0/27 Questions Answered

HW 9 (Electronic Component)

STUDENT NAME

Search students by name or email...

Q1 Maximum Likelihood Estimation

16 Points

We will begin with a short derivation. Consider a probability distribution with a domain that consists of $|X|$ different values. We get to observe N total samples from this distribution. We use n_i to represent the number of the N samples for which outcome i occurs. Our goal is to estimate the probabilities θ_i for each of the events $i = 1, 2 \ldots |X| - 1$. The probability of the last outcome, $|X|$, equals $1 - \sum_{i=1}^{|X|-1} \theta_i$.

In *maximum likelihood estimation* , we choose the θ_i that maximize the likelihood of the observed samples,

$$L(\text{samples}, \theta) \propto (1 - \theta_1 - \theta_2 - \ldots - \theta_{|X|-1})^{n_{|X|}} \prod_{i=1}^{|X|-1} \theta_i^{n_i}$$

For this derivation, it is easiest to work with the log of the likelihood. Maximizing log-likelihood also maximizes likelihood, since the quantities are related by a monotonic transformation. Taking logs we obtain

$$\theta^{\text{ML}} = \underset{\theta}{\text{argmax}} n_{|X|} \log(1 - \theta_1 - \theta_2 - \ldots - \theta_{|X|-1}) + \sum_{i=1}^{|X|-1} n_i \log \theta_i$$

Setting derivatives with respect to θ_i equal to zero, we obtain $|X| - 1$ equations in the $|X| - 1$ unknowns, $\theta_1, \theta_2, \ldots, \theta_{|X|-1}$:

$$\frac{-n_{|X|}}{1-\theta_1^{\mathrm{ML}}-\theta_2^{\mathrm{ML}}-\ldots-\theta_{|X|-1}^{\mathrm{ML}}}+\frac{n_i}{\theta_i^{\mathrm{ML}}}=0$$

Multiplying by $\theta_i\left(1-\theta_1-\theta_2-\ldots-\theta_{|X|-1}\right)$ makes the original $|X|-1$ nonlinear equations into $|X|-1$ linear equations:

$$-n_{|X|}\theta_i^{\mathrm{ML}}+n_i\left(1-\theta_1^{\mathrm{ML}}-\theta_2^{\mathrm{ML}}-\ldots-\theta_{|X|-1}^{\mathrm{ML}}\right)=0$$

That is, the maximum likelihood estimate of θ can be found by solving a linear system of
$|X|-1$ equations in $|X|-1$ unknowns. Doing so shows that the maximum likelihood estimate corresponds to simply the count for each outcome divided by the total number of samples. I.e., we have that:

$$\theta_i^{\mathrm{ML}}=\frac{n_i}{N}$$

Q1.1

2 Points

Now, consider a sampling process with 3 possible outcomes: R, G, and B. We observe the following sample counts:

outcome	R	G	B
count	0	3	10

What is the total sample count N?

Enter your answer here

Save Answer

Q1.2

6 Points

Following the setting in part 1, what are the maximum likelihood estimates for the probabilities of each outcome?

$\theta_R^{\mathrm{ML}} =$

Enter your answer here

$\theta_G^{\mathrm{ML}} =$

Enter your answer here

$\theta_B^{\mathrm{ML}} =$

Enter your answer here

Save Answer

Q1.3

4 Points

Now, use Laplace smoothing with strength $k = 4$ to estimate the probabilities of each outcome.

$\theta_R^{\mathrm{LAP},4} =$

Enter your answer here

$\theta_G^{\mathrm{LAP},4} =$

Enter your answer here

$\theta_B^{\mathrm{LAP},4} =$

Enter your answer here

Save Answer

Q1.4
4 Points

Now, consider Laplace smoothing in the limit $k \to \infty$. Fill in the corresponding probability estimates.

$\theta_R^{\mathrm{LAP},\infty} =$

Enter your answer here

$\theta_G^{\mathrm{LAP},\infty} =$

Enter your answer here

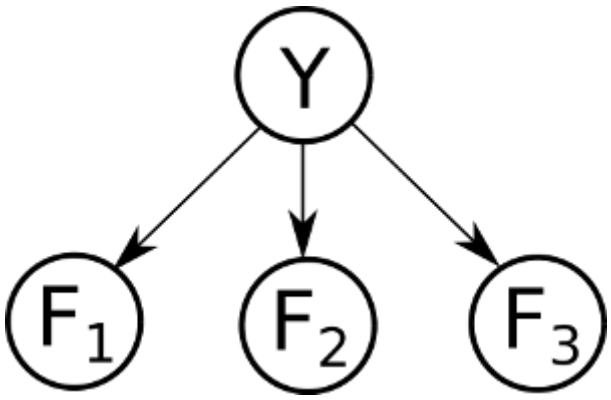
$\theta_B^{\mathrm{LAP},\infty} =$

Enter your answer here

Save Answer

Q2 Naïve Bayes
36 Points

In this question, we will train a Naive Bayes classifier to predict class labels Y as a function of input features F_i.



We are given the following 15 training points:

F_1	0	0	1	1	0	0	1	0	0	0	0	1	0	1	0
F_2	0	1	0	1	0	0	1	0	1	1	0	0	0	0	0
F_3	1	1	1	0	1	0	0	0	0	1	1	1	0	1	0
Y	A	A	A	A	A	A	A	A	A	A	A	B	C	C	C

Q2.1
4 Points

What is the maximum likelihood estimate of the prior $P(Y)$?

$P(Y = A)$:

Enter your answer here

$P(Y = B)$:

Enter your answer here

$P(Y = C)$:

Enter your answer here

Save Answer

Q2.2
5 Points

What are the maximum likelihood estimates of the conditional probability distributions? Fill in the probability values below (the tables for the second and third features are done for you).

$P(F_1 = 0|Y = A)$:

Enter your answer here

$P(F_1 = 1|Y = A)$:

Enter your answer here

$P(F_1 = 0|Y = B)$:

Enter your answer here

$P(F_1 = 1|Y = B)$:

Enter your answer here

$P(F_1 = 0|Y = C)$:

Enter your answer here

$P(F_1 = 1|Y = C)$:

Enter your answer here

F_2	Y	$P(F_2\|Y)$
0	A	0.545
1	A	0.455
0	B	1.000
1	B	0.000
0	C	1.000
1	C	0.000

F_3	Y	$P(F_3\|Y)$
0	A	0.455
1	A	0.545
0	B	0.000
1	B	1.000
0	C	0.667
1	C	0.333

Save Answer

Q2.3

3 Points

Now consider a new data point $(F_1 = 1, F_2 = 1, F_3 = 1)$. Use your classifier to determine the joint probability of causes Y and this new

data point, along with the posterior probability of Y given the new data:

$P(Y = A, F_1 = 1, F_2 = 1, F_3 = 1)$

Enter your answer here

$P(Y = B, F_1 = 1, F_2 = 1, F_3 = 1)$

Enter your answer here

$P(Y = C, F_1 = 1, F_2 = 1, F_3 = 1)$

Enter your answer here

Save Answer

Q2.4

3 Points

$P(Y = A|F_1 = 1, F_2 = 1, F_3 = 1)$

Enter your answer here

$P(Y = B|F_1 = 1, F_2 = 1, F_3 = 1)$

Enter your answer here

$P(Y = C|F_1 = 1, F_2 = 1, F_3 = 1)$

Enter your answer here

Save Answer

Q2.5

3 Points

What label does your classifier give to the new data point in part 3? (Break ties alphabetically)

- O A
- O B
- O C

Save Answer

Q2.6

4 Points

The training data is repeated here for your convenience:

F_1	0	0	1	1	0	0	1	0	0	0	0	1	0	1	0
F_2	0	1	0	1	0	0	1	0	1	1	0	0	0	0	0
F_3	1	1	1	0	1	0	0	0	0	1	1	1	0	1	0
Y	A	A	A	A	A	A	A	A	A	A	A	B	C	C	C

Now use Laplace Smoothing with strength k = 2 to estimate the prior $P(Y)$ for the same data.

$P(Y = A)$:

Enter your answer here

$P(Y = B)$:

Enter your answer here

$P(Y = C)$:

Enter your answer here

Save Answer

Q2.7

5 Points

Use Laplace Smoothing with strength k = 2 to estimate the conditional probability distributions below (again, the second two are done for you).

$P(F_1 = 0|Y = A)$:

Enter your answer here

$P(F_1 = 1|Y = A)$:

Enter your answer here

$P(F_1 = 0|Y = B)$:

Enter your answer here

$P(F_1 = 1|Y = B)$:

Enter your answer here

$P(F_1 = 0|Y = C)$:

Enter your answer here

$P(F_1 = 1|Y = C)$:

Enter your answer here

F_2	Y	$P(F_2\|Y)$
0	A	0.533
1	A	0.467
0	B	0.600
1	B	0.400
0	C	0.714
1	C	0.286

F_3	Y	$P(F_3\|Y)$
0	A	0.467
1	A	0.533
0	B	0.400
1	B	0.600
0	C	0.571
1	C	0.429

Save Answer

Q2.8

9 Points

Now consider again the new data point (F_1=1, F_2=1, F_3=1). Use the Laplace-Smoothed version of your classifier to determine the joint probability of causes Y and this new data point, along with the posterior probability of Y given the new data:

$P(Y = A, F_1 = 1, F_2 = 1, F_3 = 1)$

Enter your answer here

$P(Y = B, F_1 = 1, F_2 = 1, F_3 = 1)$

Enter your answer here

$P(Y = C, F_1 = 1, F_2 = 1, F_3 = 1)$

Enter your answer here

$P(Y = A|F_1 = 1, F_2 = 1, F_3 = 1)$

Enter your answer here

$P(Y = B|F_1 = 1, F_2 = 1, F_3 = 1)$

Enter your answer here

$P(Y = C|F_1 = 1, F_2 = 1, F_3 = 1)$

Enter your answer here

What label does your (Laplace-Smoothed) classifier give to the new data point? (Break ties alphabetically)

- A
- B
- C

Save Answer

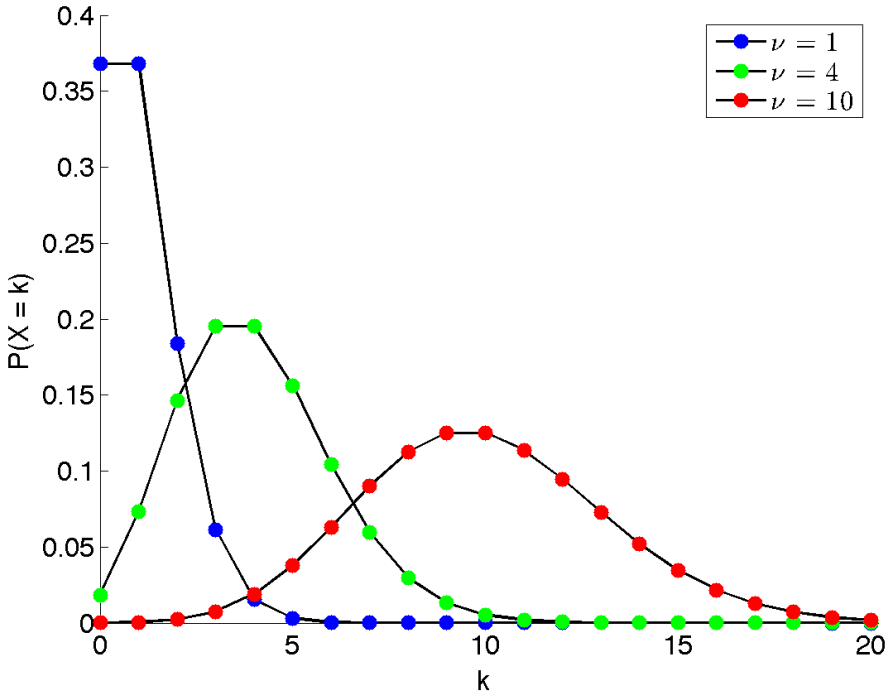
Q3 Poisson Parameter Evaluation

6 Points

We will now consider maximum likelihood estimation in the context of a different probability distribution. Under the Poisson distribution, the probability of an event occuring $X = k$ times is:

$$P(X = k) = \frac{\nu^k e^{-\nu}}{k!}$$

Here ν is the *parameter* we wish to estimate. The distribution is plotted for several values of ν below.



On a sheet of scratch paper, work out the maximum likelihood estimate for ν, given observations of several k_i. Hints: start by taking the product of the equation above over all the k_i, and then taking the log. Then, differentiate with respect to ν, set the result equal to 0, and solve for ν in terms of the k_i.

You observe the samples $k_1 = 5, k_2 = 6, k_3 = 2, k_4 = 2, k_5 = 5$. What is your maximum likelihood estimate of ν?

Enter your answer here

Save Answer

Q4 Datasets

9 Points

When training a classifier, it is common to split the available data into a training set, a hold-out set, and a test set, each of which has a different role.

Q4.1
3 Points

Which data set is used to learn the conditional probabilities?

- Training Data
- Hold-Out Data
- Test Data

Save Answer

Q4.2
3 Points

Which data set is used to tune the Laplace Smoothing hyperparameters?

- Training Data
- Hold-Out Data
- Test Data

Save Answer

Q4.3
3 Points

Which data set is used for quantifying performance results?

- Training Data
- Hold-Out Data
- Test Data

Save Answer

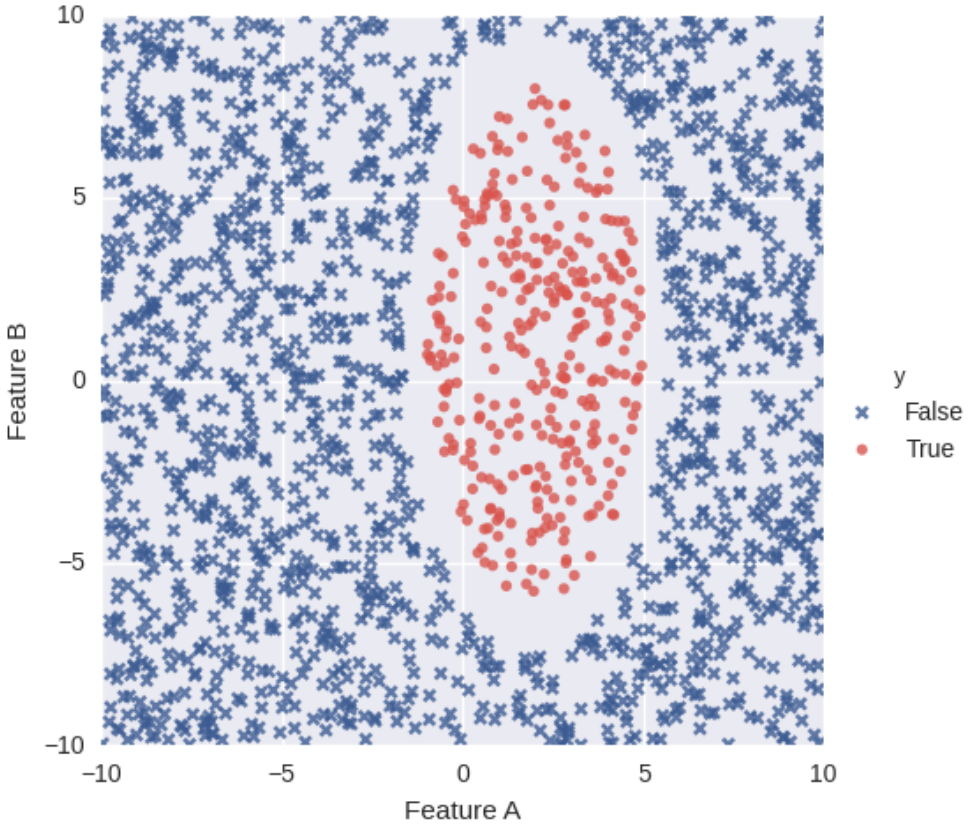
Q5 Separability

20 Points

Q5.1

10 Points

It is well known that Pactronic LLC is the premier manufacturer of Pacmen. At Pactronic LLC, quality control is currently done manually -- a group of scientists decide whether a Pacman is ready to be released into the wild based on (Feature (A)) a Pacman's intelligence score and (Feature (B)) a Pacman's empathy score. Here are many examples of Pacmen that have been released and withheld in the past. Each dot corresponds to a Pacman, and responds to the following question as true or false: this Pacman is ready to be released.



As the Vice President of Science, you would like to automate the decision making process, and decide to use the perceptron algorithm. Which of the following subsets of features would allow you to perfectly classify whether or not a Pacman can be released in the wild?

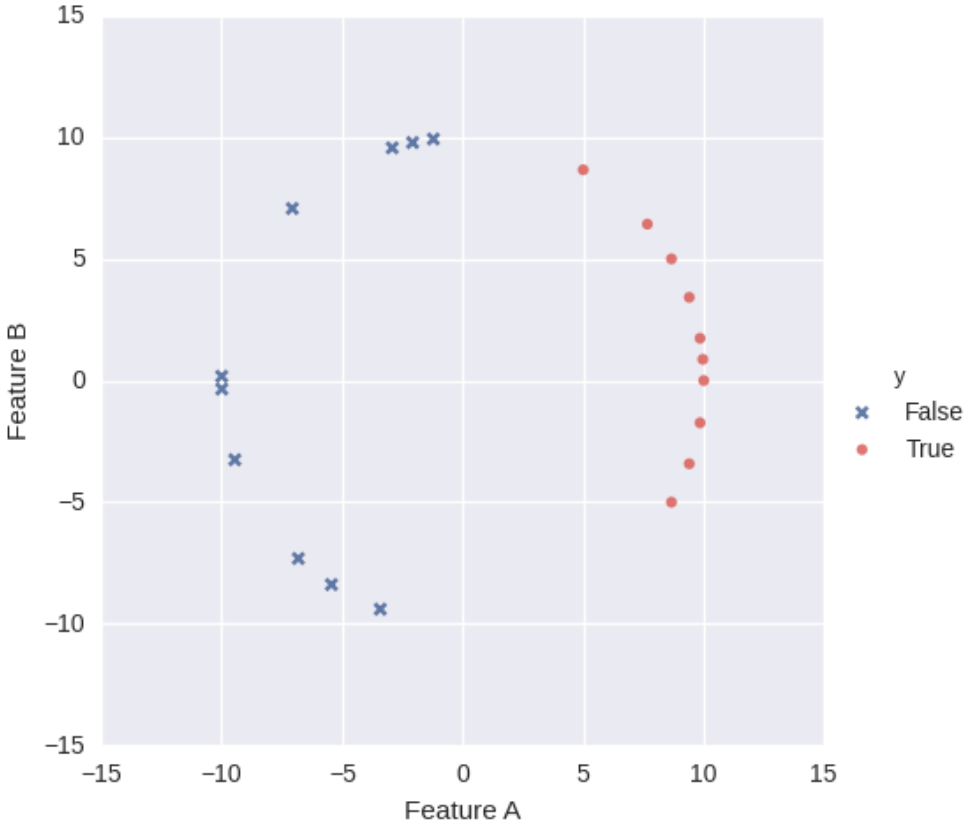
- [ ] (A, B)
- [ ] (A^2, AB, B^2, A, B)
- [ ] (A, B, X), where $(X = (A \geq C_1) \wedge (B \geq C_2)$ for some fixed (C_1, C_2) that you are allowed to pick.
- [ ] (A)
- [ ] (B)

Save Answer

Q5.2

10 Points

The CEO of Pactronic soon decides that the company will be focusing on creating fewer, but much better Pacmen. This calls for an entire re-design of the Pacman. Accordingly, the scientists come up with the latest and greatest generation of Pacmen, and once again seek your advice in quality control. Here are the newest Pacman, and their respective features:



Which of the following subsets of features would allow you to perfectly classify whether or not a Pacman can be released in the wild?

- [ ] (A, B)
- [ ] (A^2, AB, B^2, A, B)
- [ ] (A, B, X), where $(X = (A \geq C_1) \wedge (B \geq C_2)$ for some fixed (C_1, C_2) that you are allowed to pick.
- [ ] (A)
- [ ] (B)

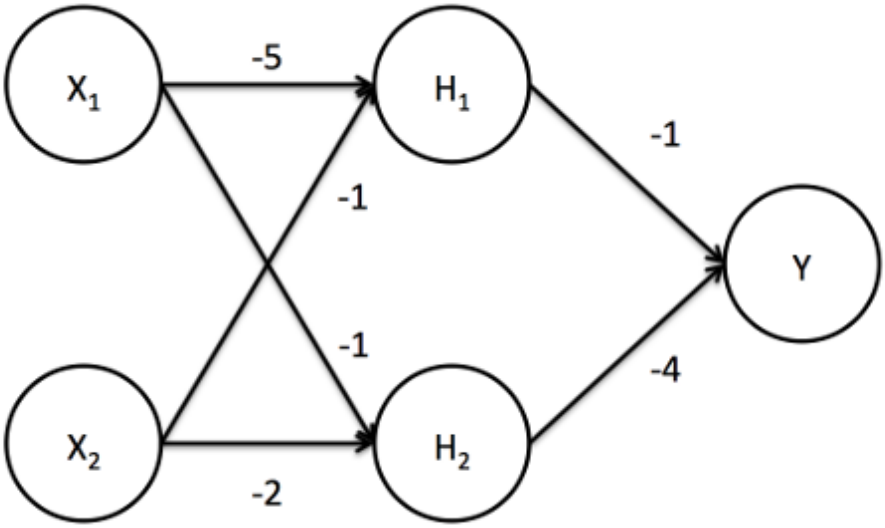
Save Answer

Q6 Neural Networks and Logic Gates

36 Points

As you probably know, Pacumus Maximus Corporation (PMC) is the most well known company that manufactures the gears that let Pacman open and close its mouth. Recently, the Vice President of Science in PMC decided to replace all of its low-level NAND, AND, and NOR gates with mini neural networks. Unfortunately, there was a Pacman uprising, where the Pacmen took over the factory, eating all of the neural network documentation. The scientists were able to salvage the following neural networks weights, but don't remember which gates these neural networks corresponded to. They've hired you to help them recover this information.

Here is the first network that you're given:



Above, the nodes H_1, H_2, and Y have sigmoid activations and each have biases of 0.5. The inputs are placed in X_1 and X_2. To convert the output Y into a boolean value, we round Y. This will be the case for all problems in this section.

Concretely, we have the following, where w_{AB} denotes the weight between nodes A and B:

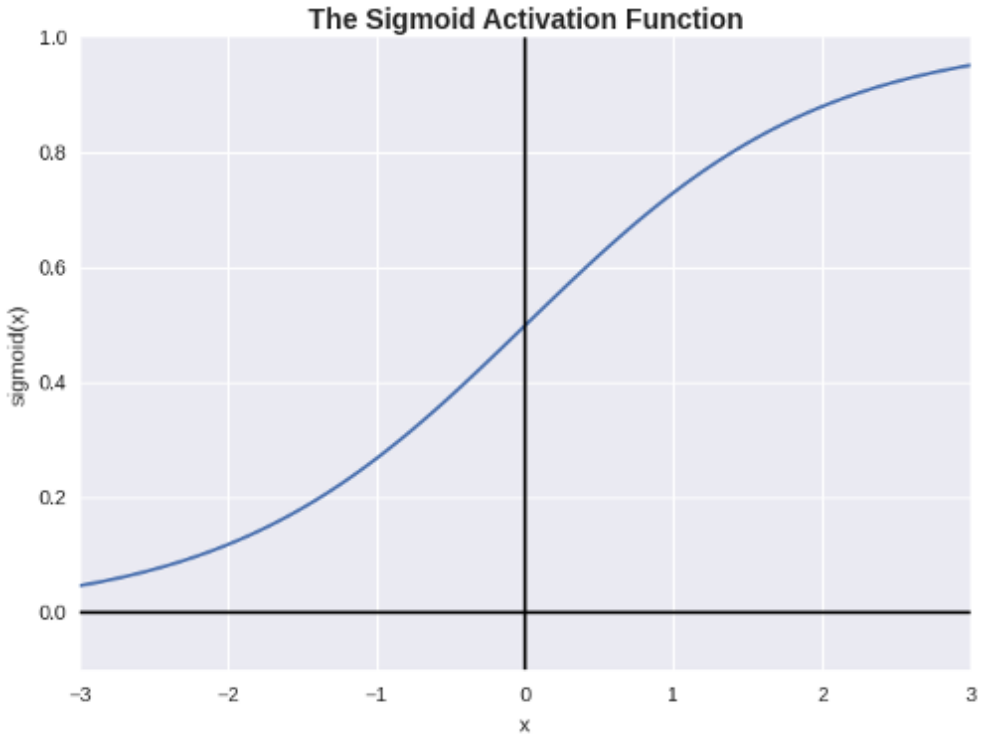
$$H_1(x) = \sigma(w_{H_1X_1} \cdot X_1 + w_{H_1X_2} \cdot X_2 + 0.5)$$

$$H_2(x) = \sigma(w_{H_2X_1} \cdot X_1 + w_{H_2X_2} \cdot X_2 + 0.5)$$

$$Y(x) = \text{round}\{\sigma(w_{YH_1} \cdot H_1 + w_{YH_2} \cdot H_2 + 0.5)\}$$

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

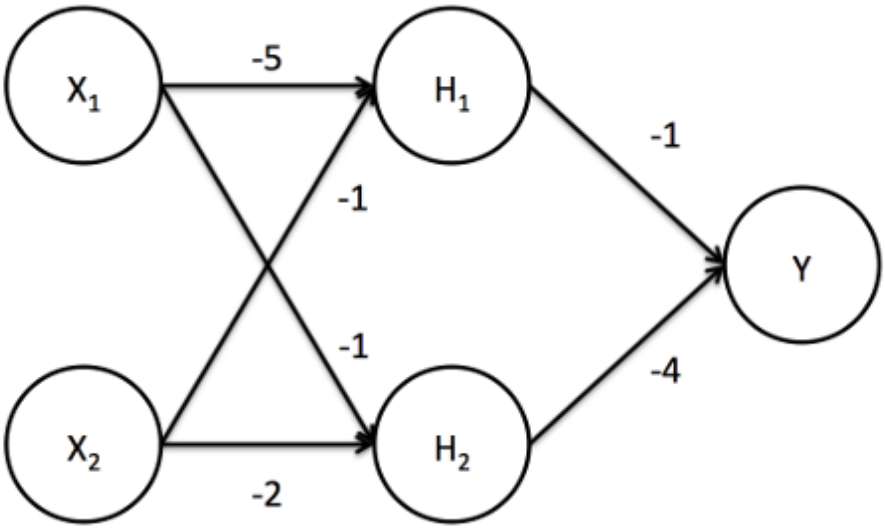
Recall that the sigmoid function, $\sigma(x)$, looks like this:



Q6.1

9 Points

Here's the first network again, for convenience:



What does this first network correspond to?

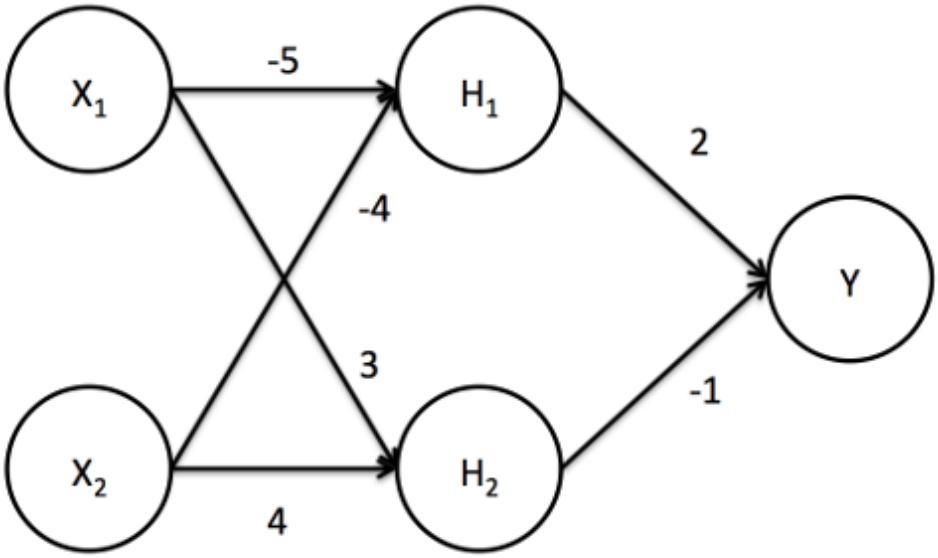
- NAND
- AND
- NOR

Save Answer

Q6.2

9 Points

Here is the second network that you're given:



What does this network correspond to?

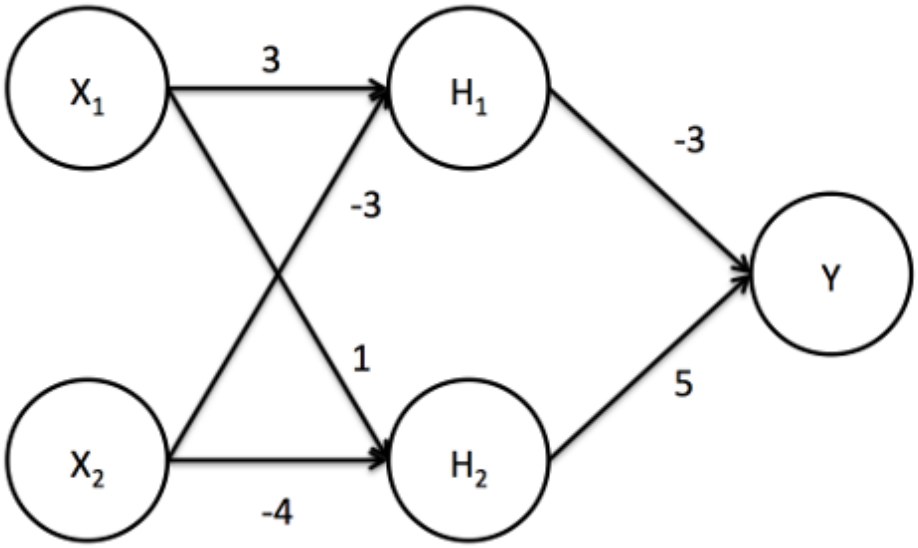
- NAND
- AND
- NOR

Save Answer

Q6.3

9 Points

Here is the third network that you're given:



What does this network correspond to?

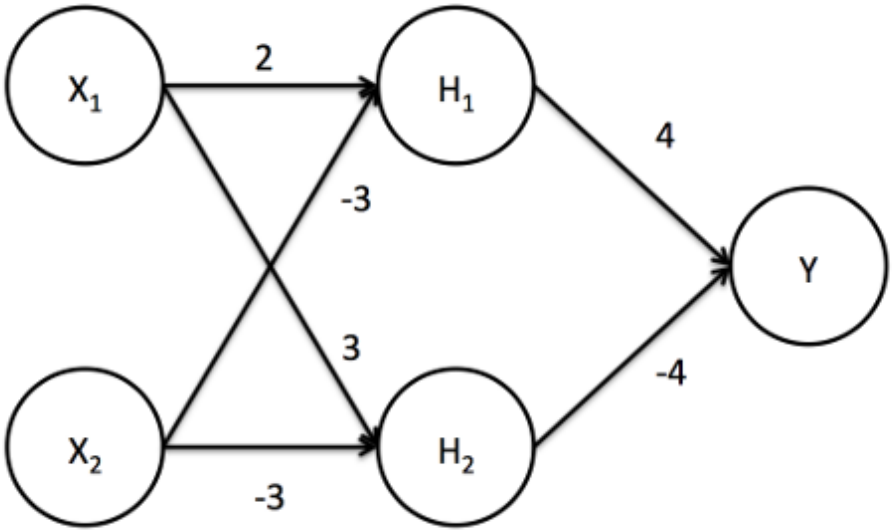
- NAND
- AND
- NOR

Save Answer

Q6.4

9 Points

Here is the final network that you're given:



What does this network correspond to?

- NAND
- AND
- NOR

Save Answer

Q7 Decision Tree

10 Points

Given the following data points: {(1, 2, 0), (4, 1, 1), (5, 4, 0), (6, 6, 0), (7, 3, 1), (9, 5, 1)}. Each data point is represented as (f_1, f_2, Y), where (f_1, f_2) are numerical features, and $Y \in \{0, 1\}$ is the label. We will build a decision tree to classify the data.

Q7.1

2 Points

Let's use f_1 to do the first split at the root of the decision tree. Suppose we want to maximize the information gain, what is the threshold value c_1 we should set when the rule is $f_1 \leq c_1$?

- 1
- 4
- 5
- 6
- 7
- 9

Save Answer

Q7.2
2 Points

With the splitting rule in part 1, what is the entropy of the subset of $f_1 \leq c_1$?

Enter your answer here

Save Answer

Q7.3
2 Points

With the splitting rule in part 1, what is the information gain?

Enter your answer here

Save Answer

Q7.4
2 Points

For the subset with $f_1 \leq c_1$, suppose the rule for the second split is $f_2 \leq c_2$. What is c_2? Again, we want to maximize the information gain.

- ○ 1
- ○ 2
- ○ 3
- ○ 4
- ○ 5
- ○ 6

Save Answer

Q7.5

2 Points

Suppose you are allowed to swap the labels of two data points, so that all data points could be classified with a decision tree of the smallest height. Which two points will you select?

- ☐ $(1, 2, 0)$
- ☐ $(4, 1, 1)$
- ☐ $(5, 4, 0)$
- ☐ $(6, 6, 0)$
- ☐ $(7, 3, 1)$
- ☐ $(9, 5, 1)$

Save Answer

Save All Answers

Submit & View Submission >