb{R}$, $\mathbf{w}^*$ is an unknown vector, and ϵ is a noise term with zero mean and σ^2 variance. Assume ϵ is independent of $\mathbf{x}$ and of all other ϵ's. If linear regression is carried out using a training data set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, and outputs the parameter vector $\mathbf{w}_{\text{lin}}$, it can be shown that the expected in-sample error E_{in} with respect to $\mathcal{D}$ is given by:

$$\mathbb{E}_{\mathcal{D}}[E_{\text{in}}(\mathbf{w}_{\text{lin}})] = \sigma^2 \left(1 - \frac{d+1}{N}\right)$$

1. For $\sigma = 0.1$ and $d = 8$, which among the following choices is the smallest number of examples N that will result in an expected E_{in} greater than 0.008?

[**a**] 10

[**b**] 25

[**c**] 100

[**d**] 500

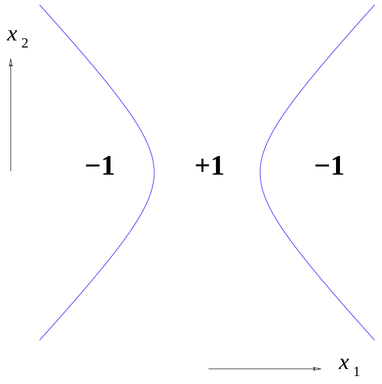
[**e**] 1000

• Nonlinear Transforms

In linear classification, consider the feature transform $\Phi : \mathbb{R}^2 \to \mathbb{R}^2$ (plus the added zeroth coordinate) given by:

$$\Phi(1, x_1, x_2) = (1, x_1^2, x_2^2)$$

2. Which of the following sets of constraints on the weights in the $\mathcal{Z}$ space could correspond to the hyperbolic decision boundary in $\mathcal{X}$ depicted in the figure?



You may assume that $\tilde{w}_0$ can be selected to achieve the desired boundary.

[a] $\tilde{w}_1 = 0, \tilde{w}_2 > 0$

[b] $\tilde{w}_1 > 0, \tilde{w}_2 = 0$

[c] $\tilde{w}_1 > 0, \tilde{w}_2 > 0$

[d] $\tilde{w}_1 < 0, \tilde{w}_2 > 0$

[e] $\tilde{w}_1 > 0, \tilde{w}_2 < 0$

Now, consider the 4th order polynomial transform from the input space $\mathbb{R}^2$:

$$\Phi_4 : \mathbf{x} \rightarrow (1, x_1, x_2, x_1^2, x_1x_2, x_2^2, x_1^3, x_1^2x_2, x_1x_2^2, x_2^3, x_1^4, x_1^3x_2, x_1^2x_2^2, x_1x_2^3, x_2^4)$$

3. What is the smallest value among the following choices that is *not* smaller than the VC dimension of a linear model in this transformed space?

[a] 3

[b] 5

[c] 15

[d] 20

[e] 21

- **Gradient Descent**

Consider the nonlinear error surface $E(u, v) = (ue^v - 2ve^{-u})^2$. We start at the point $(u, v) = (1, 1)$ and minimize this error using gradient descent in the uv space. Use $\eta = 0.1$ (learning rate, not step size).

4. What is the partial derivative of $E(u, v)$ with respect to u, i.e., $\frac{\partial E}{\partial u}$?

[a] $(ue^v - 2ve^{-u})^2$

[b] $2(ue^v - 2ve^{-u})$

[c] $2(e^v + 2ve^{-u})$

[d] $2(e^v - 2ve^{-u})(ue^v - 2ve^{-u})$

[e] $2(e^v + 2ve^{-u})(ue^v - 2ve^{-u})$

5. How many iterations (among the given choices) does it take for the error $E(u, v)$ to fall below 10^{-14} for the first time? In your programs, make sure to use double precision to get the needed accuracy.

[a] 1

[**b**] 3

[**c**] 5

[**d**] 10

[**e**] 17

6. After running enough iterations such that the error has just dropped below 10^{-14}, what are the closest values (in Euclidean distance) among the following choices to the final (u, v) you got in Problem 5?

[**a**] $(1.000, 1.000)$

[**b**] $(0.713, 0.045)$

[**c**] $(0.016, 0.112)$

[**d**] $(-0.083, 0.029)$

[**e**] $(0.045, 0.024)$

7. Now, we will compare the performance of "coordinate descent." In each iteration, we have two steps along the 2 coordinates. Step 1 is to move only along the u coordinate to reduce the error (assume first-order approximation holds like in gradient descent), and step 2 is to reevaluate and move only along the v coordinate to reduce the error (again, assume first-order approximation holds). Use the same learning rate of $\eta = 0.1$ as we did in gradient descent. What will the error $E(u, v)$ be closest to after 15 full iterations (30 steps)?

[**a**] 10^{-1}

[**b**] 10^{-7}

[**c**] 10^{-14}

[**d**] 10^{-17}

[**e**] 10^{-20}

• Logistic Regression

In this problem you will create your own target function f (probability in this case) and data set $\mathcal{D}$ to see how Logistic Regression works. For simplicity, we will take f to be a 0/1 probability so y is a deterministic function of $\mathbf{x}$.

Take $d = 2$ so you can visualize the problem, and let $\mathcal{X} = [-1, 1] \times [-1, 1]$ with uniform probability of picking each $\mathbf{x} \in \mathcal{X}$. Choose a line in the plane as the boundary between $f(\mathbf{x}) = 1$ (where y has to be $+1$) and $f(\mathbf{x}) = 0$ (where y has to be -1) by taking two random, uniformly distributed points from $\mathcal{X}$ and taking the line passing through them as the boundary between $y = \pm 1$. Pick $N = 100$ training points at random from $\mathcal{X}$, and evaluate the outputs y_n for each of these points $\mathbf{x}_n$.

Run Logistic Regression with Stochastic Gradient Descent to find g, and estimate E_{out} (the **cross entropy** error) by generating a sufficiently large, separate set of points to evaluate the error. Repeat the experiment for 100 runs with different targets and take the average. Initialize the weight vector of Logistic Regression to all zeros in each run. Stop the algorithm when $\|\mathbf{w}^{(t-1)} - \mathbf{w}^{(t)}\| < 0.01$, where $\mathbf{w}^{(t)}$ denotes the weight vector at the end of epoch t. An epoch is a full pass through the N data points (use a random permutation of $1, 2, \cdots, N$ to present the data points to the algorithm within each epoch, and use different permutations for different epochs). Use a learning rate of 0.01.

8. Which of the following is closest to E_{out} for $N = 100$?

[**a**] 0.025
[**b**] 0.050
[**c**] 0.075
[**d**] 0.100
[**e**] 0.125

9. How many epochs does it take on average for Logistic Regression to converge for $N = 100$ using the above initialization and termination rules and the specified learning rate? Pick the value that is closest to your results.

[**a**] 350
[**b**] 550
[**c**] 750
[**d**] 950
[**e**] 1750

• PLA as SGD

10. The Perceptron Learning Algorithm can be implemented as SGD using which of the following error functions $\mathrm{e}_n(\mathbf{w})$? Ignore the points $\mathbf{w}$ at which $\mathrm{e}_n(\mathbf{w})$ is not twice differentiable.

[**a**] $\mathrm{e}_n(\mathbf{w}) = e^{-y_n \mathbf{w}^\intercal \mathbf{x}_n}$
[**b**] $\mathrm{e}_n(\mathbf{w}) = -y_n \mathbf{w}^\intercal \mathbf{x}_n$
[**c**] $\mathrm{e}_n(\mathbf{w}) = (y_n - \mathbf{w}^\intercal \mathbf{x}_n)^2$
[**d**] $\mathrm{e}_n(\mathbf{w}) = \ln(1 + e^{-y_n \mathbf{w}^\intercal \mathbf{x}_n})$
[**e**] $\mathrm{e}_n(\mathbf{w}) = -\min(0, y_n \mathbf{w}^\intercal \mathbf{x}_n)$

Learning From Data
Yaser Abu-Mostafa, *Caltech*
http://work.caltech.edu/telecourse
Self-paced version

Answer Key To Homework # 5

1. **[c]**
2. **[d]**
3. **[c]**
4. **[e]**
5. **[d]**
6. **[e]**
7. **[a]**
8. **[d]**
9. **[a]**
10. **[e]**

Visit the forum (http://book.caltech.edu/bookforum) for discussion. Please follow the forum guidelines for posting answers (see the “BEFORE posting answers” announcement at the top thread there).

Learning From Data
Yaser Abu-Mostafa, *Caltech*
http://work.caltech.edu/telecourse
Self-paced version

Homework # 6

All questions have multiple-choice answers ([**a**], [**b**], [**c**], **...**)*. You can collaborate with others, but do not discuss the selected or excluded choices in the answers. You can consult books and notes, but not other people's solutions. Your solutions should be based on your own work. Definitions and notation follow the lectures.*

Note about the homework

- The goal of the homework is to facilitate a deeper understanding of the course material. The questions are not designed to be puzzles with catchy answers. They are meant to make you roll up your sleeves, face uncertainties, and approach the problem from different angles.

- The problems range from easy to difficult, and from practical to theoretical. Some problems require running a full experiment to arrive at the answer.

- The answer may not be obvious or numerically close to one of the choices, but one (and only one) choice will be correct if you follow the instructions precisely in each problem. You are encouraged to explore the problem further by experimenting with variations on these instructions, for the learning benefit.

- You are also encouraged to take part in the forum

  http://book.caltech.edu/bookforum

  where there are many threads about each homework set. We hope that you will contribute to the discussion as well. Please follow the forum guidelines for posting answers (see the "BEFORE posting answers" announcement at the top there).

• Overfitting and Deterministic Noise

1. Deterministic noise depends on $\mathcal{H}$, as some models approximate f better than others. Assume $\mathcal{H}' \subset \mathcal{H}$ and that f is fixed. **In general** (but not necessarily in all cases), if we use $\mathcal{H}'$ instead of $\mathcal{H}$, how does deterministic noise behave?

[**a**] In general, deterministic noise will decrease.

[**b**] In general, deterministic noise will increase.

[**c**] In general, deterministic noise will be the same.

[**d**] There is deterministic noise for only one of $\mathcal{H}$ and $\mathcal{H}'$.

• Regularization with Weight Decay

In the following problems use the data provided in the files

http://work.caltech.edu/data/in.dta

http://work.caltech.edu/data/out.dta

as a training and test set respectively. Each line of the files corresponds to a two-dimensional input $\mathbf{x} = (x_1, x_2)$, so that $\mathcal{X} = \mathbb{R}^2$, followed by the corresponding label from $\mathcal{Y} = \{-1, 1\}$. We are going to apply Linear Regression with a non-linear transformation for classification. The nonlinear transformation is given by

$$\phi(x_1, x_2) = (1, x_1, x_2, x_1^2, x_2^2, x_1x_2, |x_1 - x_2|, |x_1 + x_2|).$$

Recall that the classification error is defined as the fraction of misclassified points.

2. Run Linear Regression on the training set after performing the non-linear transformation. What values are closest (in Euclidean distance) to the in-sample and out-of-sample classification errors, respectively?

[**a**] 0.03, 0.08

[**b**] 0.03, 0.10

[**c**] 0.04, 0.09

[**d**] 0.04, 0.11

[**e**] 0.05, 0.10

3. Now add weight decay to Linear Regression, that is, add the term $\frac{\lambda}{N}\sum_{i=0}^{7} w_i^2$ to the squared in-sample error, using $\lambda = 10^k$. What are the closest values to the in-sample and out-of-sample classification errors, respectively, for $k = -3$? Recall that the solution for Linear Regression with Weight Decay was derived in class.

[**a**] 0.01, 0.02

[**b**] 0.02, 0.04

[**c**] 0.02, 0.06

[**d**] 0.03, 0.08

[**e**] 0.03, 0.10

4. Now, use $k = 3$. What are the closest values to the new in-sample and out-of-sample classification errors, respectively?

[**a**] 0.2, 0.2

[**b**] 0.2, 0.3

[**c**] 0.3, 0.3

[**d**] 0.3, 0.4

[**e**] 0.4, 0.4

5. What value of k, among the following choices, achieves the smallest out-of-sample classification error?

[**a**] 2

[**b**] 1

[**c**] 0

[**d**] -1

[**e**] -2

6. What value is closest to the minimum out-of-sample classification error achieved by varying k (limiting k to integer values)?

[**a**] 0.04

[**b**] 0.06

[**c**] 0.08

[**d**] 0.10

[**e**] 0.12

• Regularization for Polynomials

Polynomial models can be viewed as linear models in a space $\mathcal{Z}$, under a nonlinear transform $\Phi : \mathcal{X} \to \mathcal{Z}$. Here, Φ transforms the scalar x into a vector $\mathbf{z}$ of Legendre polynomials, $\mathbf{z} = (1, L_1(x), L_2(x), ..., L_Q(x))$. Our hypothesis set will be expressed as a linear combination of these polynomials,

$$\mathcal{H}_Q = \left\{ h \mid h(x) = \mathbf{w}^{\mathrm{T}}\mathbf{z} = \sum_{q=0}^{Q} w_q L_q(x) \right\},$$

where $L_0(x) = 1$.

7. Consider the following hypothesis set defined by the constraint:

$$\mathcal{H}(Q, C, Q_o) = \{h \mid h(x) = \mathbf{w}^{\mathrm{T}}\mathbf{z} \in \mathcal{H}_Q; w_q = C \text{ for } q \geq Q_o\},$$

which of the following statements is correct:

[a] $\mathcal{H}(10, 0, 3) \cup \mathcal{H}(10, 0, 4) = \mathcal{H}_4$

[b] $\mathcal{H}(10, 1, 3) \cup \mathcal{H}(10, 1, 4) = \mathcal{H}_3$

[c] $\mathcal{H}(10, 0, 3) \cap \mathcal{H}(10, 0, 4) = \mathcal{H}_2$

[d] $\mathcal{H}(10, 1, 3) \cap \mathcal{H}(10, 1, 4) = \mathcal{H}_1$

[e] None of the above

• Neural Networks

8. A fully connected Neural Network has $L = 2$; $d^{(0)} = 5$, $d^{(1)} = 3$, $d^{(2)} = 1$. If only products of the form $w_{ij}^{(l)} x_i^{(l-1)}$, $w_{ij}^{(l)} \delta_j^{(l)}$, and $x_i^{(l-1)} \delta_j^{(l)}$ count as operations (even for $x_0^{(l-1)} = 1$), without counting anything else, which of the following is the closest to the total number of operations in a single iteration of backpropagation (using SGD on one data point)?

[a] 30

[b] 35

[c] 40

[d] 45

[e] 50

Let us call every 'node' in a Neural Network a unit, whether that unit is an input variable or a neuron in one of the layers. Consider a Neural Network that has 10 input units (the constant $x_0^{(0)}$ is counted here as a unit), one output unit, and 36 hidden units (each $x_0^{(l)}$ is also counted as a unit). The hidden units can be arranged in any number of layers $l = 1, \cdots, L-1$, and each layer is fully connected to the layer above it.

9. What is the minimum possible number of weights that such a network can have?

[**a**] 46

[**b**] 47

[**c**] 56

[**d**] 57

[**e**] 58

10. What is the maximum possible number of weights that such a network can have?

[**a**] 386

[**b**] 493

[**c**] 494

[**d**] 509

[**e**] 510

Learning From Data
Yaser Abu-Mostafa, *Caltech*
http://work.caltech.edu/telecourse
Self-paced version

Answer Key To Homework # 6

1. **[b]**
2. **[a]**
3. **[d]**
4. **[e]**
5. **[d]**
6. **[b]**
7. **[c]**
8. **[d]**
9. **[a]**
10. **[e]**

Visit the forum (http://book.caltech.edu/bookforum) for discussion. Please follow the forum guidelines for posting answers (see the “BEFORE posting answers” announcement at the top thread there).

Learning From Data
Yaser Abu-Mostafa, *Caltech*
http://work.caltech.edu/telecourse
Self-paced version

Homework # 7

All questions have multiple-choice answers ([**a**], [**b**], [**c**], ...*). You can collaborate with others, but do not discuss the selected or excluded choices in the answers. You can consult books and notes, but not other people's solutions. Your solutions should be based on your own work. Definitions and notation follow the lectures.*

Note about the homework

- The goal of the homework is to facilitate a deeper understanding of the course material. The questions are not designed to be puzzles with catchy answers. They are meant to make you roll up your sleeves, face uncertainties, and approach the problem from different angles.

- The problems range from easy to difficult, and from practical to theoretical. Some problems require running a full experiment to arrive at the answer.

- The answer may not be obvious or numerically close to one of the choices, but one (and only one) choice will be correct if you follow the instructions precisely in each problem. You are encouraged to explore the problem further by experimenting with variations on these instructions, for the learning benefit.

- You are also encouraged to take part in the forum

  http://book.caltech.edu/bookforum

  where there are many threads about each homework set. We hope that you will contribute to the discussion as well. Please follow the forum guidelines for posting answers (see the "BEFORE posting answers" announcement at the top there).

- **Validation**

In the following problems, use the data provided in the files `in.dta` and `out.dta` for Homework # 6. We are going to apply linear regression with a nonlinear transformation for classification (without regularization). The nonlinear transformation is given by ϕ_0 through ϕ_7 which transform (x_1, x_2) into

$$1 \qquad x_1 \qquad x_2 \qquad x_1^2 \qquad x_2^2 \qquad x_1x_2 \qquad |x_1 - x_2| \qquad |x_1 + x_2|$$

To illustrate how taking out points for validation affects the performance, we will consider the hypotheses trained on $\mathcal{D}_{\text{train}}$ (without restoring the full $\mathcal{D}$ for training after validation is done).

1. Split `in.dta` into training (first 25 examples) and validation (last 10 examples). Train on the 25 examples only, using the validation set of 10 examples to select between five models that apply linear regression to ϕ_0 through ϕ_k, with $k = 3, 4, 5, 6, 7$. For which model is the classification error on the validation set smallest?

[**a**] $k = 3$

[**b**] $k = 4$

[**c**] $k = 5$

[**d**] $k = 6$

[**e**] $k = 7$

2. Evaluate the out-of-sample classification error using `out.dta` on the 5 models to see how well the validation set predicted the best of the 5 models. For which model is the out-of-sample classification error smallest?

[**a**] $k = 3$

[**b**] $k = 4$

[**c**] $k = 5$

[**d**] $k = 6$

[**e**] $k = 7$

3. Reverse the role of training and validation sets; now training with the last 10 examples and validating with the first 25 examples. For which model is the classification error on the validation set smallest?

[**a**] $k = 3$

[**b**] $k = 4$

[c] $k = 5$

[d] $k = 6$

[e] $k = 7$

4. Once again, evaluate the out-of-sample classification error using `out.dta` on the 5 models to see how well the validation set predicted the best of the 5 models. For which model is the out-of-sample classification error smallest?

[a] $k = 3$

[b] $k = 4$

[c] $k = 5$

[d] $k = 6$

[e] $k = 7$

5. What values are closest in Euclidean distance to the out-of-sample classification error obtained for the model chosen in Problems 1 and 3, respectively?

[a] 0.0, 0.1

[b] 0.1, 0.2

[c] 0.1, 0.3

[d] 0.2, 0.2

[e] 0.2, 0.3

- **Validation Bias**

6. Let e_1 and e_2 be independent random variables, distributed uniformly over the interval $[0, 1]$. Let $\mathsf{e} = \min(\mathsf{e}_1, \mathsf{e}_2)$. The expected values of $\mathsf{e}_1, \mathsf{e}_2, \mathsf{e}$ are closest to

[a] 0.5, 0.5, 0

[b] 0.5, 0.5, 0.1

[c] 0.5, 0.5, 0.25

[d] 0.5, 0.5, 0.4

[e] 0.5, 0.5, 0.5

- **Cross Validation**

7. You are given the data points (x, y): $(-1, 0), (\rho, 1), (1, 0)$, $\rho \geq 0$, and a choice between two models: constant $\{\, h_0(x) = b \,\}$ and linear $\{\, h_1(x) = ax + b \,\}$. For which value of ρ would the two models be tied using leave-one-out cross-validation with the squared error measure?

[a] $\sqrt{\sqrt{3}+4}$

[b] $\sqrt{\sqrt{3}-1}$

[c] $\sqrt{9+4\sqrt{6}}$

[d] $\sqrt{9-\sqrt{6}}$

[e] None of the above

• PLA vs. SVM

Notice: Quadratic Programming packages sometimes need tweaking and have numerical issues, and this is characteristic of packages you will use in practical ML situations. Your understanding of support vectors will help you get to the correct answers.

In the following problems, we compare PLA to SVM with hard margin[1] on linearly separable data sets. For each run, you will create your own target function f and data set $\mathcal{D}$. Take $d = 2$ and choose a random line in the plane as your target function f (do this by taking two random, uniformly distributed points on $[-1,1] \times [-1,1]$ and taking the line passing through them), where one side of the line maps to $+1$ and the other maps to -1. Choose the inputs $\mathbf{x}_n$ of the data set as random points in $\mathcal{X} = [-1,1] \times [-1,1]$, and evaluate the target function on each $\mathbf{x}_n$ to get the corresponding output y_n. If all data points are on one side of the line, discard the run and start a new run.

Start PLA with the all-zero vector and pick the misclassified point for each PLA iteration at random. Run PLA to find the final hypothesis g_{PLA} and measure the disagreement between f and g_{PLA} as $\mathbb{P}[f(\mathbf{x}) \neq g_{\text{PLA}}(\mathbf{x})]$ (you can either calculate this exactly, or approximate it by generating a sufficiently large, separate set of points to evaluate it). Now, run SVM on the same data to find the final hypothesis g_{SVM} by solving

$$\begin{array}{ll} \displaystyle\min_{\mathbf{w},b} & \displaystyle\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \\ \text{s.t.} & y_n\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b\right) \geq 1 \end{array}$$

using quadratic programming on the primal or the dual problem. Measure the disagreement between f and g_{SVM} as $\mathbb{P}[f(\mathbf{x}) \neq g_{\text{SVM}}(\mathbf{x})]$, and count the number of support vectors you get in each run.

8. For $N = 10$, repeat the above experiment for 1000 runs. How often is g_{SVM} better than g_{PLA} in approximating f? The percentage of time is closest to:

[a] 20%

[1] For hard margin in SVM packages, set $C \to \infty$.

[**b**] 40%

[**c**] 60%

[**d**] 80%

[**e**] 100%

9. For $N = 100$, repeat the above experiment for 1000 runs. How often is g_{SVM} better than g_{PLA} in approximating f? The percentage of time is closest to:

[**a**] 10%

[**b**] 30%

[**c**] 50%

[**d**] 70%

[**e**] 90%

10. For the case $N = 100$, which of the following is the closest to the average number of support vectors of g_{SVM} (averaged over the 1000 runs)?

[**a**] 2

[**b**] 3

[**c**] 5

[**d**] 10

[**e**] 20

Learning From Data
Yaser Abu-Mostafa, *Caltech*
http://work.caltech.edu/telecourse
Self-paced version

Answer Key To Homework # 7

1. **[d]**
2. **[e]**
3. **[d]**
4. **[d]**
5. **[b]**
6. **[d]**
7. **[c]**
8. **[c]**
9. **[d]**
10. **[b]**

Visit the forum (http://book.caltech.edu/bookforum) for discussion. Please follow the forum guidelines for posting answers (see the "BEFORE posting answers" announcement at the top thread there).

Learning From Data
Yaser Abu-Mostafa, *Caltech*
http://work.caltech.edu/telecourse
Self-paced version

Homework # 8

*All questions have multiple-choice answers (***[a]***,* **[b]***,* **[c]***, ...). You can collaborate with others, but do not discuss the selected or excluded choices in the answers. You can consult books and notes, but not other people's solutions. Your solutions should be based on your own work. Definitions and notation follow the lectures.*

Note about the homework

- The goal of the homework is to facilitate a deeper understanding of the course material. The questions are not designed to be puzzles with catchy answers. They are meant to make you roll up your sleeves, face uncertainties, and approach the problem from different angles.

- The problems range from easy to difficult, and from practical to theoretical. Some problems require running a full experiment to arrive at the answer.

- The answer may not be obvious or numerically close to one of the choices, but one (and only one) choice will be correct if you follow the instructions precisely in each problem. You are encouraged to explore the problem further by experimenting with variations on these instructions, for the learning benefit.

- You are also encouraged to take part in the forum

  http://book.caltech.edu/bookforum

  where there are many threads about each homework set. We hope that you will contribute to the discussion as well. Please follow the forum guidelines for posting answers (see the "BEFORE posting answers" announcement at the top there).

• Primal versus Dual Problem

1. Recall that N is the size of the data set and d is the dimensionality of the input space. The original formulation of the hard-margin SVM problem (minimize $\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$ subject to the inequality constraints), without going through the Lagrangian dual problem, is

[**a**] a quadratic programming problem with N variables

[**b**] a quadratic programming problem with $N+1$ variables

[**c**] a quadratic programming problem with d variables

[**d**] a quadratic programming problem with $d+1$ variables

[**e**] not a quadratic programming problem

Notice: The following problems deal with a real-life data set. In addition, the computational packages you use may employ different heuristics and require different tweaks. This is a typical situation that a Machine Learning practitioner faces. There are uncertainties, and the answers may or may not match our expectations. Although this situation is not as 'sanitized' as other homework problems, it is important to go through it as part of the learning experience.

SVM with Soft Margins

In the rest of the problems of this homework set, we apply soft-margin SVM to handwritten digits from the processed US Postal Service Zip Code data set. Download the data (extracted features of intensity and symmetry) for training and testing:

http://www.amlbook.com/data/zip/features.train

http://www.amlbook.com/data/zip/features.test

(the format of each row is: **digit intensity symmetry**). We will train two types of binary classifiers; one-versus-one (one digit is class $+1$ and another digit is class -1, with the rest of the digits disregarded), and one-versus-all (one digit is class $+1$ and the rest of the digits are class -1).

The data set has thousands of points, and some quadratic programming packages cannot handle this size. We recommend that you use the packages in libsvm:

http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Implement SVM with soft margin on the above zip-code data set by solving

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m K(\mathbf{x}_n, \mathbf{x}_m) - \sum_{n=1}^{N} \alpha_n \\ \text{s.t.} \quad & \sum_{n=1}^{N} y_n \alpha_n = 0 \\ & 0 \le \alpha_n \le C \quad n = 1, \cdots, N \end{aligned}$$

When evaluating E_{in} and E_{out} of the resulting classifier, use binary classification error.

Practical remarks:

(i) For the purpose of this homework, do not scale the data when you use libsvm or other packages, otherwise you may inadvertently change the (effective) kernel and get different results.

(ii) In some packages, you need to specify double precision.

(iii) In 10-fold cross validation, if the data size is not a multiple of 10, the sizes of the 10 subsets may be off by 1 data point.

(iv) Some packages have software parameters whose values affect the outcome. ML practitioners have to deal with this kind of added uncertainty.

- **Polynomial Kernels**

Consider the polynomial kernel $K(\mathbf{x}_n, \mathbf{x}_m) = (1 + \mathbf{x}_n^{\mathrm{T}}\mathbf{x}_m)^Q$, where Q is the degree of the polynomial.

2. With $C = 0.01$ and $Q = 2$, which of the following classifiers has the **highest** E_{in}?

[a] 0 versus all

[b] 2 versus all

[c] 4 versus all

[d] 6 versus all

[e] 8 versus all

3. With $C = 0.01$ and $Q = 2$, which of the following classifiers has the **lowest** E_{in}?

[a] 1 versus all

[b] 3 versus all

[c] 5 versus all

[**d**] 7 versus all

[**e**] 9 versus all

4. Comparing the two selected classifiers from Problems 2 and 3, which of the following values is the closest to the difference between the number of support vectors of these two classifiers?

[**a**] 600

[**b**] 1200

[**c**] 1800

[**d**] 2400

[**e**] 3000

5. Consider the 1 versus 5 classifier with $Q = 2$ and $C \in \{0.001, 0.01, 0.1, 1\}$. Which of the following statements is correct? Going up or down means strictly so.

[**a**] The number of support vectors goes down when C goes up.

[**b**] The number of support vectors goes up when C goes up.

[**c**] E_{out} goes down when C goes up.

[**d**] Maximum C achieves the lowest E_{in}.

[**e**] None of the above

6. In the 1 versus 5 classifier, comparing $Q = 2$ with $Q = 5$, which of the following statements is correct?

[**a**] When $C = 0.0001$, E_{in} is higher at $Q = 5$.

[**b**] When $C = 0.001$, the number of support vectors is lower at $Q = 5$.

[**c**] When $C = 0.01$, E_{in} is higher at $Q = 5$.

[**d**] When $C = 1$, E_{out} is lower at $Q = 5$.

[**e**] None of the above

• Cross Validation

In the next two problems, we will experiment with 10-fold cross validation for the polynomial kernel. Because E_{cv} is a random variable that depends on the random partition of the data, we will try 100 runs with different partitions and base our answer on how many runs lead to a particular choice.

7. Consider the 1 versus 5 classifier with $Q = 2$. We use E_{cv} to select $C \in \{0.0001, 0.001, 0.01, 0.1, 1\}$. If there is a tie in E_{cv}, select the smaller C. Within the 100 random runs, which of the following statements is correct?

[a] $C = 0.0001$ is selected most often.

[b] $C = 0.001$ is selected most often.

[c] $C = 0.01$ is selected most often.

[d] $C = 0.1$ is selected most often.

[e] $C = 1$ is selected most often.

8. Again, consider the 1 versus 5 classifier with $Q = 2$. For the winning selection in the previous problem, the average value of E_{cv} over the 100 runs is closest to

(a) 0.001

(b) 0.003

(c) 0.005

(d) 0.007

(e) 0.009

• RBF Kernel

Consider the radial basis function (RBF) kernel $K(\mathbf{x}_n, \mathbf{x}_m) = \exp\left(-||\mathbf{x}_n - \mathbf{x}_m||^2\right)$ in the soft-margin SVM approach. Focus on the 1 versus 5 classifier.

9. Which of the following values of C results in the lowest E_{in}?

[a] $C = 0.01$

[b] $C = 1$

[c] $C = 100$

[d] $C = 10^4$

[e] $C = 10^6$

10. Which of the following values of C results in the lowest E_{out}?

[a] $C = 0.01$

[b] $C = 1$

[c] $C = 100$

[d] $C = 10^4$

[e] $C = 10^6$

Learning From Data
Yaser Abu-Mostafa, *Caltech*
http://work.caltech.edu/telecourse
Self-paced version

Answer Key To Homework # 8

1. **[d]**
2. **[a]**
3. **[a]**
4. **[c]**
5. **[d]**
6. **[b]**
7. **[b]**
8. **[c]**
9. **[e]**
10. **[c]**

Visit the forum (http://book.caltech.edu/bookforum) for discussion. Please follow the forum guidelines for posting answers (see the "BEFORE posting answers" announcement at the top thread there).

Learning From Data
Yaser Abu-Mostafa, *Caltech*
http://work.caltech.edu/telecourse
Self-paced version

Final Exam

All questions have multiple-choice answers ([**a**], [**b**], [**c**], **...***). You can collaborate with others, but do not discuss the selected or excluded choices in the answers. You can consult books and notes, but not other people's solutions. Your solutions should be based on your own work. Definitions and notation follow the lectures.*

Note about the final

- There are twice as many problems in this final as there are in a homework set, and some problems require packages that will need time to get to work properly.
- Problems cover different parts of the course. To facilitate your search for relevant lecture parts, an indexed version of the lecture video segments can be found at the Machine Learning Video Library:

  http://work.caltech.edu/library
- To discuss the final, you are encouraged to take part in the forum

  http://book.caltech.edu/bookforum

  where there is a dedicated subforum for this final.
- Please follow the forum guidelines for posting answers (see the "BEFORE posting answers" announcement at the top there).

- **Nonlinear transforms**

1. The polynomial transform of order $Q = 10$ applied to $\mathcal{X}$ of dimension $d = 2$ results in a $\mathcal{Z}$ space of what dimensionality (not counting the constant coordinate $x_0 = 1$ or $z_0 = 1$)?

[a] 12

[b] 20

[c] 35

[d] 100

[e] None of the above

- **Bias and Variance**

2. Recall that the average hypothesis $\bar{g}$ was based on training the same model $\mathcal{H}$ on different data sets $\mathcal{D}$ to get $g^{(\mathcal{D})} \in \mathcal{H}$, and taking the expected value of $g^{(\mathcal{D})}$ w.r.t. $\mathcal{D}$ to get $\bar{g}$. Which of the following models $\mathcal{H}$ could result in $\bar{g} \notin \mathcal{H}$?

[a] A singleton $\mathcal{H}$ ($\mathcal{H}$ has one hypothesis)

[b] $\mathcal{H}$ is the set of all constant, real-valued hypotheses

[c] $\mathcal{H}$ is the linear regression model

[d] $\mathcal{H}$ is the logistic regression model

[e] None of the above

- **Overfitting**

3. Which of the following statements is *false*?

[a] If there is overfitting, there must be two or more hypotheses that have different values of E_{in}.

[b] If there is overfitting, there must be two or more hypotheses that have different values of E_{out}.

[c] If there is overfitting, there must be two or more hypotheses that have different values of $(E_{\text{out}} - E_{\text{in}})$.

[d] We can always determine if there is overfitting by comparing the values of $(E_{\text{out}} - E_{\text{in}})$.

[e] We cannot determine overfitting based on one hypothesis only.

4. Which of the following statements is true?

[**a**] Deterministic noise cannot occur with stochastic noise.

[**b**] Deterministic noise does not depend on the hypothesis set.

[**c**] Deterministic noise does not depend on the target function.

[**d**] Stochastic noise does not depend on the hypothesis set.

[**e**] Stochastic noise does not depend on the target distribution.

● Regularization

5. The regularized weight $\mathbf{w}_{\text{reg}}$ is a solution to:

$$\text{minimize} \;\; \frac{1}{N}\sum_{n=1}^{N}(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n - y_n)^2 \;\; \text{subject to} \;\; \mathbf{w}^{\mathrm{T}}\Gamma^{\mathrm{T}}\Gamma\mathbf{w} \le C,$$

where Γ is a matrix. If $\mathbf{w}_{\text{lin}}^{\mathrm{T}}\Gamma^{\mathrm{T}}\Gamma\mathbf{w}_{\text{lin}} \le C$, where $\mathbf{w}_{\text{lin}}$ is the linear regression solution, then what is $\mathbf{w}_{\text{reg}}$?

[**a**] $\mathbf{w}_{\text{reg}} = \mathbf{w}_{\text{lin}}$

[**b**] $\mathbf{w}_{\text{reg}} = \Gamma\mathbf{w}_{\text{lin}}$

[**c**] $\mathbf{w}_{\text{reg}} = \Gamma^{\mathrm{T}}\Gamma\mathbf{w}_{\text{lin}}$

[**d**] $\mathbf{w}_{\text{reg}} = C\Gamma\mathbf{w}_{\text{lin}}$

[**e**] $\mathbf{w}_{\text{reg}} = C\mathbf{w}_{\text{lin}}$

6. Soft-order constraints that regularize polynomial models can be

[**a**] written as hard-order constraints

[**b**] translated into augmented error

[**c**] determined from the value of the VC dimension

[**d**] used to decrease both E_{in} and E_{out}

[**e**] None of the above is true

● Regularized Linear Regression

We are going to experiment with linear regression for classification on the processed US Postal Service Zip Code data set from Homework 8. Download the data (extracted features of intensity and symmetry) for training and testing:

http://www.amlbook.com/data/zip/features.train http://www.amlbook.com/data/zip/features.test

(the format of each row is: **digit intensity symmetry**). We will train two types of binary classifiers; one-versus-one (one digit is class +1 and another digit is class −1, with the rest of the digits disregarded), and one-versus-all (one digit is class +1 and the rest of the digits are class −1). When evaluating E_{in} and E_{out}, use binary classification error. Implement the regularized least-squares linear regression for classification that minimizes

$$\frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\mathbf{z}_n - y_n\right)^2 \;+\; \frac{\lambda}{N}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

where $\mathbf{w}$ includes w_0.

7. Set $\lambda = 1$ and do not apply a feature transform (i.e., use $\mathbf{z} = \mathbf{x} = (1, x_1, x_2)$). Which among the following classifiers has the lowest E_{in}?

[**a**] 5 versus all

[**b**] 6 versus all

[**c**] 7 versus all

[**d**] 8 versus all

[**e**] 9 versus all

8. Now, apply a feature transform $\mathbf{z} = (1, x_1, x_2, x_1x_2, x_1^2, x_2^2)$, and set $\lambda = 1$. Which among the following classifiers has the lowest E_{out}?

[**a**] 0 versus all

[**b**] 1 versus all

[**c**] 2 versus all

[**d**] 3 versus all

[**e**] 4 versus all

9. If we compare using the transform versus not using it, and apply that to '0 versus all' through '9 versus all', which of the following statements is correct for $\lambda = 1$?

[**a**] Overfitting always occurs when we use the transform.

[**b**] The transform always improves the out-of-sample performance by at least 5% (E_{out} with transform $\leq 0.95E_{\text{out}}$ without transform).

[**c**] The transform does not make any difference in the out-of-sample performance.

[d] The transform always worsens the out-of-sample performance by at least 5%.

[e] The transform improves the out-of-sample performance of '5 versus all,' but by less than 5%.

10. Train the '1 versus 5' classifier with $\mathbf{z} = (1, x_1, x_2, x_1x_2, x_1^2, x_2^2)$ with $\lambda = 0.01$ and $\lambda = 1$. Which of the following statements is correct?

[a] Overfitting occurs (from $\lambda = 1$ to $\lambda = 0.01$).

[b] The two classifiers have the same E_{in}.

[c] The two classifiers have the same E_{out}.

[d] When λ goes up, both E_{in} and E_{out} go up.

[e] When λ goes up, both E_{in} and E_{out} go down.

• Support Vector Machines

11. Consider the following training set generated from a target function $f : \mathcal{X} \to \{-1, +1\}$ where $\mathcal{X} = \mathbb{R}^2$

$$\begin{array}{ccc} \mathbf{x}_1 = (1,0), y_1 = -1 & \mathbf{x}_2 = (0,1), y_2 = -1 & \mathbf{x}_3 = (0,-1), y_3 = -1 \\ \mathbf{x}_4 = (-1,0), y_4 = +1 & \mathbf{x}_5 = (0,2), y_5 = +1 & \mathbf{x}_6 = (0,-2), y_6 = +1 \\ & \mathbf{x}_7 = (-2,0), y_7 = +1 & \end{array}$$

Transform this training set into another two-dimensional space $\mathcal{Z}$

$$z_1 = x_2^2 - 2x_1 - 1 \qquad z_2 = x_1^2 - 2x_2 + 1$$

Using geometry (not quadratic programming), what values of $\mathbf{w}$ (without w_0) and b specify the separating plane $\mathbf{w}^{\mathrm{T}}\mathbf{z} + b = 0$ that maximizes the margin in the $\mathcal{Z}$ space? The values of w_1, w_2, b are:

[a] $-1,\ 1,\ -0.5$

[b] $1,\ -1,\ -0.5$

[c] $1,\ 0,\ -0.5$

[d] $0,\ 1,\ -0.5$

[e] None of the above would work.

12. Consider the same training set of the previous problem, but instead of explicitly transforming the input space $\mathcal{X}$, apply the hard-margin SVM algorithm with the kernel

$$K(\mathbf{x}, \mathbf{x}') = (1 + \mathbf{x}^{\mathrm{T}}\mathbf{x}')^2$$

(which corresponds to a second-order polynomial transformation). Set up the expression for $\mathcal{L}(\alpha_1...\alpha_7)$ and solve for the optimal $\alpha_1, ..., \alpha_7$ (numerically, using a quadratic programming package). The number of support vectors you get is in what range?

[a] 0-1

[b] 2-3

[c] 4-5

[d] 6-7

[e] >7

● Radial Basis Functions

We experiment with the RBF model, both in regular form (Lloyd + pseudo-inverse) with K centers:

$$\operatorname{sign}\left(\sum_{k=1}^{K} w_k \; \exp\left(-\gamma \; ||\mathbf{x} - \mu_k||^2\right) + \; b\right)$$

(notice that there is a bias term), and in kernel form (using the RBF kernel in hard-margin SVM):

$$\operatorname{sign}\left(\sum_{\alpha_n > 0} \alpha_n y_n \; \exp\left(-\gamma \; ||\mathbf{x} - \mathbf{x}_n||^2\right) + \; b\right).$$

The input space is $\mathcal{X} = [-1, 1] \times [-1, 1]$ with uniform probability distribution, and the target is

$$f(\mathbf{x}) = \operatorname{sign}(x_2 - x_1 + 0.25 \sin(\pi x_1))$$

which is slightly nonlinear in the $\mathcal{X}$ space. In each run, generate 100 training points at random using this target, and apply both forms of RBF to these training points. Here are some guidelines:

- Repeat the experiment for as many runs as needed to get the answer to be stable (statistically away from flipping to the closest competing answer).
- In case a data set is not separable in the '$\mathcal{Z}$ space' by the RBF kernel using hard-margin SVM, discard the run but keep track of how often this happens, if ever.

- When you use Lloyd's algorithm, initialize the centers to random points in $\mathcal{X}$ and iterate until there is no change from iteration to iteration. If a cluster becomes empty, discard the run and repeat.

13. For $\gamma = 1.5$, how often do you get a data set that is not separable by the RBF kernel (using hard-margin SVM)? *Hint: Run the hard-margin SVM, then check that the solution has* $E_{\text{in}} = 0$.

[**a**] $\leq 5\%$ of the time

[**b**] $> 5\%$ but $\leq 10\%$ of the time

[**c**] $> 10\%$ but $\leq 20\%$ of the time

[**d**] $> 20\%$ but $\leq 40\%$ of the time

[**e**] $> 40\%$ of the time

14. If we use $K = 9$ for regular RBF and take $\gamma = 1.5$, how often does the kernel form beat the regular form (excluding runs mentioned in Problem 13 and runs with empty clusters, if any) in terms of E_{out}?

[**a**] $\leq 15\%$ of the time

[**b**] $> 15\%$ but $\leq 30\%$ of the time

[**c**] $> 30\%$ but $\leq 50\%$ of the time

[**d**] $> 50\%$ but $\leq 75\%$ of the time

[**e**] $> 75\%$ of the time

15. If we use $K = 12$ for regular RBF and take $\gamma = 1.5$, how often does the kernel form beat the regular form (excluding runs mentioned in Problem 13 and runs with empty clusters, if any) in terms of E_{out}?

[**a**] $\leq 10\%$ of the time

[**b**] $> 10\%$ but $\leq 30\%$ of the time

[**c**] $> 30\%$ but $\leq 60\%$ of the time

[**d**] $> 60\%$ but $\leq 90\%$ of the time

[**e**] $> 90\%$ of the time

16. Now we focus on regular RBF only, with $\gamma = 1.5$. If we go from $K = 9$ clusters to $K = 12$ clusters (only 9 and 12), which of the following 5 cases happens most often in your runs (excluding runs with empty clusters, if any)? Up or down means strictly so.

[**a**] E_{in} goes down, but E_{out} goes up.

[**b**] E_{in} goes up, but E_{out} goes down.

[**c**] Both E_{in} and E_{out} go up.

[**d**] Both E_{in} and E_{out} go down.

[**e**] E_{in} and E_{out} remain the same.

17. For regular RBF with $K = 9$, if we go from $\gamma = 1.5$ to $\gamma = 2$ (only 1.5 and 2), which of the following 5 cases happens most often in your runs (excluding runs with empty clusters, if any)? Up or down means strictly so.

[**a**] E_{in} goes down, but E_{out} goes up.

[**b**] E_{in} goes up, but E_{out} goes down.

[**c**] Both E_{in} and E_{out} go up.

[**d**] Both E_{in} and E_{out} go down.

[**e**] E_{in} and E_{out} remain the same.

18. What is the percentage of time that regular RBF achieves $E_{\text{in}} = 0$ with $K = 9$ and $\gamma = 1.5$ (excluding runs with empty clusters, if any)?

[**a**] $\leq 10\%$ of the time

[**b**] $> 10\%$ but $\leq 20\%$ of the time

[**c**] $> 20\%$ but $\leq 30\%$ of the time

[**d**] $> 30\%$ but $\leq 50\%$ of the time

[**e**] $> 50\%$ of the time

- **Bayesian Priors**

19. Let $f \in [0, 1]$ be the unknown probability of getting a heart attack for people in a certain population. Notice that f is just a constant, not a function, for simplicity. We want to model f using a hypothesis $h \in [0, 1]$. Before we see any data, we assume that $P(h = f)$ is uniform over $h \in [0, 1]$ (the prior). We pick one person from the population, and it turns out that he or she had a heart attack. Which of the following is true about the posterior probability that $h = f$ given this sample point?

[**a**] The posterior is uniform over $[0, 1]$.

[**b**] The posterior increases linearly over $[0, 1]$.

[**c**] The posterior increases nonlinearly over $[0, 1]$.

[**d**] The posterior is a delta function at 1 (implying f has to be 1).

[**e**] The posterior cannot be evaluated based on the given information.

- **Aggregation**

20. Given two learned hypotheses g_1 and g_2, we construct the aggregate hypothesis g given by $g(\mathbf{x}) = \frac{1}{2}\left(g_1(\mathbf{x}) + g_2(\mathbf{x})\right)$ for all $\mathbf{x} \in \mathcal{X}$. If we use mean-squared error, which of the following statements is true?

[a] $E_{\text{out}}(g)$ cannot be worse than $E_{\text{out}}(g_1)$.

[b] $E_{\text{out}}(g)$ cannot be worse than the smaller of $E_{\text{out}}(g_1)$ and $E_{\text{out}}(g_2)$.

[c] $E_{\text{out}}(g)$ cannot be worse than the average of $E_{\text{out}}(g_1)$ and $E_{\text{out}}(g_2)$.

[d] $E_{\text{out}}(g)$ has to be between $E_{\text{out}}(g_1)$ and $E_{\text{out}}(g_2)$ (including the end values of that interval).

[e] None of the above

Learning From Data
Yaser Abu-Mostafa, *Caltech*
http://work.caltech.edu/telecourse
Self-paced version

Answer Key To Final Exam

1. **[e]**

2. **[d]**

3. **[d]**

4. **[d]**

5. **[a]**

6. **[b]**

7. **[d]**

8. **[b]**

9. **[e]**

10. **[a]**

11. **[c]**

12. **[c]**

13. **[a]**

14. **[e]**

15. **[d]**

16. **[d]**

17. **[c]**

18. **[a]**

19. **[b]**

20. **[c]**

Visit the forum (http://book.caltech.edu/bookforum) for discussion. Please follow the forum guidelines for posting answers (see the "BEFORE posting answers" announcement at the top thread there).

UNIVERSITY OF PENNSYLVANIA

CIS 520: Machine Learning

Final Exam, Fall 2017

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

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.

For all questions, select exactly one answer and fill in the corresponding circle on the bubble form. 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.*

The exam has 57 questions, totalling 98 points. These are distributed as follows:

Problems 1–31 are worth 1 point each (total 31 points).
Problems 32–42 are worth 2 points each (total 22 points).
Problems 43–57 are worth 3 points each (total 45 points).

Name: ________________________________

1. [1 points] *True or False?* Iterating between the E-step and M-step of EM algorithms always converges to a local optimum of the likelihood.

   (a) True
   (b) False

   ★ **SOLUTION:** A

2. [1 points] *True or False?* Lasso selects a subset (not necessarily strict) of the original features.

   (a) True
   (b) False

   ★ **SOLUTION:** A

3. [1 points] *True or False?* The features selected by PCA are linear combinations of the original features.

   (a) True
   (b) False

   ★ **SOLUTION:** A

4. [1 points] *True or False?* The solution to principal component analysis (PCA) can always be found using singular value decomposition (SVD).

   (a) True
   (b) False

   ★ **SOLUTION:** A

5. [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.

   (a) True
   (b) False

   ★ **SOLUTION:** A

6. [1 points] *True or False?* Principal Components Regression (PCR) generally yields models in which some of the original features do not affect the prediction.

   (a) True
   (b) False

★ **SOLUTION:** B

7. [1 points] *True or False?* The eigenvectors of AA^T and A^TA are the same for any matrix A.

   (a) True

   (b) False

★ **SOLUTION:** B

8. [1 points] Suppose you learn a model for binary classification using an algorithm $\mathcal{A}$. After learning this model you observe an additional instance-label pair $(\mathbf{x}, y)$, but you suspect that one of the components x_i has been corrupted with noise. To check this, you want to infer the probability $\Pr(x_i|y)$ using your model. Which of the following algorithms $\mathcal{A}$ would work best to find this probability?

   (a) Naive Bayes

   (b) Logistic Regression

★ **SOLUTION:** A

9. [1 points] *True or False?* K-means clustering can be kernelized using the kernel trick.

   (a) True

   (b) False

★ **SOLUTION:** A

10. [1 points] *True or False?* PCA does nonlinear orthogonal transformation of data into a lower dimensional space.

    (a) True

    (b) False

★ **SOLUTION:** B

11. [1 points] Consider data with n samples $D = \{\mathbf{x}^1, \ldots, \mathbf{x}^n\}$ with $\mathbf{x}^i \in \mathbb{R}^p$ with $n \gg p \gg k$, where k is the number of components to be kept. If computation speed is **not** as issue, it is advisable to do PCA by

    (a) computing the covariance matrix of X.

    (b) computing the SVD of X.

    (c) either (a) or (b). It won't make any difference.

★ **SOLUTION:** C

12. [1 points] *True or False?* PCA is a type of linear autoencoder.

    (a) True

    (b) False

★ **SOLUTION:** A

13. [1 points] *True or False?* Consider data with n samples $\mathbf{x}^1, \ldots, \mathbf{x}^n$ with $\mathbf{x}^i \in \mathbb{R}^p$. Given the number of principal components, the covariance matrix given by

$$\Sigma = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}^i - \overline{\mathbf{x}})(\mathbf{x}^i - \overline{\mathbf{x}})^\top$$

is **sufficient** to compute the reconstruction accuracy after applying PCA to the data.

    (a) True

    (b) False

★ **SOLUTION:** A

14. [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.

    (a) True

    (b) False

★ **SOLUTION:** A

15. [1 points] *True or False?* We observe data sampled from the model $y \sim \mathcal{N}(w^T x, \sigma^2)$. Since w is unknown, we would like to estimate it using linear regression. Adding priors to the parameters w, i.e. $w_j \sim \mathcal{N}(0, \lambda^2)$ is equivalent to adding an L_2 penalty on the parameters in the objective function defined by log-likelihood.

    (a) True

    (b) False

★ **SOLUTION:** A

16. [1 points] *True or False?* If some given data is not linearly separable, we can design a linear classifier that separates the data in a higher dimension as long as no point x appears twice with different y values.

    (a) True
    (b) False

    ★ **SOLUTION:** A

17. [1 points] *True or False?* L_1-penalized linear regression is unbiased.

    (a) True
    (b) False

    ★ **SOLUTION:** B

18. [1 points] *True or False?* Stepwise regression finds the global optimum, minimizing its loss function (squared error plus the usual L_0 penalty).

    (a) True
    (b) False

    ★ **SOLUTION:** B

19. [1 points] *True or False?* Inverting $X^T X$ for data sets with many more observations than features ($n \gg p$, where X is $n \times p$) is, in general, significantly slower than computing $X^T X$.

    (a) True
    (b) False

    ★ **SOLUTION:** B

20. [1 points] *True or False?* It is difficult to implement 'data-parallel' (e.g. map-reduce) algorithms for linear regression, as simple data-parallel methods come at a large cost in accuracy.

    (a) True
    (b) False

    ★ **SOLUTION:** B

21. [1 points] *True or False?* The complexity of contemporary deep learning systems for vision (as measured, for example, by the number of bits of information to specify them) is coming close to that of the human visual cortex.

    (a) True
    (b) False

★ **SOLUTION:** B

22. [1 points] *True or False?* Progress in machine learning has continued at a rapid pace over the past decade in spite of the fact that the cost of computing (as, for example, measured by the number of multiplies per second that can be done for a dollar) is no longer decreasing by a factor of two roughly every 18 months.

    (a) True

    (b) False

★ **SOLUTION:** B

23. [1 points] *True or False?* When doing machine learning on large data sets, it is good practice to test which algorithms work best on a small subset of the data before running the best model on the whole data set, since the same algorithms that work best on small data sets almost always also work best on big sets of the same data.

    (a) True

    (b) False

    ★ **SOLUTION:** B

24. [1 points] *True or False?* In the video assigned in class, the speaker (Killian Weinberger) argued that deep learning systems for image recognition approximately map images to a manifold in which certain directions correspond to meaningful directions such as faces appearing younger/older or more male/female.

    (a) True

    (b) False

    ★ **SOLUTION:** A

25. [1 points] *True or False?* The elastic net tends to select more features than well-optimized L_0 penalty methods.

    (a) True

    (b) False

    ★ **SOLUTION:** A

26. [1 points] *True or False?* PCA, when the data are mean centered, but not standardized, is scale invariant.

    (a) True

    (b) False

    ★ **SOLUTION:** B

27. [1 points] *True or False?* Stepwise regression (linear regression with an L_0 penalty) is scale invariant.

    (a) True

    (b) False

★ **SOLUTION:** A

28. [1 points] *True or False?* K-means clustering (using standard Euclidean distance) is scale invariant.

    (a) True

    (b) False

★ **SOLUTION:** B

29. [1 points] *True or False?* MLE is more likely to overfit than MAP since MAP tends to shrink parameters.

    (a) True

    (b) False

★ **SOLUTION:** A

30. [1 points] *True or False?* Consider a "true" distribution p given by

$$p(A) = 0.5,\ p(B) = 0.25,\ p(C) = 0.25$$

and an "approximating" distribution q given by

$$q(A) = 0.5,\ q(B) = 0.5,\ q(C) = 0\,.$$

The KL divergence $\mathrm{KL}(p\|q)$ is $\frac{1}{2}\log(2)$.

(a) True

(b) False

★ **SOLUTION:** B

31. [1 points] *True or False?* When you do principal components analysis on an $n * p$ observation matrix and keep k components, you get dimensions as follows: loadings: $n * k$, scores: $k * p$

(a) True

(b) False

★ **SOLUTION:** B

32. [2 points] Duplicating a feature in linear regression

(a) Does not reduce the L2-Penalized Residual Sum of Squares.

(b) Does not reduce the Residual Sum of Squares (RSS).

(c) Can reduce the L1-Penalized Residual Sum of Squares (RSS).

(d) None of the above

★ **SOLUTION:** B

33. [2 points] *True or False?* Given a set of data points $x_1, \cdots, x_n$, the K-means objective for finding cluster centers $\mu_1, \ldots, \mu_K$, and cluster assignments r_{ik}'s is as follows:

$$J(\mu, r) = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} \|\mu_k - x_i\|_2^2\,.$$

Minimizing this objective is a convex optimization problem.

(a) True

(b) False

★ **SOLUTION:** B

34. [2 points] We apply the Expectation Maximization algorithm to $f(D, Z, \theta)$ where D denotes the data, Z denotes the hidden variables and θ the variables we seek to optimize. Which of the following are correct?

    (a) EM will always return the same solution which may not be optimal

    (b) EM will always return the same solution which must be optimal

    (c) The solution depends on the initialization

★ **SOLUTION:** C

35. [2 points] In a convolutional neural net with an image of size 5x5x3 (where 3 is red/green/blue), we pad with a single zero all around the image and then use 4 local receptive fields ('filters') of size 3x3x3 and a stride of size 2. The outputs of these local receptive fields are sent to a single output. Assuming that there are no bias terms in this model, the total number of parameters (degrees of freedom) in the network is:

    (a) 3*3*4 + 4
    (b) 3*3*3*4 + 4
    (c) 3*3*3*4 + 4*9
    (d) 7*7*7*4 + 4
    (e) none of the above

    ★ **SOLUTION:** C

36. [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

37. [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

38. [2 points] The AdaBoost algorithm can be viewed as minimizing the:

    (a) exponential loss
    (b) logistic loss
    (c) hinge loss
    (d) squared loss
    (e) none of the above

    ★ **SOLUTION:** A

39. [2 points] In a supervised learning problem, the true quantity we really want to minimize is:

    (a) the training error
    (b) the test error
    (c) the generalization error
    (d) the cross-validation error
    (e) none of the above

★ **SOLUTION:** C

40. [2 points] Suppose you train two binary classifiers, h_1 and h_2, on the same training data, from two function classes $\mathcal{H}_1$ and $\mathcal{H}_2$ with $\text{VCdim}(\mathcal{H}_1) < \text{VCdim}(\mathcal{H}_2)$. Suppose h_1 and h_2 have the same training error. Then the VC-dimension based generalization error bound for h_1 is:

    (a) smaller than that for h_2
    (b) larger than that for h_2
    (c) equal to that for h_2
    (d) We can't say anything about the relationship between the two

★ **SOLUTION:** A

41. [2 points] Which of the following provides a discriminative learning algorithm/model for structured prediction?

    (a) Sum-product algorithm
    (b) Conditional random fields
    (c) Viterbi algorithm
    (d) Baum-Welch algorithm
    (e) None of the above

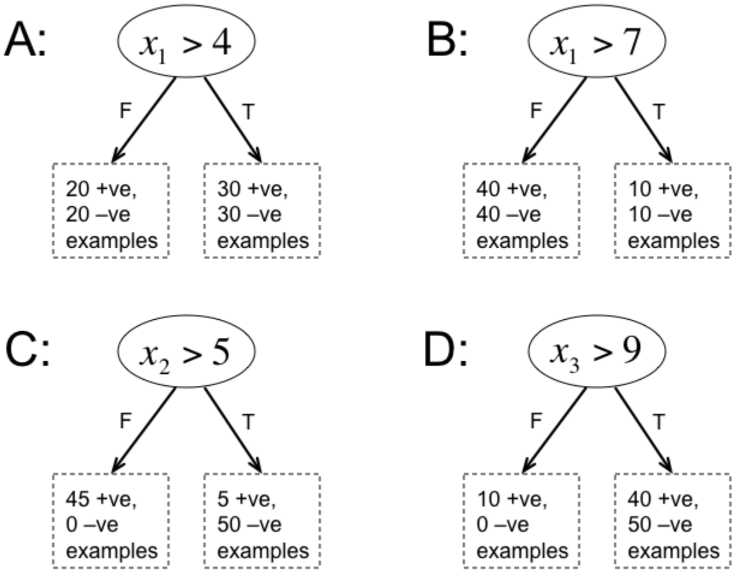
★ **SOLUTION:** B

42. [2 points] In reinforcement learning, a deterministic policy is

    (a) a mapping from states to states
    (b) a mapping from state-action pairs to states
    (c) a mapping from actions to states
    (d) a mapping from states to actions
    (e) none of the above

★ **SOLUTION:** D

43. [3 points] Suppose you have a binary classification problem with 3-dimensional feature vectors $\mathbf{x} \in \mathbb{R}^3$. You are given 50 positive and 50 negative training examples, and want to build a decision tree classifier. Consider 4 possible splits at the root node:



Which of the above splits gives the highest information gain?

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

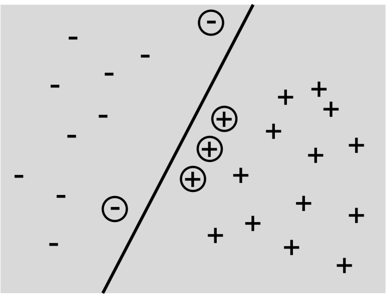
★ **SOLUTION:** C

44. [3 points] Consider modeling observations with 2 features using a Gaussian mixture model with 3 mixture components, where each component can have a different (full) covariance matrix. How many parameters are needed?

(a) 12
(b) 14
(c) 15
(d) 17
(e) none of the above

★ **SOLUTION:** D (17 = 3*2 (means) + 3*3 (covariance matrices) + 2 (mixing coefficients))

45. [3 points] Consider a binary classification problem in a 2-dimensional instance space $\mathcal{X} = \mathbb{R}^2$. You are given a linearly separable training set. You run the hard-margin SVM algorithm and obtain the separating hyperplane below (support vectors are circled):



What is the smallest number of data points that would have to be removed from the training set in order for the SVM solution to change?

(a) 1
(b) 2
(c) 3
(d) 4
(e) None of the above

★ **SOLUTION:** A

46. [3 points] Consider a binary classification problem in a d-dimensional instance space $\mathcal{X} = \mathbb{R}^d$. Your friend has a training set containing m labeled examples. She computes two $m \times m$ kernel matrices, K_1 and K_2, and gives them to you. The first matrix K_1 is obtained by applying an RBF kernel $K(\mathbf{x}, \mathbf{x}') = e^{-\|\mathbf{x}-\mathbf{x}'\|^2}$ to the original d-dimensional data points. The second matrix K_2 is obtained by selecting $r < d$ features and applying the RBF kernel (with the same width parameter as before) to the reduced r-dimensional data points. You are given these two kernel matrices and are asked to train an SVM classifier in each case. The training time for K_2 will be:

(a) smaller than that for K_1
(b) greater than that for K_1
(c) roughly similar to that for K_1

★ **SOLUTION:** C

47. [3 points] Consider running the perceptron algorithm for an online binary classification task. Recall that on each round t, the algorithm receives an instance $\mathbf{x}_t$ and uses the current weight vector $\mathbf{w}_t$ to predict $\widehat{y}_t = \text{sign}(\mathbf{w}_t^\top \mathbf{x}_t)$; it then receives the true label y_t, and if it made a mistake, it updates the weight vector as

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_t \, \mathbf{x}_t \, .$$

Suppose that on a particular round t, the algorithm predicts $\widehat{y}_t = -1$ and receives the true label $y_t = +1$. Assume $\mathbf{x}_t \neq \mathbf{0}$. In this case, after updating the weight vector, it is guaranteed that:

(a) $\mathbf{w}_{t+1}^\top \mathbf{x}_t > \mathbf{w}_t^\top \mathbf{x}_t$
(b) $\mathbf{w}_{t+1}^\top \mathbf{x}_t < \mathbf{w}_t^\top \mathbf{x}_t$
(c) $\mathbf{w}_{t+1}^\top \mathbf{x}_t > 0$
(d) $\mathbf{w}_{t+1}^\top \mathbf{x}_t < 0$
(e) none or more than one of the above

★ **SOLUTION:** A

48. [3 points] Consider running the AdaBoost algorithm for a binary classification problem in which you are given a small training set of 5 examples, $\{(x_i, y_i)\}_{i=1}^5$. In the first round, all 5 examples have equal weight, $D_1(i) = 1/5$. Suppose that the true labels and the predictions made by the weak classifier h_1 learned in the first round are as follows:

i	y_i	$h_1(x_i)$
1	-1	-1
2	-1	$+1$
3	$+1$	-1
4	$+1$	$+1$
5	$+1$	$+1$

In the second round, how many of the 5 examples will receive a higher weight than they had in the first round?

(a) 1
(b) 2
(c) 3
(d) 4
(e) none of the above

★ **SOLUTION:** B

49. [3 points] Consider a binary classification problem with the following loss function:

		$\widehat{y}$	
		-1	$+1$
y	-1	0	0.8
	$+1$	0.2	0

For a particular instance x, your class probability estimation (CPE) model $\widehat{\eta}$ predicts the probability of a positive label to be $\widehat{\eta}(x) = 0.75$. To minimize expected loss, the predicted label $\widehat{y}$ for this instance should be:

(a) $+1$

(b) -1

(c) Both are equally good

★ **SOLUTION:** B

50. [3 points] Consider a binary classification problem in which the label $+1$ is rare. You have learned a binary classifier $h : \mathcal{X} \rightarrow \{\pm 1\}$ from some training data. On a test set of 100 data points, the classifier's predictions, as well as the true labels, are as follows:

		$h(x)$	
		-1	$+1$
y	-1	75	15
	$+1$	3	7

What is the true positive rate (TPR) of the classifier h on the above test set?

(a) 3/10

(b) 7/10

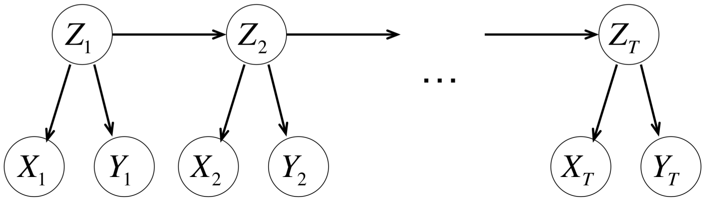
(c) 3/78

(d) 7/22

(e) None of the above

★ **SOLUTION:** B

51. [3 points] Consider a variant of a hidden Markov model in which each hidden state Z_t generates two conditionally independent observations X_t, Y_t:



Suppose each hidden state Z_t takes one of K possible values, each observation X_t takes one of M_1 possible values, and each observation Y_t takes one of M_2 possible values. Assume the model is homogeneous, so that transition and emission probabilities are the same for all t. What is the total number of parameters in this model? Choose the tightest expression below.

(a) $O(K^2 M_1 M_2)$
(b) $O(K^2(M_1 + M_2))$
(c) $O(K^2 + K M_1 M_2)$
(d) $O(K^2 + K(M_1 + M_2))$
(e) None of the above

★ **SOLUTION:** D

52. [3 points] Consider learning a (homogeneous) hidden Markov model for a part-of-speech tagging task, where observations X_t are words and hidden states Z_t are parts of speech such as 'noun' (N), 'verb' (V), 'determiner' (D), etc. You are given labeled training data consisting of the following two sentences with corresponding parts of speech:

N	V	N
Mary	likes	mountains

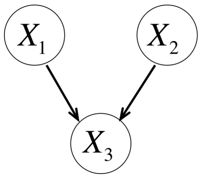
D	N	V	D	N
The	dog	ate	the	candy

What is the maximum likelihood estimate of the transition probability $A_{\mathsf{V},\mathsf{N}} = P(Z_{t+1} = \mathsf{N} \mid Z_t = \mathsf{V})$?

(a) 1
(b) 1/3
(c) 1/4
(d) 0
(e) None of the above

★ **SOLUTION:** E (Answer: 1/2)

53. [3 points] Consider three random variables X_1, X_2, X_3, each of which takes one of 2 possible values. Suppose their joint probability distribution is known to factor according to the Bayesian network structure below:

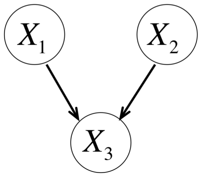


Given this information, how many parameters are needed to specify the joint probability distribution?

(a) 7
(b) 6
(c) 5
(d) 4
(e) None of the above

★ **SOLUTION:** B

54. [3 points] Consider a probability distribution that factors according to the Bayesian network structure below:

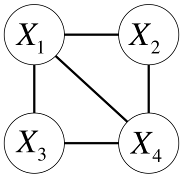


Which of the following (conditional) independence statements *must* be true?

(a) $X_1 \perp\!\!\!\perp X_3 \,|\, X_2$
(b) $X_1 \perp\!\!\!\perp X_2 \,|\, X_3$
(c) $X_1 \perp\!\!\!\perp X_3$
(d) $X_1 \perp\!\!\!\perp X_2$
(e) None of the above

★ **SOLUTION:** D

55. [3 points] Consider a probability distribution that factors according to the Markov network structure below:



Which of the following (conditional) independence statements *must* be true?

(a) $X_2 \perp\!\!\!\perp X_3 \,|\, X_1$

(b) $X_1 \perp\!\!\!\perp X_4 \,|\, \{X_2, X_3\}$

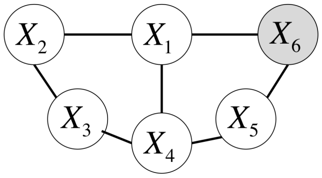
(c) $X_2 \perp\!\!\!\perp X_3 \,|\, \{X_1, X_4\}$

(d) $X_2 \perp\!\!\!\perp X_3$

(e) None of the above

★ **SOLUTION:** C

56. [3 points] Consider 6 random variables $X_1, \ldots, X_6$, each of which takes one of K possible values. You are given their joint probability distribution, which factors according to the Markov network structure below; you are also told that X_6 takes the value $\overline{x}_6$.



You are asked to find the posterior probability distribution $p(x_1 \,|\, \overline{x}_6)$. You decide to use the variable elimination algorithm and eliminate variables in the order $(5, 4, 3, 2)$. As a function of K, how many computations will you need? Choose the tightest expression below.

(a) $O(K^2)$

(b) $O(K^3)$

(c) $O(K^4)$

(d) $O(K^5)$

(e) None of the above

★ **SOLUTION:** B

57. [3 points] Consider an active learning setup for binary classification with labels $\{\pm 1\}$ and 0-1 loss. 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, $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, and $\mathbf{x}_4$, 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:

$$\widehat{\eta}(\mathbf{x}_1) = 0.30\,; \quad \widehat{\eta}(\mathbf{x}_2) = 0.40\,; \quad \widehat{\eta}(\mathbf{x}_3) = 0.55\,; \quad \widehat{\eta}(\mathbf{x}_4) = 0.75\,.$$

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

★ **SOLUTION:** C

UNIVERSITY OF PENNSYLVANIA

CIS 520: Machine Learning
Midterm, 2018

Exam policy: This exam allows one one-page, two-sided cheat sheet; No other materials.

Time: 80 minutes. Be sure to write your name and Penn student ID (the 8 bigger digits on your ID card) on the answer form and fill in the associated bubbles *in pencil*.

If you think a question is ambiguous, mark what you think is the best answer. 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.*

There are **46** questions, worth a total of **70** points.

1. [1 points] MLE estimation of a model $y = f(x;\theta) + \epsilon$ where ϵ is mean zero Gaussian noise is the same as minimizing the ___ error.

   (a) L_0
   (b) L_1
   (c) L_2
   (d) L_∞

   ★ **SOLUTION:** B

2. [2 points] Increasing k in k-nearest neighbor models will:

   (a) Increase bias, increase variance
   (b) Increase bias, decrease variance
   (c) Decrease bias, increase variance
   (d) Decrease bias, decrease variance

   ★ **SOLUTION:** B

3. [2 points] In a least-squares linear regression problem, adding an L_2 regularization penalty always decreases the expected sum of squares error of the solution w on unseen test data.

   (a) True
   (b) False

   ★ **SOLUTION:** B

4. [1 points] In a least-squares linear regression problem, adding an L_2 regularization penalty cannot decrease the sum of squares error of the solution w on the training data.

   (a) True
   (b) False

★ **SOLUTION:** A

5. [2 points] We have some data $D = \{x_i, y_i\}$, and we assume a simple linear model of this data with Gaussian noise as follows:

   $Y = w^\top X + b + Z$, with $Z \sim N(0, \sigma^2)$

   We will further assume a prior on w, that means $w_j \sim N(0, \lambda^2)$. Then, in which case does MAP **not** reduce to MLE?

   (a) $\lambda \to \infty$

   (b) $\sigma \to \infty$

   (c) $N \to \infty$ (Here N means the number of samples)

   (d) $\frac{\lambda}{\sigma} \to \infty$

★ **SOLUTION:** B

6. [2 points] Suppose we want to fit the following regression prediction model: $h(x) = c$, which is constant for all x. Suppose the actual underlying model that generated the data is $y = ax$, where a is a constant slope. In other words, we are modeling the underlying linear relation with a constant model. Let us now try to compute the bias and variance of our method. Assume that $x \sim N(\mu, \sigma^2)$. Compute the **average** hypothesis $h(x)$ over datasets $D = \{x_1, ..., x_n\}$ (Here we use the ordinary least squares estimate to estimate $h(x; D)$):

   (a) $a\mu$

   (b) $\frac{a}{\sigma^2}$

   (c) $\frac{a}{\sigma}$

   (d) $\frac{a}{\mu}$

★ **SOLUTION:** A

7. [1 points] For any probability model, the log-likelihood function of data generated from this model is always guaranteed to be concave.

   (a) True

   (b) False

★ **SOLUTION:** False

8. [1 points] Any Naïve Bayes classifier that assumes the features are drawn from Gaussian distributions can always be written as a linear classifier.

   (a) True

   (b) False

★ **SOLUTION:** False

9. [1 points] Which type of regularization leads to sparser solutions (fewer non-zero weights)?

   (a) L_2

   (b) L_1

   (c) neither

★ **SOLUTION:** B

10. [1 points] If the complexity of a model increases, then which of the following is expected to increase?

   (a) Bias

   (b) Variance

★ **SOLUTION:** B

11. [2 points] The L_0 pseudo-norm of a vector $\mathbf{w}$ of length n is defined as

   (a) $\sum_{i=1}^{n} |w_i|$

   (b) $\sum_{i=1}^{n} \mathbb{I}(w_i \neq 0)$, where $\mathbb{I}$ takes value 1 when $w_i \neq 0$, and 0 otherwise.

   (c) $\sqrt{\sum_{i=1}^{n} w_i^2}$

   (d) None of the above

★ **SOLUTION:** B

12. [2 points] The solution to the following L_1 regularized least-squares regression

$$\operatorname*{argmin}_{\mathbf{w}} \|Y - X\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1 \,,$$

for $\lambda > 0$ is:

(a) $(X^\top X)^{-1} X^\top Y$

(b) $(X^\top X + \lambda I)^{-1} X^\top Y$

(c) The objective is unbounded, i.e. the solution is $-\infty$.

(d) None of the above

★ **SOLUTION:** D

13. [1 points] Which of the following sequences has a higher description length (in bits) – a) 111111111111.... or b) 010101010101.....?

(a) a

(b) b

★ **SOLUTION:** B

14. [2 points] In MDL, reducing the description length of a model also reduces the description length of the residual error of that model.

(a) True

(b) False

★ **SOLUTION:** False

15. [1 points] Kernels such as radial basis functions can often be used with support vector machines to make data linearly separable, even when the data was not linearly separable before the kernel was applied.

(a) True

(b) False

★ **SOLUTION:** True

16. [1 points] When doing 10-fold cross-validation, standard practice for making predictions on new points is to take an ensemble (e.g. average or majority vote) of the ten models which are built.

    (a) True

    (b) False

★ **SOLUTION:** False

17. [1 points] Given features $X = [X_1, X_2, \ldots, X_p]$ and output Y, the reduction in the entropy of Y from observing feature X_p is given by $H(X_p) - H(X_p|Y)$

    (a) True

    (b) False

★ **SOLUTION:** True

18. [1 points] The entropy for a Gaussian, Y, given by $H(Y) = -\int p(y)\log(p(y))dy$ is independent of its mean, $E[Y]$.

    (a) True

    (b) False

★ **SOLUTION:** True

19. [2 points] $\mathbf{E}(f(X))$ for a continuous random variable X is given by

    (a) $f(\mathbf{E}(X))$ for all functions f

    (b) $f(\mathbf{E}(X))$ if $f(X)$ is a linear function of X

    (c) $\int f(x)p(x)dx$ where $p(x)$ is the probability density function of X

    (d) both (b) and (c)

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

★ **SOLUTION:** D

20. [1 points] In logistic regression, adding Gaussian priors to the parameters w, i.e. $w_j \sim \mathcal{N}(0, \lambda^2)$ is equivalent to adding a quadratic penalty on the parameters in the objective function defined by the loglikelihood.

    (a) True

    (b) False

★ **SOLUTION:** True

21. [1 points] Adding an L_1 penalty on the parameters in a regression problem is equivalent to a prior that the weights are small and will, in general, shrink all of the parameters.

    (a) True

    (b) False

★ **SOLUTION:** True

22. [1 points] The curve defined by $\|x\|_{0.5} = 2$ (where $\|x\|_{0.5}$ is the $L_{0.5}$ norm) is a convex set.

    (a) True

    (b) False

★ **SOLUTION:** False

23. [2 points] The following function can be interpreted as a probability density:

$$f(x) = \begin{cases} 2, & |x| \leq 1/2 \\ 0, & |x| > 1/2 \end{cases}$$

    (a) True

    (b) False

★ **SOLUTION:** False

24. [1 points] The expected value of the testing error in approximating a true y by a model $\widehat{y} = f(x;\theta)$ is equal to the sum of the expected value of the bias on the training set plus the expected value of the variance on the training set, plus the irreducible uncertainty (the variance of the noise).

    (a) True

    (b) False

★ **SOLUTION:** False

25. [2 points] Assume a variable can take on three values, A, B, and C with probabilities either given by $p = [p_A, p_B, p_C] = [1/2, 1/4, 1/4]$ (I.e., $p_A = 1/2, p_B = 1/4, p_C = 1/4)$) or by $q = [q_A, q_B, q_C] = [1/2, 1/2, 0]$. The KL divergence $KL(p, q)$ is equal to

    (a) 0

    (b) $-(1/2)\log(1/2) - (1/2)\log(1/4)$

    (c) infinity

    (d) none of the above

★ **SOLUTION:** C

26. [2 points] The MLE estimate of weights $\widehat{w}$ for ordinary least squares gives estimates of the (unknown) true weight w that are distributed as $\widehat{w} \sim N(w, \sigma^2/n)$. If we use Ridge regression instead, the expected value of $\widehat{w}$ will

    (a) remain $\widehat{w}$

    (b) become larger in magnitude

    (c) become smaller in magnitude

    (d) we can't say

★ **SOLUTION:** C

27. [2 points] The MLE estimate of weights for ordinary least squares gives $\widehat{w} \sim N(w, \sigma^2/n)$. If we use Ridge regression instead, the variance of $\widehat{w}$ will

    (a) remain σ^2/n

    (b) become larger in magnitude

    (c) become smaller in magnitude

    (d) we can't say

★ **SOLUTION:** C

28. [1 points] The training error of 1-NN is always zero.

    (a) True

    (b) False

★ **SOLUTION:** either

29. [1 points] A classifier trained on less training data is less likely to overfit.

    (a) True

    (b) False

★ **SOLUTION:** False

30. [1 points] A gradient descent algorithm with a properly chosen fixed step size for training a logistic regression model almost always converges to the exact value of the optimal regression weights.

    (a) True

    (b) False

★ **SOLUTION:** either

31. [2 points] Let $X_1, X_2 \ldots X_n$ be iid samples from Uniform$(-w, w)$, i.e.

$$f_X(x) = \begin{cases} 0 & \text{if } |x| > w \\ \frac{1}{2w} & \text{if } |x| \le w \end{cases}$$

where $w > 0$ is an unknown parameter. The MLE estimate of w is

(a) $\frac{\sum_{i=1}^{n} |X_i|}{n}$

(b) $\frac{n}{\sum_{i=1}^{n} 2|X_i|}$

(c) $\max_i |X_i|$

(d) $\max_i \frac{1}{2|X_i|}$

★ **SOLUTION:** C

32. [2 points] Consider the two class problem where class label $y \in \{T, F\}$ and each training example X has 2 binary attributes X_1, $X_2 \in \{T, F\}$. How many parameters will you need to know/estimate if you are to classify an example using the Naïve Bayes classifier?

(a) 5

(b) 8

(c) 3

(d) 7

★ **SOLUTION:** A

33. [2 points] Again, consider the two class problem where class label $y \in \{T, F\}$ and each training example X has 2 binary attributes X_1, X_2 $\in \{T, F\}$. How many parameters will you need to estimate if we do **not** make the Naïve Bayes conditional independence assumption?

(a) 3

(b) 5

(c) 7

(d) 8

★ **SOLUTION:** C

34. [2 points] Consider the following two sets in the two-dimensional plane:

$$C = \{\mathbf{x} \in \mathbb{R}^2 \mid 0 \leq x_1 \leq 2\} \cap \{\mathbf{x} \in \mathbb{R}^2 \mid 2 \leq x_1 \leq 4\}$$
$$D = \{\mathbf{x} \in \mathbb{R}^2 \mid 0 \leq x_1 \leq 2\} \cup \{\mathbf{x} \in \mathbb{R}^2 \mid 2 \leq x_1 \leq 4\}$$

Select the most accurate statement:

(a) Both C and D are convex.
(b) C is convex, D is not convex.
(c) C is not convex, D is convex.
(d) Neither C nor D is convex.

★ **SOLUTION:** A

35. [2 points] Consider the following functions of two variables:

$$f(\mathbf{x}) = \max(x_1, -x_2)$$
$$g(\mathbf{x}) = -x_1^2 - x_2^2$$

Select the most accurate statement:

(a) Both f and g are convex.
(b) f is convex, g is concave.
(c) f is convex, g is neither convex nor concave.
(d) Neither f nor g is convex.

★ **SOLUTION:** B

36. [2 points] Consider a primal optimization problem with two inequality constraints, and suppose you form the Lagrange dual problem over two dual variables λ_1, λ_2. Which of the following functions of two variables *cannot* be the objective function of the dual problem?

(a) $\phi(\boldsymbol{\lambda}) = \lambda_1 - \lambda_2$
(b) $\phi(\boldsymbol{\lambda}) = \lambda_1^2 + \lambda_2^2$
(c) $\phi(\boldsymbol{\lambda}) = \lambda_1^2 - \lambda_2^2$
(d) All three can be dual objective functions.

★ **SOLUTION:** *

37. [2 points] Let $K_1 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $K_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be two symmetric, positive definite kernel functions. Which of the following *cannot* be a valid kernel function?

    (a) $K(x, x') = 5 \cdot K_1(x, x')$

    (b) $K(x, x') = K_1(x, x') + K_2(x, x')$

    (c) $K(x, x') = K_1(x, x') + \dfrac{1}{K_2(x, x')}$

    (d) All three are valid kernels.

★ **SOLUTION:** *

38. [2 points] You are trying to impress your friend with your photographs. Over the last few months, you have observed which photographs she gives a 'like' to and which she does not. Based on these examples, you want to estimate the probability that she will like your new photograph. Which of the following machine learning methods would be **least** useful for this problem?

    (a) Logistic regression

    (b) k-nearest neighbor

    (c) Support vector machines

★ **SOLUTION:** either B or C

39. [2 points] The linear (soft margin) support vector machine algorithm learns a weight vector $\mathbf{w}$ (and possibly a bias term b). What sort of regularization does it effectively perform on $\mathbf{w}$?

    (a) L_1 regularization

    (b) L_2 regularization

    (c) regularization other than L_1 and L_2

    (d) no regularization

★ **SOLUTION:** B

40. [2 points] Suppose you are training a support vector machine classifier using a polynomial kernel $K(\mathbf{x}, \mathbf{x}') = (\mathbf{x} \cdot \mathbf{x}' + 1)^q$ for degrees $q = 1, 2, 3, 4$. Assuming that for each value of q, you select the SVM parameter C by cross-validation over a large enough range of values, which of the following scenarios are realistic (could be expected to arise in practice)?

(a)

q	1	2	3	4
Train error	0.35	0.30	0.24	0.17
Test error	0.25	0.21	0.18	0.23

(b)

q	1	2	3	4
Train error	0.17	0.24	0.29	0.33
Test error	0.29	0.25	0.32	0.36

(c)

q	1	2	3	4
Train error	0.24	0.21	0.19	0.13
Test error	0.31	0.26	0.23	0.27

(d) All three scenarios are realistic.

★ **SOLUTION:** C

41. [2 points] Suppose you are solving a regression problem. You have 1000 labeled examples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{1000}$ and decide to use ridge regression:

$$\min_{\mathbf{w}} \sum_{i=1}^{1000} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2 + \lambda \|\mathbf{w}\|_2^2 .$$

To choose λ, you perform cross-validation over some range of values; let's say a value λ_1 is selected. After some time, you get another 1000 labeled examples $\{(\mathbf{x}_i, y_i)\}_{i=1001}^{2000}$. You decide to re-learn your model using all the data; you again select λ using cross-validation. Let's call this second value λ_2. What is the relationship you would expect between λ_1 and λ_2 ?

(a) λ_1 is expected to be smaller than λ_2

(b) λ_2 is expected to be smaller than λ_1

(c) λ_1 and λ_2 will be approximately the same size

(d) there is not enough information to tell

★ **SOLUTION:** *

42. [2 points] In a convolutional neural net with an image of size 5x5x3 (where 3 is red/green/blue), we pad with a single zero all around the image and then use 4 local receptive fields ('filters') of size 3x3x3 and a stride of size 2. The outputs of these local receptive fields are sent to a single output. Assuming that there are no bias terms in this model, the total number of weights in the network is.

    (a) 3*3*4 + 4

    (b) 3*3*3*4 + 4

    (c) 7*7*4 + 4

    (d) 7*7*7*4 + 4

    (e) none of the above

★ **SOLUTION:** E

43. [1 points] When learning neural networks, one should always use an L_2 loss function rather than a log-likelihood.

    (a) True

    (b) False

★ **SOLUTION:** False

44. [1 points] When learning a decision tree, a feature x_j which is not correlated with the label y (i.e., $corr(x_j, y) = 0$) will never be split on and hence will never to used in the tree.

    (a) True

    (b) False

★ **SOLUTION:** False

45. [1 points] When learning a decision tree, features that have many possible values (e.g. colors or countries) tend to be more likely to be selected than binary features.

    (a) True

    (b) False

★ **SOLUTION:** True

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.

   ★ **SOLUTION:** True

2. [1 points] *True or False?* The largest eigenvector of the covariance matrix is the direction of minimum variance in the data.

   ★ **SOLUTION:** False

3. [1 points] *True or False?* The non-zero eigenvalues of $AA^\top$ and $A^\top A$ are the same.

   ★ **SOLUTION:** True

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 $A^\top A$

   ★ **SOLUTION:** C

5. [1 points] *True or False?* PCA is a type of linear autoencoder.

   ★ **SOLUTION:** True

6. [1 points] *True or False?* A GAN may be trained via backpropogation alone.

   ★ **SOLUTION:** True

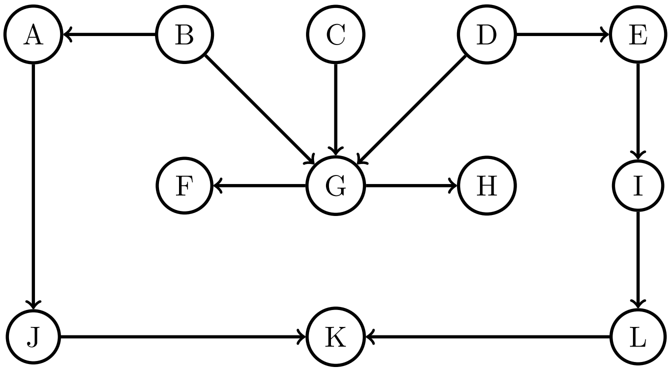
7. [1 points] *True or False?* For $\mathbf{x} \in \mathbb{R}^d$, $f(\mathbf{x}) = \max\big(||\mathbf{x}||, \log(e^{x_1} + \cdots + e^{x_d})\big)$ is convex.

   ★ **SOLUTION:** True

8. [1 points] *True or False?* $k(x, y) = \exp(-||x - y||)$ is a valid kernel.

★ **SOLUTION:** True

The following seven questions refer to this figure:



9. [1 points] *True or False?* $C \perp\!\!\!\perp D \mid F$

★ **SOLUTION:** False

10. [1 points] *True or False?* $D \perp\!\!\!\perp I \mid E, F, K$

★ **SOLUTION:** False

11. [1 points] *True or False?* $C \perp\!\!\!\perp J \mid A, F, L$

★ **SOLUTION:** True

12. [1 points] *True or False?* $F \perp\!\!\!\perp L \mid G$

★ **SOLUTION:** True

13. [1 points] *True or False?* $\neg(G \perp\!\!\!\perp E \mid D, K)$

★ **SOLUTION:** True

14. [1 points] *True or False?* I d-separates E and L

★ **SOLUTION:** True

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

★ **SOLUTION:** E

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.

★ **SOLUTION:** A

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

★ **SOLUTION:** B â

18. [2 points] You are processing the data of a survey where people have the option to report their income. We know 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

★ **SOLUTION:** C

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 $O(|S|^2|A|)$.

★ **SOLUTION:** EITHER

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

★ **SOLUTION:** A

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

★ **SOLUTION:** A

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.

★ **SOLUTION:** False

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 $\epsilon$-soft policy
Repeat forever:
    Generate an episode using $\pi$
    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

★ **SOLUTION:** D

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

★ **SOLUTION:** B

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.

★ **SOLUTION:** EITHER Full credit as it was too obscure, but the correct answer is C.

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

★ **SOLUTION:** B

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

★ **SOLUTION:** A

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 $x_1, \dots x_T$.

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

★ **SOLUTION:** False

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

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

★ **SOLUTION:** True

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.

★ **SOLUTION:** False

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

★ **SOLUTION:** E The question has nothing to do with EM; it's just counting how many

parameters are in this naive Bayes model so you have 1 for $p(y = 1)$, 7 for $p(x_1|y = 1)$, 7 for $p(x_1|y = -1)$, 7 for $p(x_2|y = 1)$, 7 for $p(x_2|y = -1)$ =29 in total

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:

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

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

★ **SOLUTION:** A

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

★ **SOLUTION:** C Both logistic regression and linear SVM give us a hyperplane for clas-

sification in original space so you only need O(p) steps for prediction. But polynomial SVM requires O(M+p) steps where M is the number of support vectors.

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

★ **SOLUTION:** E

35. [1 points] *True or False?* After the $i - th$ 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?

★ **SOLUTION:** False

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

★ **SOLUTION:** E HMMs also treat words as independent given latent variable so its not

'unlike' (sorry; his is perhaps a bit confusing in how it was phrased)

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

★ **SOLUTION:** C

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

★ **SOLUTION:** D

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

★ **SOLUTION:** A or D

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

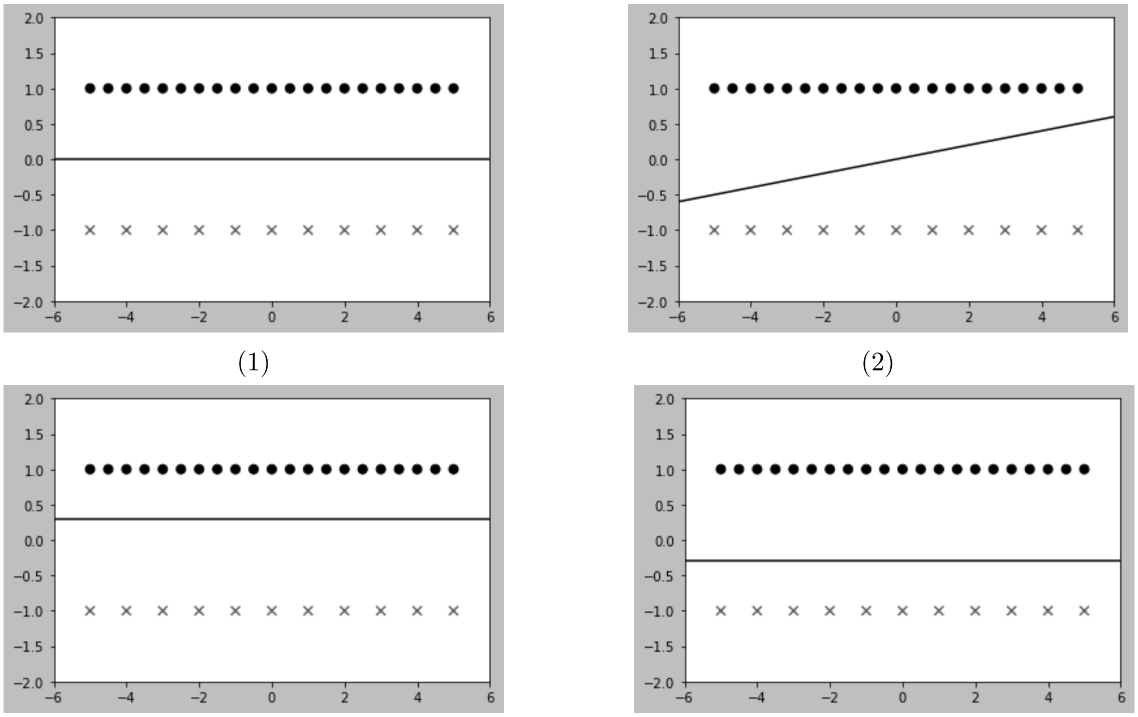
    (c) Adding a fourth hidden layer

    (d) Using ReLU activations instead of sigmoid

    (e) Using an L_1 penalty instead of L_2

★ **SOLUTION:** B **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

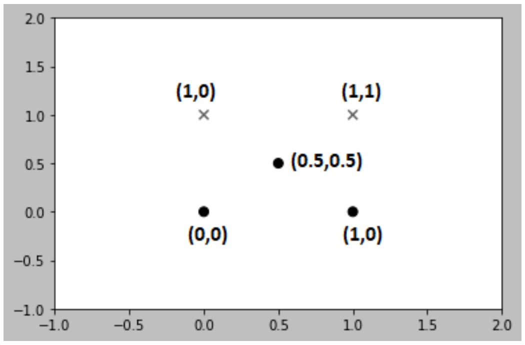
    ★ **SOLUTION:** D

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

    ★ **SOLUTION:** B

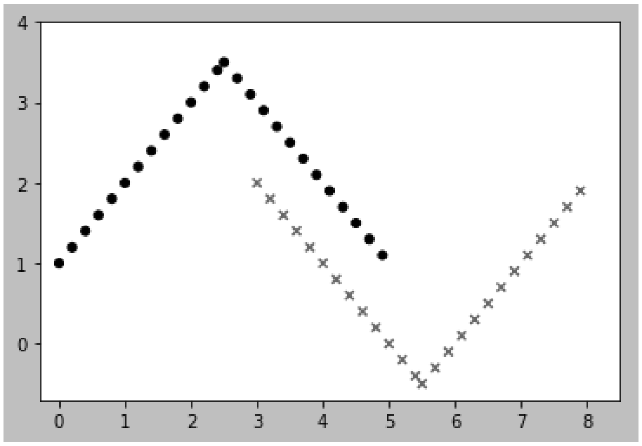
43. [1 points] The following data set consists of 5 points: each corner of a unit square and its center. Can this data set be made separable by an SVM with an RBF kernel using **only two** support vectors? (There is no restriction on the kernel width or choice of support vectors.)



    (a) True
    (b) False

★ **SOLUTION:** EITHER Too confusing.

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

★ **SOLUTION:** C Example solution: the point (x, y) is classified as dot if $f(x, y) =$

$2ReLU(y - x) - ReLU(6 - x - y) - ReLU(x - 4 - y) \geq 0$

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

★ **SOLUTION:** A

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

★ **SOLUTION:** B

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.

★ **SOLUTION:** D

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

★ **SOLUTION:** B

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

★ **SOLUTION:** C

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

★ **SOLUTION:** D

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.

★ **SOLUTION:** False

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

★ **SOLUTION:** C 'query by majority' does not exist; it was 'query by committee'

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,

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

★ **SOLUTION:** C

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.

★ **SOLUTION:** B

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

★ **SOLUTION:** B

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).

★ **SOLUTION:** D It is generally a bad idea to mean center sparse data (a, b, and c)

as that destroys the sparse structure. In general, in other cases, it usually is good to mean center. For the medical data one probably wants to fully standardize, but we didn't ask that.

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

★ **SOLUTION:** C

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

★ **SOLUTION:** A

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

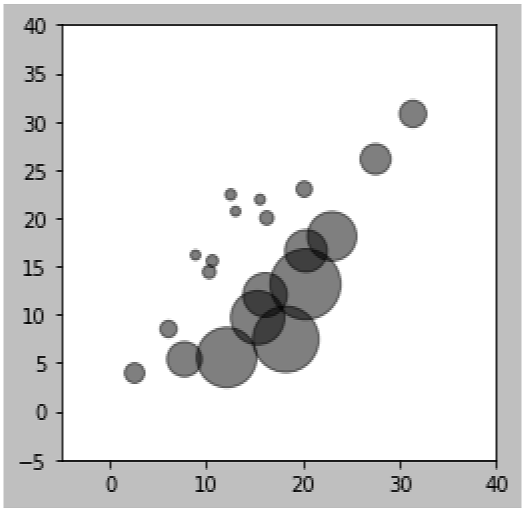
(c) 15*22

(d) 16*23

(e) None of the above

★ **SOLUTION:** B

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

★ **SOLUTION:** D

$$\text{V} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

(1)

new point $P' =< 2, 3 >$, after mean centering we get $P =< 2, 0 >$
$P^T \cdot V =< \sqrt{2}, -\sqrt{2} >$

UNIVERSITY OF PENNSYLVANIA
CIS 520: Machine Learning
Midterm 2019

Exam policy: This exam allows one one-page, two-sided cheat sheet; No other materials.

Time: 80 minutes. Be sure to write your name and Penn student ID (the 8 bigger digits on your ID card) on the answer form and fill in the associated bubbles *in pencil*.

If you think a question is ambiguous, mark what you think is the best answer. As always, we will consider written regrade requests if your interpretation of a question differed from what we intended. *We will only grade the answer 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.

Questions follow the course convention that n or N represents the number of observations and p the number of features in the model.

There are **40** questions and **61** points (19 one point questions, 21 two point questions).

1. [2 points] Compared to MLE, MAP solutions tend to have

   (a) lower bias and lower variance
   (b) higher bias and lower variance
   (c) lower bias and higher variance
   (d) higher bias and higher variance

   ★ **SOLUTION:** B

2. [2 points] We trained a three-way logistic regression and obtained weights

   $$w_a = (1, 1, 0), w_b = (-1, 1, 1), w_c = (2, 1, 2)$$

   What label would be given to the point $x = (0, 1, 1)$?

   (a) A
   (b) B
   (c) C

   ★ **SOLUTION:** C

3. [2 points] As part of building a decision tree, we want to measure the information gain from asking a question about a binary split on feature X_1 of the following 4 samples:

X_1	y
T	0
F	1
F	1
F	0

   Which of the following measures the information gain of splitting the sample on the value of X_1?

   (a) $-(\frac{2}{3})\log_2(\frac{2}{3}) + (\frac{1}{2})\log_2(\frac{1}{2})$
   (b) $-\log_2(\frac{1}{2}) + \frac{1}{4}(\frac{2}{3}\log_2(\frac{2}{3}) + \frac{1}{3}\log_2(\frac{1}{3})) + \frac{3}{4}(\frac{2}{3} + \log_2(\frac{2}{3}) + \frac{1}{3}\log_2(\frac{1}{3}))$
   (c) $-\log_2(\frac{1}{2}) + \frac{3}{4}(\frac{2}{3}\log_2(\frac{2}{3}) + \frac{1}{3}\log_2(\frac{1}{3}))$
   (d) $-\log_2(\frac{1}{2}) + \frac{1}{4}(\frac{2}{3}\log_2(\frac{2}{3}) + \frac{1}{3}\log_2(\frac{1}{3}))$
   (e) none of the above

★ **SOLUTION:** C

4. [2 points] The depth of a decision tree is most likely to scale as

   (a) the number of training examples

   (b) the square root of the number of training examples

   (c) the log of the number of the training examples

★ **SOLUTION:** C

5. [2 points] You are using a dataset for housing with one feature (area) in squared meters. Your friend Margaret is also using the same data set but with the mentioned feature in squared feet. (A meter is roughly 3 feet.) You both are using nearest neighbors with L2 distance. What do you expect to see?

   (a) The two models will both give the exact same predictions.

   (b) Area is **more** important when measured in feet than in meters.

   (c) Area is **less** important when measured in feet than in meters.

   (d) The two models will give different predictions, but we can't know the relative effect of area.

★ **SOLUTION:** B

6. [2 points] Similar to the question above, you are using a dataset for housing with one feature (area) in squared meters. Your friend Margaret is also using the same data set but with the mentioned feature in squared feet. However, this time, you are both are experimenting with decision trees. What do you expect to see?

   (a) The two models will both give the exact same predictions.

   (b) Area is **more** important when measured in feet than in meters.

   (c) Area is **less** important when measured in feet than in meters.

   (d) The two models will give different predictions, but we can't know the relative effect of area.

★ **SOLUTION:** A

7. [2 points] Traditionally, when we have a real-valued input attribute during decision-tree learning, we consider 9 binary splits according to whether the attribute is in the lowest 10%, lowest 20%, etc. Your friend Pat suggests that instead we should just have a 10-way split with one branch for each of the 10 "bins" of the attribute (lowest 10%, next 10%, etc). The single biggest problem with Pat's suggestion is:

   (a) It is too computationally expensive.

   (b) It would probably result in a decision tree that scores badly on the training set and a test set.

   (c) It would probably result in a decision tree that scores well on the training set but badly on a test set.

   (d) It would probably result in a decision tree that scores well on a test set but badly on a training set.

   ★ **SOLUTION:** C

8. [2 points] Suppose we compute the MAP estimate of the mean μ of a Gaussian with a fixed variance, n samples and the following prior on μ:

$$p(\mu|\mu_0, \sigma_0) = \frac{1}{\sigma_0\sqrt{2\pi}} e^{\frac{-(\mu-\mu_0)^2}{2\sigma_0^2}}$$

   Under what condition(s) does the MAP estimate of μ converge to the MLE estimate of μ?

   (a) $\mu_0 \to 0$

   (b) $n \to \infty$

   (c) $\sigma_0 \to \infty$

   (d) two of the above

   (e) all of the above

   ★ **SOLUTION:** D

9. [2 points] As k in k-NN increases:

   (a) the bias increases and the variance increases

   (b) the bias increases and the variance decreases

   (c) the bias decreases and the variance increases

   (d) the bias decreases and the variance decreases

   ★ **SOLUTION:** B

10. [1 points] *True or False?* The larger the regularization penalty λ in penalized regression, the more the model will tend to overfit the training data.

    ★ **SOLUTION:** False

11. [1 points] Which norm most heavily shrinks large weights (weights that are much bigger than 1)?

    (a) L_0

    (b) L_1

    (c) L_2

    (d) not enough information to tell

    ★ **SOLUTION:** C

12. [2 points] Suppose you have a friend who is an esteemed doctor. Your friend has a dataset containing gene expression data for patients. For each patient, there is a vector containing the expression levels of $10,000$ genes, along with a label indicating whether the patient developed a certain disease. Using this, your friend's goal is to learn a model where, given a new patient (also represented by a similar $10,000$-dimensional vector), the model can output the probability of this patient developing the disease. Your friend also believes a reasonable model will depend on at least $9,000$ of the genes. Which of the following methods would be most suitable for this problem?

(a) Linear least squares regression with L_2 penalty

(b) Linear least squares regression with L_1 penalty

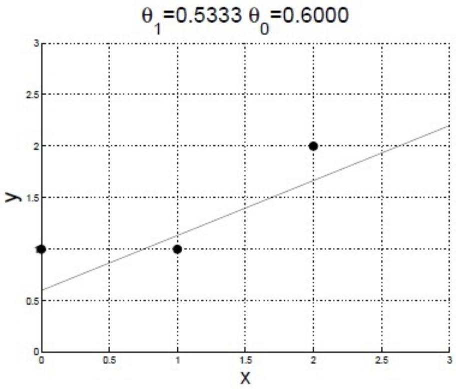
(c) Logistic regression with L_2 penalty

(d) Logistic regression with L_1 penalty

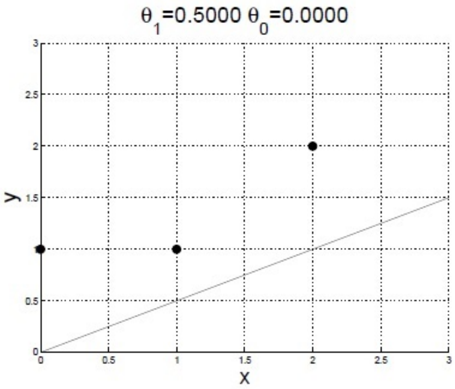
(e) Decision Tree

★ **SOLUTION:** C

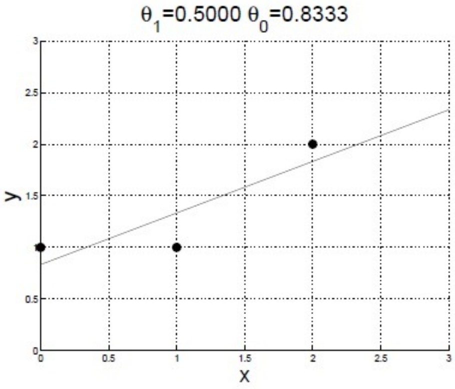
Models A-D below are fit by linear regression on a simple three-example data set $\{(1,1),(0,1),(2,2)\}$ with one independent variable x and one dependent variable y, using OLS or L_1 or L_2-penalized regression.



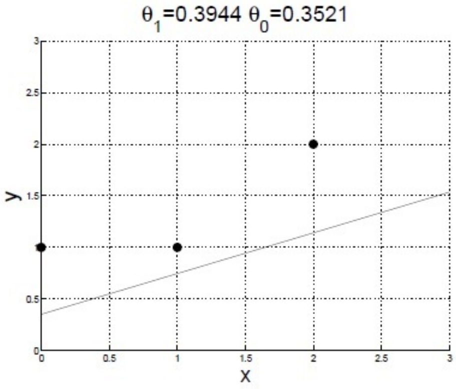
A)



B)



C)



D)

Use the plots above to answer **the next three questions**:

13. [2 points] Which model was most likely to have been generated using OLS?

   (a) Model A
   (b) Model B
   (c) Model C
   (d) Model D
   (e) Not enough information

★ **SOLUTION:** (C)

14. [2 points] Which model was most likely to have been generated using L_1 regularization?

    (a) Model A

    (b) Model B

    (c) Model C

    (d) Model D

    (e) Not enough information

★ **SOLUTION:** (B)

15. [2 points] Which model was most likely to have been generated using L_2 regularization?

    (a) Model A

    (b) Model B

    (c) Model C

    (d) Model D

    (e) Not enough information

★ **SOLUTION:** (D)

16. [2 points] In order to check whether a function $k(x_1, x_2)$ is a kernel function, we used a matrix X to generate a potential kernel matrix K. If the matrix K has eigenvalues 1.0, 2.3, 3.7, and 4.2 then we can conclude that

    (a) $k(x_1, x_2)$ is definitely a kernel function

    (b) $k(x_1, x_2)$ cannot be a kernel function

    (c) $k(x_1, x_2)$ might be a kernel function

★ **SOLUTION:** C

17. [2 points] In order to check whether a function $k(x_1, x_2)$ is a kernel function, we used a matrix X to generate a potential kernel matrix K. If the matrix K has eigenvalues -1.0, 2.3, 3.7, and 4.2 then we can conclude that

    (a) $k(x_1, x_2)$ is definitely a kernel function

    (b) $k(x_1, x_2)$ cannot be a kernel function

    (c) $k(x_1, x_2)$ might be a kernel function

★ **SOLUTION:** B

18. [1 points] *True or False?* Stepwise linear regression will always find at least as accurate a model as streamwise regression, assuming the same regularization penalty is used in both cases.

★ **SOLUTION:** False

19. [1 points] *True or False?* The result of streamwise regression depends on the order in which features are added.

★ **SOLUTION:** True

20. [1 points] *True or False?* In stepwise regression, the model which includes all the features will always be considered.

★ **SOLUTION:** False

21. [2 points] You want to fit a linear regression model with "the best" 10 features out of 200 candidate features. Which of the following will work best?

    (a) Perform L_2 regularization

    (b) Perform L_1 regularization

    (c) Use the hinge loss instead of square error loss.

★ **SOLUTION:** B

22. [1 points] *True or False?* For logistic regression, gradient descent can converge to a local minimum and fail to find a global minimum. More advanced optimization methods are thus often used.

★ **SOLUTION:** False

23. [1 points] *True or False?* A larger stepsize in gradient descent can lead to faster convergence, but lower accuracy.

★ **SOLUTION:** True

24. [1 points] *True or False?* The key idea of AdaGrad is to learn slowly from frequent features but pay attention to rare but informative features.

★ **SOLUTION:** True

25. [1 points] *True or False?* The decision boundary for logistic regression in p dimensions is a $p-1$-dimensional hyperplane.

★ **SOLUTION:** True

26. [2 points] The radial basis functions used in RBFs transform a feature vector of dimension p to a new space of dimension k such that

    (a) $k < p$

    (b) $k = p$

    (c) $k > p$

    (d) there is no constraint on how k and p are related

    ★ **SOLUTION:** D

27. [1 points] *True or False?* SVMs, unlike neural nets, have a global optimum.

    ★ **SOLUTION:** True

28. [1 points] *True or False?* The error surface followed by the gradient descent backpropagation algorithm changes if we change the training data.

    ★ **SOLUTION:** True

29. [2 points] Which of the following statement about KL-divergence is **not** true?

    (a) KL-divergence is a measure of how different two distributions are

    (b) KL-divergence is 0 if and only if two distributions P and Q are equal

    (c) KL-divergence can sensibly be measured between any two vectors of equal length

    (d) KL-divergence is always non-negative.

    ★ **SOLUTION:** (C)

30. [1 points] *True or False?* A max pooling layer in a ConvNet has an equal number of weights to learn as a filter of the same size.

★ **SOLUTION:** False

31. [1 points] *True or False?* (Standard) GANS are neural networks composed of two subcomponent neural nets. One neural net (the generator) takes as input an image and generates a different, fake image. The other neural net (the discriminator) takes as input either a real image or a fake image and tries to determine whether it was real or fake.

★ **SOLUTION:** False

32. [2 points] Assume a convolutional neural net where the input is a 4×4 RGB image (i.e., 4×4×3), with 2 filters, each of size 2×2×3, a stride of 2, and we zero pad the image with zeros on all four sides (giving a 6×6×3 "input").

    How many nodes are there in the first hidden layer?

    (a) 6×6×3×2
    (b) 5×5×2
    (c) 4×4×2
    (d) 3×3×2
    (e) none of the above

    ★ **SOLUTION:** D

33. [1 points] *True or False?* When fitting a model using Adaboost, training should stop before the training error reaches zero.

    ★ **SOLUTION:** False

34. [2 points] Suppose we are fitting a model using gradient boosting, where we call η the learning rate of the model update:

$$F_m(x) = F_{m-1}(x) + \eta \cdot \gamma_{m-1} h_{m-1}(x),\ 0 < \eta \leq 1$$

    Additionally, each base learner $h_{m-1}(x)$ is fit with a randomly subsampled fraction f, $0 < f \leq 1$ of the training data.

    Which of the following would most reduce overfitting?

    (a) increasing η, increasing f
    (b) decreasing η, increasing f
    (c) increasing η, decreasing f
    (d) decreasing η, decreasing f

★ **SOLUTION:** D

35. [1 points] *True or False?* When fitting a model using gradient boosting and a squared loss, we fit the base learner to the (standard) residuals at each iteration.

★ **SOLUTION:** True

36. [1 points] *True or False?* In AdaBoost, in each iteration the weights of the misclassified examples on any given iteration all go up by the same multiplicative factor.

★ **SOLUTION:** True

37. [1 points] *True or False?* AdaBoost will eventually give zero training error regardless of the type of weak classifier it uses, provided enough iterations are performed.

★ **SOLUTION:** False

38. [2 points] Which of the following **least** penalizes misclassifications as a function of how far they are from the decision boundary?

    (a) hinge loss
    (b) logistic loss
    (c) exponential loss

★ **SOLUTION:** A

39. [1 points] *True or False?* Hinge loss functions only penalize points for which the predicted classification label is wrong.

★ **SOLUTION:** False

40. [1 points] *True or False?* The number of support vectors found by an SVM is independent of the magnitude of the penalty on the slack variables.

★ **SOLUTION:** False

UNIVERSITY OF PENNSYLVANIA

CIS 520: Machine Learning
Final Exam, 2019

Exam policy: This exam allows one one-page, two-sided cheat sheet; No other materials.

Time: 120 minutes. Be sure to write your name and Penn student ID (the 8 bigger digits on your ID card) on the answer form and fill in the associated bubbles *in pencil.*

If you think a question is ambiguous, mark what you think is the best answer. As always, we will consider written regrade requests if your interpretation of a question differed from what we intended. *We will only grade the answer 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.*

Questions follow the course convention that n or N represents the number of observations and p the number of features in the model.

There are **68** questions with a total of **90** points.

1. [2 points] For very large training data sets, which of the following will usually have the lowest training time?
   (a) logistic regression
   (b) neural nets
   (c) K-nearest neighbors
   (d) random forests
   (e) a linear SVM

   ★ **SOLUTION:** C

2. [2 points] For very large training data sets, which of the following will usually produce the smallest models (requiring the fewest parameters)?
   (a) logistic regression
   (b) neural nets
   (c) K-nearest neighbors
   (d) random forests
   (e) a linear SVM

   ★ **SOLUTION:** * A, C or E Logistic regression or SVM produces

   the smallest models, but k-NN has the fewest parameters (none).

3. [1 points] *True or False?* We train a decision tree model on a dataset that is not feature scaled. We standardize features on the test set (set all to mean zero, standard deviation 1). We expect the model to accurately predict on the feature-scaled test set because decision trees are scale-invariant.

   ★ **SOLUTION:** False

4. [2 points] A K-NN classifier will give accuracies most similar to
   (a) logistic regression
   (b) decision trees

(c) k-means (with each cluster labeled with the majority class of the items in it)

(d) neural networks

(e) Naive Bayes

★ **SOLUTION:** C

5. [1 points] *True or False?* The KL-divergence $D_{KL}(p||q)$ measures how different an approximate distribution p is from a "true" (or at least more accurate) distribution q.

★ **SOLUTION:** False

6. [1 points] *True or False?* The distribution $P(A) = 1/2, P(B) = 1/2$ has higher entropy than the distribution $P(A) = 1/3, P(B) = 1/3, P(C) = 1/3$

★ **SOLUTION:** False

7. [1 points] *True or False?* L2 loss is more robust to outliers than L1 loss.

★ **SOLUTION:** False

8. [2 points] Suppose you want to analyze a data set containing gene expression data for patients. For each patient, there is a feature vector containing expression levels of 20,000 genes, together with a label indicating whether the patient developed a certain disease. Your goal is to learn a model which, given a new patient (also represented by a similar 20,000-dimensional vector), can estimate the probability of this patient developing the disease. You expect the eventual model will depend on only a small number of the genes. Which of the following methods would be most suitable?

(a) Logistic regression with L2 regularization

(b) Linear regression with L2 regularization

(c) Logistic regression with L0 regularization

(d) Linear regression with L0 regularization

(e) Linear support vector machine

★ **SOLUTION:** C

9. [1 points] *True or False?* Naive Bayes is generally preferable to logistic regression if there are very little data (compared to the number of parameters), while logistic regression gives more accurate models in the limit of large training sets.

★ **SOLUTION:** True

10. [1 points] *True or False?* Generative Adversarial Networks (GANs) often converge to poor solutions (local optima) because the Generator tends to converge much more quickly than the Discriminator.

★ **SOLUTION:** False

For the next 3 questions, assume you have classification data with classes Y being +1 or -1 and features x_j also being +1 or -1 for $j \in 1, ..., p$.

In an attempt to turbocharge your classifier, you duplicate each feature, so now each example has $2p$ features, with $x_{p+j} = x_j$ for $j \in 1, ..., p$. The following questions compare the original feature set with the doubled one. You may assume that in the case of ties, class +1 is always chosen. Assume that there are equal numbers of training examples in each class.

11. [1 points] For a Naive Bayes classifier:

(a) The test accuracy will usually be higher with the original features.

(b) The test accuracy will usually be higher with the doubled features.

(c) The test accuracy will be the same with either feature set.

★ **SOLUTION:** * C This is tricky, but since we assumed equal class

probabilities, doubling the features squares the P(x$_j|class)terms, sodoesn'tchangetheclasspredict$

12. [1 points] For a Naive Bayes classifier:

(a) On a given training instance, the conditional probability $P(Y|x_1, ...)$ on a training instance will be more extreme (i.e. closer to 0 or 1) with the original features.

(b) On a given training instance, the conditional probability $P(Y|x_1, ...)$ on a training instance will be more extreme (i.e. closer to 0 or 1) with the doubled features.

(c) On a given training instance, the conditional probability $P(Y|x_1, ...)$ on a training instance will be the same with either feature set.

★ **SOLUTION:** B

13. [2 points] For a perceptron classifier:

(a) The test accuracy will, in general, be higher with the original features.

(b) The test accuracy will, in general, be higher with the doubled features.

(c) The test accuracy will always be the same with either feature set.

★ **SOLUTION:** C

14. [2 points] In neural networks, what is the benefit of the ReLU activation function over a Sigmoid activation function?

    (a) ReLUs allow the model to learn non-linear decision boundaries
    (b) ReLUs allow for faster backpropogation gradient calculations
    (c) The ReLUs activation function can be used in output layers while a sigmoidal activation function cannot.
    (d) All of the above

★ **SOLUTION:** B

15. [2 points] Consider a convolutional net where the p-dimensional input data is laid out in a one-dimensional fashion (e.g. for text or speech), and where there are k filters (kernels), each of size $m \times 1$. Assume no padding, and a stride of size s. Which of the following **best** approximates the number of outputs of this layer?

    (a) ksp/m
    (b) $kspm$
    (c) ksm/p
    (d) spm
    (e) kmp/s

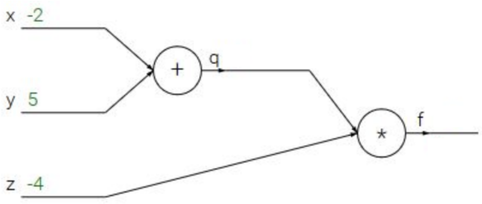
★ **SOLUTION:** E

16. [2 points] Suppose you have inputs as $x = -2$, $y = 5$, and $z = -4$. You have a neuron q and neuron f with functions:

$$q = x + y$$
$$f = q * z$$

What is the gradient of f with respect to x, y, and z? See the figure below.

    (a) $(-3, 4, 4)$
    (b) $(4, 4, 3)$
    (c) $(-4, -4, 3)$
    (d) $(3, -4, -4)$



★ **SOLUTION:** C

17. [2 points] In which of the following neural net architectures do some of the weights get reused more than once in each single forward pass?

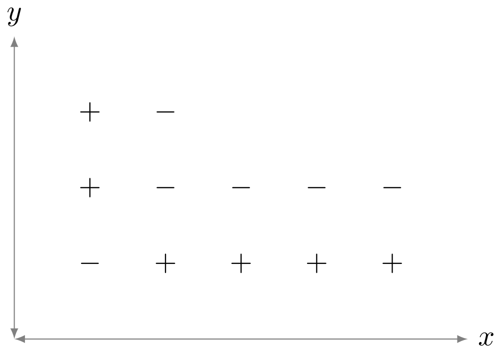
    (a) convolutional neural network
    (b) recurrent neural network
    (c) fully connected neural network
    (d) autoencoder
    (e) Both A and B

★ **SOLUTION:** E

18. [2 points] In AdaBoost, we choose α_t as the weight of the t-th weak learner, where

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \epsilon_t}{\epsilon_t} \right)$$

$\varepsilon_t = P_{x \sim D_t} [h_t(x) \neq y]$ is the weighted fraction of examples misclassified by the t-th weak learner. Here our weak learners are depth 1 decision trees that yield vertical or horizontal half-plane decision boundaries. If we conduct two iterations of boosting on the following dataset, Which is larger, α_1 or α_2?



(a) $\alpha_1 > \alpha_2$

(b) $\alpha_2 > \alpha_1$

(c) $\alpha_1 = \alpha_2$

(d) Not enough information

★ **SOLUTION:** A

19. [2 points] Suppose you have a classification problem where you want to penalize misclassifications more the farther they are from the decision boundary. How many of the following loss functions would be appropriate?

- $0-1$ loss
- hinge loss
- logistic loss
- exponential loss

(a) none

(b) 1

(c) 2

(d) 3

(e) all 4

★ **SOLUTION:** D All but 0/1

20. [2 points] Suppose you encountered a classification problem where you **do not** want to additionally reward highly confident correct classifications. What choice(s) of loss would be appropriate?

    (a) $0-1$ loss only
    (b) hinge loss only
    (c) exponential loss only
    (d) $0-1$ or hinge loss
    (e) hinge or exponential loss

★ **SOLUTION:** D

21. [2 points] Which of the following losses are **not** convex?

    (a) $0-1$ loss
    (b) hinge loss
    (c) exponential loss
    (d) (a) and (b)
    (e) (a), (b) and (c)

★ **SOLUTION:** A

22. [1 points] *True or False?* Removal of a support vector will always change the SVM decision boundary.

★ **SOLUTION:** False

23. [1 points] *True or False?* Radial Basis Functions (RBFs) use a Gaussian kernel to transform a p-dimensional feature space (x) to a (k-dimensional) transformed feature space where $k \leq p$.

★ **SOLUTION:** False

24. [1 points] *True or False?* Principle Component Regression (PCR) uses the right singular vectors of the feature matrix X to transform a p-dimensional feature space (x) to a (k-dimensional) transformed feature space where $k \leq p$.

★ **SOLUTION:** True

25. [1 points] *True or False?* A perceptron is guaranteed to learn a perfect decision boundary within a finite number of iterations for linearly separable data.

★ **SOLUTION:** True

26. [1 points] *True or False?* Least Mean Squares (LMS) is an online approximation to linear regression and perceptrons are online approximations to SVMs.

★ **SOLUTION:** True

27. [2 points] After performing SVD on a dataset with 5 features, you retrieve eigenvalues 6, 5, 4, 3, 2. How many components should we include to explain at least 75% of the variance of the dataset?

    (a) 1
    (b) 2
    (c) 3
    (d) 4

★ **SOLUTION:** C

28. [1 points] *True or False?* After performing SVD on a dataset, you notice the eigenvalues returned are all approximately equal. You expect variance explained to be approximately linear to the number of components used for PCA.

★ **SOLUTION:** True

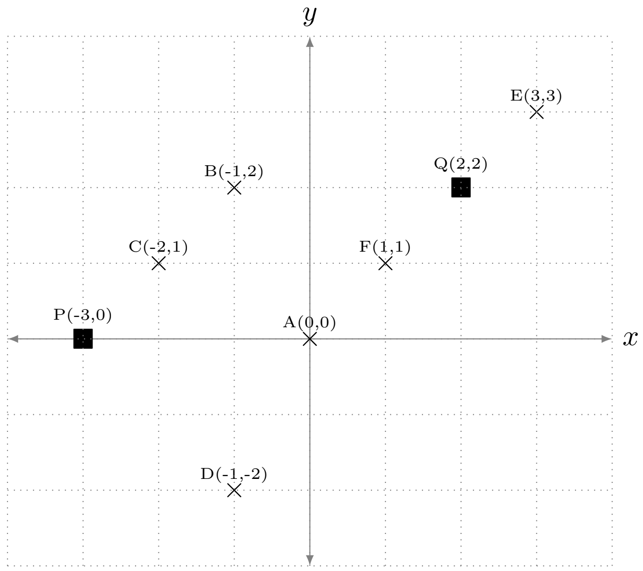
29. [1 points] *True or False?* Given a dataset with n features, if we know $(n-1)$ principal components of the dataset, then we can determine the missing principal component.

★ **SOLUTION:** True

30. [1 points] *True or False?* Changing one feature of a dataset with multiple features from centimeters to inches will not affect the outcome of PCA.

★ **SOLUTION:** False

For the next 4 questions, refer the the graph presented.



31. [2 points] In the first step of K-means with the standard Euclidean distance metric, which points will be assigned to the cluster centered at P?

    (a) C, D
    (b) A, C, D
    (c) B, C, D
    (d) A, B, C
    (e) A, B, C, D

★ **SOLUTION:** C

32. [2 points] Continue running K-means with the standard Euclidean distance metric. What does the cluster center P get updated to? (Do not include P as a point).

    (a) $(-\frac{3}{2}, \frac{1}{2})$
    (b) $(-\frac{4}{3}, \frac{1}{3})$
    (c) $(-\frac{4}{3}, 0)$
    (d) $(-1, \frac{1}{3})$
    (e) $(-\frac{3}{2}, \frac{1}{3})$

★ **SOLUTION:** B

While K-means used Euclidean distance in class, we can extend it to other distance functions, where the assignment and update phases still iteratively minimize the total (non-Euclidian) distance. Here, consider the Manhattan distance:

$$d'((A_1, A_2), (B_1, B_2)) = |A_1 - B_1| + |A_2 - B_2|$$

Again start from the original locations for P and Q as shown in the figure, and perform the update assignment step and the update cluster center step using Manhattan distance as the distance function:

33. [2 points] Starting from the same initial configuration, select all points that get assigned to the cluster with center at P, under this new distance function $d'(A, B)$.

    (a) C, D

    (b) A, C, D

    (c) B, C, D

    (d) A, B, C

    (e) A, B, C, D

★ **SOLUTION:** B

34. [2 points] What does cluster center P now get updated to, under this new distance function $d'(A, B)$? (Do not include P as a point).

    (a) $(-\frac{3}{2}, \frac{1}{2})$

    (b) $(-\frac{4}{3}, \frac{1}{3})$

    (c) $(-\frac{4}{3}, 0)$

    (d) $(-1, -\frac{1}{3})$

    (e) $(-\frac{3}{2}, \frac{1}{3})$

★ **SOLUTION:** * The exam as given did not have the correct solution option; it is now D

35. [1 points] *True or False?* The F1 score is generally a better performance measure than accuracy when there is extreme class imbalance in the labels.

★ **SOLUTION:** True

36. [1 points] *True or False?* An AUC (Area under the ROC curve) of 0.4 on test data suggests overfitting.

★ **SOLUTION:** True

37. [1 points] *True or False?* LIME (Local Interpretable Model-Agnostic Explanations) fits a linear model to observations close a point of interest and determines which features in that linear model are most influential in making the prediction.

★ **SOLUTION:** * The use of ”observations” here is ambiguous; it fits points that are created as perturbations of the original point.

38. [1 points] *True or False?* When doing boosting to compute ensembles of trees, using complex (high depth) trees generally helps to improve test set accuracy.

★ **SOLUTION:** False

39. [1 points] *True or False?* When running linear regressions, it is a good idea to look at the largest (in absolute value) regression weights to see which features are most influential in determining the predictions.

★ **SOLUTION:** False

40. [2 points] For Naive Bayes, what happens to our document posteriors as we increase our pseudo-count parameter?

    (a) They approach 0

    (b) They approach 1

    (c) They approach the document priors

    (d) None of the above

★ **SOLUTION:** C

41. [2 points] We are trying to use Naive Bayes to classify a Facebook meme as either funny or sad. Suppose out of 100 training memes, we see that 60 of these memes are funny while 40 are sad. Also, assume that our dictionary consists of only three words, and the counts of the words for funny and sad memes are listed below.

Word	Funny Count	Sad Count
Yum	40	5
Friends	40	15
Cry	20	30

If we see a meme post with one occurrence of the word friends and one occurrence of the word cry (with no other words), which class has higher posterior probability?

    (a) Funny

    (b) Sad

★ **SOLUTION:** A There are 100 memes, 60 funny and 40 sad. (as stated

in the question text)

Of the 60 funny memes, 40 contain ”yum”. 40 contain ”friends” and 20 contain ”cry”. (The memes have more than one word in them, so a meme like ”froyo with my friends, yum!” has both ”friends” and ”yum” in it.)

When we compute p(friends—funny), that is short for the probability a meme with class label ”funny” contains the word ”friends”, which is 40/60, not 40/(40+40+20).

42. [1 points] *True or False?* Because LDA has "hidden variables" representing the mixture of topics within each document and the topic that each word in each document come from, it is often solved using the EM algorithm.

★ **SOLUTION:** True

43. [1 points] *True or False?* EM algorithms are attractive because for problems such as estimating Gaussian Mixture Models, they are guaranteed to find a global optimum in likelihood.

★ **SOLUTION:** False

44. [1 points] *True or False?* Power methods for estimating eigenvectors are attractive because they are guaranteed to find a global optimum in reconstruction error when used in PCA.

★ **SOLUTION:** True

45. [2 points] Which of the following statement about Hidden Markov Model (HMM) is **not** true?

    (a) An HMM is a deterministic model because it explicitly describes the conditional distribution of the output given the current state.

    (b) The Markov process assumes that future is independent of the past given the present.

    (c) HMMs allow us to compute the joint probability of a set of hidden states given a set of observed states.

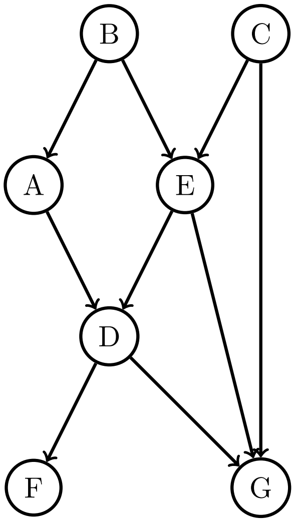
    (d) In HMM, given initial hidden states, constant transition matrix, emission matrix, and no other information, the observation will converge to only one state after a long sequence of prediction.

★ **SOLUTION:** * A or D

46. [1 points] *True or False?* The EM algorithm does a kind of "gradient descent" in likelihood, since both steps are guaranteed to decrease the negative log-likelihood.

★ **SOLUTION:** True

The following 5 questions are related to this graph:



47. [1 points] *True or False?* The joint probability of this graph can be represented as:

$P(A \mid B)P(B)P(C)P(D \mid A, E)P(E \mid B, C)P(F \mid D)P(G \mid C, E, D)$

★ **SOLUTION:** True

48. [1 points] *True or False?* The class of joint probability distributions that can be represented by the resulting Bayesian network:

$P(A \mid B)P(B)P(C)P(D \mid A, E)P(E \mid B, C)P(F \mid A, B, C, D, E)P(G \mid A, B, C, E, D)$

is smaller than the original network shown above.

★ **SOLUTION:** False

49. [1 points] *True or False?*

$$B \perp C \mid F$$

★ **SOLUTION:** False

50. [1 points] *True or False?*

$$A \perp C \mid G$$

★ **SOLUTION:** False

51. [1 points] *True or False?*

$$A \perp G \mid C, E$$

★ **SOLUTION:** False

52. [1 points] *True or False?* Although speech-to-text and text-to-speech is usually modeled using LSTMs or other RNNs, one could also use CNNs with one-dimensional filters.

★ **SOLUTION:** True

53. [1 points] *True or False?* CNNs work well on high dimensional problems like medical diagnosis from health records (which contain varied features like age, weight, temperature, lab results, disease history, etc.)

★ **SOLUTION:** False

54. [1 points] *True or False?* Vanilla RNNs, unlike HMMs, do not forget things exponentially quickly.

★ **SOLUTION:** False

55. [1 points] *True or False?* Q-learning is guaranteed to converge (for discrete states and actions) so long as all (state, action) pairs are visited infinitely often.

★ **SOLUTION:** * It is, but only given specific constraints on the learning

rate.

56. [1 points] *True or False?* In Q-learning, $Q(s, a)$ represents the expected discounted reward of taking action a in state s and subsequently following an optimal policy.

★ **SOLUTION:** False

57. [1 points] *True or False?* Epsilon-greedy Reinforcement Learning methods "exploit" by using an optimal policy a (small) fraction, given by ϵ, and "explore" a large fraction $(1 - \epsilon)$ of the time.

★ **SOLUTION:** False

58. [1 points] *True or False?* Current RL methods for game play, such as alphaZero, unlike earlier methods that trained a “new” Q-function by playing against an older one, now just play the “new” network against itself.

★ **SOLUTION:** True

59. [1 points] *True or False?* Value Iteration iteratively updates V using Bellman's equation, and is guaranteed to converge to the unique optimum represented by the solution to Bellman's equation (if all states are visited infinite numbers of times).

★ **SOLUTION:** True

60. [1 points] *True or False?* Autoencoders always take an input and pass it through an “encoder” which produces a lower dimensional representation which is then passed through a “decoder” to reconstruct the input as accurately as possible.

★ **SOLUTION:** False

61. [1 points] *True or False?* When picking which additional points, x, to label for linear regression, it is desirable to pick points that are "as spread out as possible" (i.e. as far away as possible from the existing points)

    ★ **SOLUTION:** True

62. [1 points] *True or False?* When picking which additional points, x, to label for SVMs, it is desirable to pick points that are "as spread out as possible" (i.e. far away from the existing points)

    ★ **SOLUTION:** False

63. [1 points] *True or False?* The most widely used experimental design methods pick new points to label such that they maximize a norm $||X^T X||_p$ for some p.

    ★ **SOLUTION:** * Too complex; everyone was given credit: They minimize the norm of the inverse of that matrix. This is often, but not always the same as maximizing that matrix.

64. [1 points] *True or False?* When doing active learning for SVMs, labeling the x's for which one is "most uncertain" will tend to select points that are closer to the separating hyperplane.

    ★ **SOLUTION:** True

65. [1 points] *True or False?* The "Query by Committee" active learning method makes more sense to use with linear regression than with random forests.

★ **SOLUTION:** False

66. [1 points] If a standard CNN model has been trained to distinguish images of men from women in a setting where the training data has 75% women, then predictions on a test set of images drawn from the same distribution is more likely to have

    (a) under 75% women

    (b) very close to 75% women

    (c) over 75% women

★ **SOLUTION:** C

67. [2 points] When training a machine learning model on a data set which is not representative of the population of interest (e.g. when using Twitter users to represent the general population) it is best to:

    (a) Use L_1 rather than L_2 loss

    (b) restratify

    (c) use imputation

    (d) none of the above

★ **SOLUTION:** B

68. [1 points] *True or False?* Least Mean Squares does stochastic gradient descent in a negative log-likelihood.

★ **SOLUTION:** True

UNIVERSITY OF PENNSYLVANIA
CIS 5200: Machine Learning
Midterm 2022

Exam policy: You are allowed one two-sided cheat sheet. No calculators.

Time: 90 minutes.

Please write your name and Penn ID on the bubble sheet.

If you think a question is ambiguous, mark what you think is the best answer. As always, we will consider written regrade requests if your interpretation of a question differed from what we intended. *We will only grade the answer 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.

Questions follow the course convention that n or N represents the number of observations and p the number of features in the model.

There are **46** questions totaling **61** points.

1. [1 points] *True or False?* Ordinary least squares (with L2 loss) is more commonly used than regression with L1 loss, because the L2 loss, unlike the L1 loss, is convex.

   ★ **SOLUTION:** False

2. [2 points] Which of the following is true of ordinary least squares regression?

   (a) It is prone to underfitting.
   (b) It always has a closed-form Solution.
   (c) The weights are usually calculated using gradient descent.
   (d) It has a closed-form solution, except when the columns of $\mathbf{X}$ are linearly dependent
   (e) none of the above

   ★ **SOLUTION:** E-also gave credit for D

3. [2 points] In a regression task, suppose one feature is in inches. If we convert that feature to centimeters, which model would in general have the same predictions as before?

   (a) OLS regression
   (b) Ridge regression
   (c) K-NN
   (d) two of the above
   (e) none of the above

   ★ **SOLUTION:** A

4. [2 points] If you had to choose whether to use the MLE or MAP estimate for a machine learning model, which is generally better to use and why?

   (a) MAP, because it generalizes MLE.
   (b) MLE, because it is unbiased.
   (c) MLE, because it avoids overfitting better than MAP.
   (d) MAP, because it is unbiased.

★ **SOLUTION:** A

5. [1 points] *True or False?* Using ridge regression will always result in all of the weights being smaller than using ordinary least squares regression on the same dataset.

★ **SOLUTION:** False - but gave credit for either

6. [1 points] *True or False?* When using the MAP estimate, collecting more data increases the effect of the prior because the prior can affect more data than before.

★ **SOLUTION:** False

7. [1 points] *True or False?* Suppose θ is the parameter we are interested in, D is the dataset that we have observed. Then θ_{MLE} maximizes $P(D|\theta)$, and θ_{MAP} maximizes $P(\theta|D)$.

★ **SOLUTION:** True

8. [1 points] *True or False?* If we double all features (i.e. multiply all x's by 2), Decision Trees will give the same prediction.

★ **SOLUTION:** True

9. [1 points] *True or False?* If we double all features (i.e. multiply all x's by 2), K-NN will give the same prediction.

★ **SOLUTION:** True

10. [1 points] *True or False?* If we double all features (i.e. multiply all x's by 2), Ridge regression will give the same prediction.

★ **SOLUTION:** False

11. [2 points] Which of the following regression methods perform feature selection?

    (a) L0

    (b) L1

    (c) L2

    (d) L-inf

    (e) Two of the above

★ **SOLUTION:** E

12. [1 points] *True or False?* We regularize a Bernoulli model for estimating the probability of getting a "heads" (vs. a "tails"), by 'pretending' we have already seen a certain number of heads and tails. The effect of this prior vanishes in the limit of infinite observations.

★ **SOLUTION:** True

13. [2 points] Which of the following is TRUE about the bias-variance decomposition of test error?

    (a) Expected test error is equal to $bias + var^2 + noise$

    (b) Noise can be reduced by using regularization

    (c) Low variance corresponds to high complexity

    (d) None of the above

★ **SOLUTION:** D

14. [1 points] In machine learning, we tend to favor estimators that are **pick the best–or least bad–answer**:

    (a) Unbiased, because we want predictions as close to their true values as possible

(b) Biased, because we don't want to treat the training data as definitive

(c) Unbiased, because at high training set size n we will converge to the optimal predictor

(d) Biased, because we want to push model parameters to as close to zero as we can

★ **SOLUTION:** B

15. [1 points] What best defines the relationship between a model's fit and its bias and variance

(a) A model with low bias and high variance is underfitting

(b) A model with high bias and high variance is underfitting

(c) A model with low bias and low variance is underfitting

(d) A model with high bias and low variance is underfitting

★ **SOLUTION:** D

16. [2 points] If Z is ""High Entropy"", this suggests that

(a) Z is from a relatively uniform distribution

(b) Z is from a highly varied distribution

(c) The value of Z is certain and known

(d) Z is conditional on other variables

★ **SOLUTION:** A

17. [1 points] *True or False?* A Gaussian distribution with $\sigma = 1$ has higher entropy than one with $\sigma = 3$.

★ **SOLUTION:** False

18. [1 points] *True or False?* Standard SVMs, as covered in class (linear, so no Gaussian kernel), have no hyperparameters that need to be set by cross-validation.

★ **SOLUTION:** False

19. [2 points] If we assume that the data is being generated from a source with no noise $y = w^T x$, choosing a classifier $f^*(x) = {w^*}^T x$ using OLS will make which of the following quantities zero?

    (a) Only Bias

    (b) Only Variance

    (c) Both Bias and Variance

    (d) None of the above

    ★ **SOLUTION:** C

20. [1 points] *True or False?* Increasing the kernel width in a RBF model will in general increase the training error.

    ★ **SOLUTION:** True

21. [1 points] *True or False?* Increasing the kernel width in a RBF model will in general increase the testing error.

    ★ **SOLUTION:** False

22. [1 points] *True or False?* Increasing k in a K-NN model will in general increase the error on the 'training set'.

    ★ **SOLUTION:** True

23. [1 points] *True or False?* If OLS gives all non-zero values for the weights, using Ridge Regression will also result in all non-zero weight values.

    ★ **SOLUTION:** True

24. [1 points] *True or False?* Leave-one-out cross-validation is often preferred over k-fold cross-validation for smaller datasets.

★ **SOLUTION:** True

25. [1 points] *True or False?* Link functions transform the feature space nonlinearly before fitting a linear model.

★ **SOLUTION:** False

26. [2 points] A discrete probability distribution has (x, probability(x)) tuples $\left(1, \frac{1}{4}\right), \left(2, \frac{1}{2}\right), \left(4, \frac{1}{4}\right)$. What is the expected value of x?

    (a) 1

    (b) 1.5

    (c) 2.25

    (d) 7

    (e) none of the above

★ **SOLUTION:** C

27. [2 points] A discrete probability distribution has (x, probability(x)) tuples $\left(1, \frac{1}{4}\right), \left(2, \frac{1}{2}\right), \left(4, \frac{1}{4}\right)$. What is the entropy of this distribution?

    (a) 1 bit

    (b) 1.5 bits

    (c) 2 bits

    (d) 3 bits

    (e) none of the above

★ **SOLUTION:** B

28. [1 points] *True or False?* The L2 penalty in ridge regression comes from the assumption of Gaussian noise in the linear regression model $y = w^T x + \epsilon$.

★ **SOLUTION:** False

29. [1 points] *True or False?* The MAP with a uniform prior (the same probability everywhere) will be different from the MLE because the prior's shrinkage is applied everywhere.

★ **SOLUTION:** False

30. [2 points] Suppose you are given training and testing datasets for a regression problem, and you train a ridge regression model on the training dataset. If you observe low training error but high testing error, what is likely to be the best thing to do to improve the testing error?

    (a) Increase the number of features used
    (b) Increase the regularization parameter λ
    (c) Increase the number of training iterations
    (d) None of the above

★ **SOLUTION:** B

31. [1 points] *True or False?* AdaGrad uses a faster learning rate for features that have been changed more in the past than for ones that have changed less.

★ **SOLUTION:** False

32. [2 points] You train an SVM on a dataset with data of x_i and labels $y_i = \pm 1$ This produces a vector of weights w and a bias b. For the first 6 training examples, the values of $w^T x_i + b$ are listed below. How many support vectors are there?

i	y_i	$w^T x_i + b$
1	-1	-1.5
2	-1	-1
3	-1	-0.5
4	$+1$	1
5	$+1$	3
6	-1	4

(a) 2
(b) 3
(c) 4
(d) 5
(e) none of the above

★ **SOLUTION:** C

33. [1 points] *True or False?* SVM solutions can be expressed purely in terms of the vectors on the "margin".

★ **SOLUTION:** False

34. [1 points] *True or False?* The hinge loss used in SVMs generally gives less weight than logistic regression to points that are **misclassified** with a high probability or score.

★ **SOLUTION:** True

35. [1 points] *True or False?* The hinge loss used in SVMs generally gives less weight than logistic regression to points that are **correctly classified** with a high probability or score.

★ **SOLUTION:** True

36. [2 points] Your random forest is overfitting. What should you do to the number of trees in the model?

(a) Increase
(b) Decrease
(c) the number of trees won't effect overfitting

★ **SOLUTION:** A

37. [1 points] *True or False?* $k(x, x') = exp(\frac{||x-x'||_2^2}{2\sigma^2})$ is a valid kernel function.

★ **SOLUTION:** False

38. [1 points] *True or False?* An ensemble will usually have lower variance than a single model that is part of the ensemble.

★ **SOLUTION:** True

39. [1 points] *True or False?* Boosting methods tend to give higher test accuracy when they use more accurate 'weak models'.

★ **SOLUTION:** False

40. [1 points] *True or False?* Random forests (as we used them in class) try to make their weak learning less accurate by (among other methods) bagging.

★ **SOLUTION:** True

41. [1 points] *True or False?* Random forests (as we used them in class) try to make their weak learning less accurate by (among other methods) boosting.

★ **SOLUTION:** False

42. [1 points] *True or False?* Random forests (as we used them in class) try to make their weak learning less accurate by (among other methods) randomly selecting subsets of features.

★ **SOLUTION:** True

43. [1 points] *True or False?* Random forests (as we used them in class) try to make their weak learning less accurate by (among other methods) adding noise to the labels.

★ **SOLUTION:** False

44. [2 points] Which of the following statements about KL-Divergence is true? Note that P refers to the true distribution and Q refers to the alternative distribution.

   (a) It measures how well (or more precisely, how badly) Q approximates P

   (b) It measures the difference between two distributions, and thus can be seen as a distance metric

   (c) It can be written as cross-entropy of P, Q minus the entropy of Q. i.e. $D_{KL}(P||Q) = H(P, Q) - H(Q)$

   (d) Two of the above

   (e) None of the above

★ **SOLUTION:** A

45. [1 points] *True or False?* Streamwise regression never tests more models than stepwise regression.

★ **SOLUTION:** True -but credit was give for either on the original version of the question

46. [2 points] Consider classification using k-NN. Suppose we try different norms and get different decision boundaries. Which norms best match figures A, B, and C, respectively?

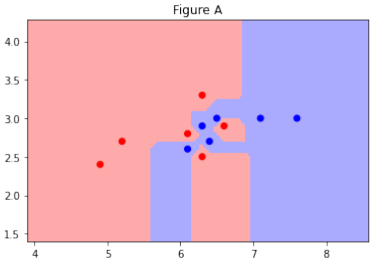
   (a) $L1, L2, L_{\text{inf}}$

   (b) $L2, L1, L_{\text{inf}}$

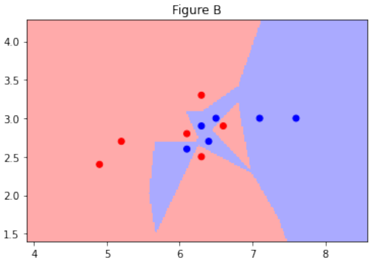
(c) $L1, L_{\text{inf}},\ L2$

(d) $L_{\text{inf}}, L2, L1$

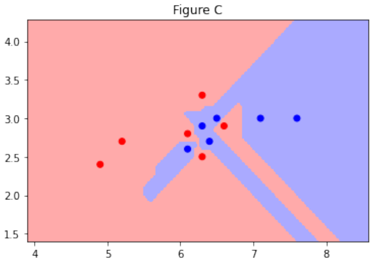
(e) none of the above



(a)



(b)



(c)

★ **SOLUTION:** A

Total Points: 60