Homework 6: Learning Theory and Generative Models

Due Apr 4 at 11:59pm **Points** 100 **Questions** 29
Available until Apr 8 at 11:59pm **Time Limit** None

Instructions

Due date: Wednesday, April 4 at 11:59PM
TAs: Bowei, Mo, Soham, Brynn

Homework 6 covers topics including Learning Theory, MLE/MAP, Naïve Bayes, and the guest lectures.

Important note on questions with numerical answers: When you enter numerical answers on Canvas, Canvas will automatically convert it to a decimal with **four** decimal places. So if you enter '7', Canvas will convert it into '7.0000'; if you enter '7.343', Canvas will convert into '7.3430'; if you enter '7.34388', Canvas will convert it into '7.3439'. Please just answer the numerical answers according to the questions, and don't worry about the extra 0s that get added at the end.

Important note on scoring on multiple answer questions: You will notice that there are questions like "Select all answers that are correct." in the assignment. The top right corner of the question shows the number of points the question is worth. Please keep in mind that the way Canvas grades this type of questions is as follows: Canvas divides the total points possible by the amount of correct answers for that question. This amount is awarded for every correct answer selected and deducted for every incorrect answer selected. For example, if the question has 2 options, is worth 2 points, and exactly one of the options is correct, then if you select both options, you would receive 2 - 2 = 0 points. However, the minimum score you could get on a problem is 0.

Important: Only **one submission** is allowed for this homework. Please make sure you're confident about your answers before you submit.

This quiz was locked Apr 8 at 11:59pm.

Attempt History

	Attempt	Time	Score
LATEST	Attempt 1	4,423 minutes	92 out of 100 *

* Some questions not yet graded

Correct answers are hidden.

Score for this quiz: **92** out of 100 *
Submitted Apr 2 at 5:13pm
This attempt took 4,423 minutes.

Learning Theory

Question 1

4 / 4 pts

Let $\delta = |H|e^{-\epsilon m}$. According to the PAC theorems discussed in class, which of the following is correct? Select one.

○ With probability at least 1-δ, every hypothesis with training error at most ϵ has true error 0.

○ With probability at least 1-ϵ, a random hypothesis with training error 0 has true error at most δ.

● With probability at least 1-δ, every hypothesis with training error 0 has true error at most ϵ.

○ With probability at least 1-ϵ, a random hypothesis with true error 0 has training error at most δ.

Question 2

4 / 4 pts

Consider a decision tree learner applied to data where each example is described by 10 boolean variables $X_1, X_2, \cdots, X_{10}$. What is the VC dimension of the hypothesis space used by this decision tree learner?

1024.0000

Question 3

4 / 4 pts

Consider instance space X which is the set of real numbers. What is the VC dimension of hypothesis class H, where each hypothesis h in H is of the form "if a < x < b or c < x < d then y = 1; otherwise y = 0"? (i.e., H is an infinite hypothesis class where a, b, c, and d are arbitrary real numbers.)

○ 2

○ 3

● 4

○ 5

○ 6

MLE/MAP

Question 4

3 / 3 pts

Suppose you place a Beta prior over the Bernoulli distribution, and attempt to learn the parameter of the Bernoulli distribution from data. Further suppose an adversary chooses "bad", but finite hyperparameters for your Beta prior in order to confuse your learning algorithm. As the number of training examples grows to infinity, the MAP estimate of θ can still converge to the MLE estimate of θ,

● True

○ False

Question 5

4 / 4 pts

Let θ be a random variable with the following probability density function (pdf):

$f(\theta) = 2\theta$ if $0 \leq \theta \leq 1$, otherwise $f(\theta) = 0$

Suppose another random variable Y, which is conditioning on θ, follows an exponential distribution with $\lambda = 3\theta$. Recall that the exponential distribution with parameter λ has the following pdf:

$f(y) = \lambda e^{-\lambda y}$ if $y \geq 0$, otherwise $f(y) = 0$

What is the MAP estimate of θ given $y = \frac{2}{3}$ is observed?

○ 0

○ 1/3

● 1

○ 2

Important Note: For simplicity, we are not considering a bias term for questions 6-13.

Question 6

3 / 3 pts

In HW3, you have derived the closed form solution for linear regression. Now, we are coming back to linear regression, viewing it as a statistical model, and deriving the MLE and MAP estimate of the parameters in the following questions.

Assume we have data $D = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^{N}$, where $\mathbf{x}^{(i)} = (x_1^{(i)}, \cdots, x_M^{(i)})$. So our data has N instances and each instance has M attributes/features. Each $y^{(i)}$ is generated given $\mathbf{x}^{(i)}$ with additive noise $\epsilon^{(i)} \sim N(0, \sigma^2)$, that is $y^{(i)} = \mathbf{w}^T \mathbf{x}^{(i)} + \epsilon^{(i)}$ where $\mathbf{w}$ is the parameter vector of linear regression. Given this assumption, what is the distribution of y?

● $y^{(i)} \sim N(\mathbf{w}^T\mathbf{x}^{(i)}, \sigma^2)$

○ $y^{(i)} \sim N(0, \sigma^2)$

○ $y^{(i)} \sim Uniform(\mathbf{w}^T\mathbf{x}^{(i)} - \sigma, \mathbf{w}^T\mathbf{x}^{(i)} + \sigma)$

○ None of the above.

Question 7

4 / 4 pts

The next step is to learn the MLE of the parameters of the linear regression model. Which expression below is the correct conditional log likelihood $\ell(\mathbf{w})$ with the given data?

● $\sum_{i=1}^{N}[-\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})^2]$

○ $\sum_{i=1}^{N}[\log(\sqrt{2\pi\sigma^2}) + \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})^2]$

○ $\sum_{i=1}^{N}[-\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})]$

○ $-\log(\sqrt{2\pi\sigma^2}) + \sum_{i=1}^{N}[-\frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})^2]$

Question 8

4 / 4 pts

Then, the MLE of the parameters is just $argmax_{\mathbf{w}}\ell(\mathbf{w})$. Among the following expressions, select ALL that can yield the correct MLE.

□ $argmax_{\mathbf{w}} \sum_{i=1}^{N}[-\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})]$

☑

$argmax_{\mathbf{w}} \sum_{i=1}^{N} [-\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})^2]$

☑ $argmax_{\mathbf{w}} \sum_{i=1}^{N} [-\frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})^2]$

□ $argmax_{\mathbf{w}} \sum_{i=1}^{N} [-\frac{1}{2}(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})]$

☑ $argmax_{\mathbf{w}} \sum_{i=1}^{N} [-\frac{1}{2}(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})^2]$

Question 9

3 / 3 pts

According to the above derivations, is the MLE for the conditional log likelihood equivalent to minimizing mean squared errors (MSE) for the linear regression model when making predictions? Why or why not?

○

Yes, because the derivative of the negative conditional log-likelihood has the same form as the derivative of the MSE loss.

●

Yes, because the parameters that maximize the conditional log-likelihood also minimize the MSE loss.

○

No, because one is doing maximization and the other is doing minimization.

○

No, because the MSE has an additional error term $\epsilon^{(i)}$ in the expression whereas the quantity to be minimized in MLE does not.

○

No, because the conditional log-likelihood has additional constant terms that do not appear in the MSE loss.

Question 10

4 / 4 pts

Now we are moving on to learn the MAP estimate of the parameters of the linear regression model. The MAP estimate is obtained through solving the following optimization problem.

$$\mathbf{w}_{MAP} = \arg\max_{\mathbf{w}} p(\mathbf{w}|D) = \arg\max_{\mathbf{w}} p(D, \mathbf{w})$$

Suppose are using a Gaussian prior distribution with mean 0 and variance $\frac{1}{\lambda}$ for each element w_m of the parameter vector $\mathbf{w}$ ($1 \leq m \leq M$), i.e. $w_m \sim N(0, \frac{1}{\lambda})$. Assume that $w_1, \cdots, w_M$ are mutually independent of each other. Which expression below is the correct log joint-probability of the data and parameters $\log p(D, \mathbf{w})$?

(For simplicity, just use $p(D|\mathbf{w})$ to denote the data likelihood.)

○ $\log p(D|\mathbf{w}) - \sum_{m=1}^{M} \log(\sqrt{2\pi\lambda}) - \lambda(w_m)^2$

○ $\log p(D|\mathbf{w}) + \sum_{m=1}^{M} - \log(\sqrt{2\pi\lambda}) - \lambda(w_m)^2$

○ $\log p(D|\mathbf{w}) - \sum_{m=1}^{M} \log(\sqrt{\frac{2\pi}{\lambda}}) - \frac{\lambda}{2}(w_m)^2$

● $\log p(D|\mathbf{w}) + \sum_{m=1}^{M} - \log(\sqrt{\frac{2\pi}{\lambda}}) - \frac{\lambda}{2}(w_m)^2$

Question 11

4 / 4 pts

A MAP estimator with a Gaussian prior $\mathcal{N}(0, \sigma^2)$ you trained gives significantly higher test error than train error. What could be a possible approach to fixing this?

○ Increase variance σ^2

● Decrease variance σ^2

○ Try MLE estimator instead

○ None of the above

Question 12

4 / 4 pts

Maximizing the log posterior probability $\ell_{MAP}(\mathbf{w})$ gives you the MAP estimate of the parameters. The MAP estimate with Gaussian prior is actually equivalent to a L$_2$ regularization on the parameters of linear regression model in minimizing an objective function $J(\mathbf{w})$ that consists of a term related to log conditional likelihood $\ell(\mathbf{w})$ and a L$_2$ regularization term. The following options specify the two terms in $J(\mathbf{w})$ explicitly. Which one is correct based on your derived log posterior probability in the previous question?

○ $-\ell(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|_2$

◉ $-\ell(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$

○ $-\ell(\mathbf{w}) + \lambda\|\mathbf{w}\|_2$

○ $\ell(\mathbf{w}) - \frac{\lambda}{2}\|\mathbf{w}\|_2^2$

Question 13

3 / 3 pts

MAP estimation with what prior is equivalent to L$_1$ regularization?

Note:

- The pdf of a Uniform distribution over [a,b] is $f(x) = \frac{1}{b-a}$ if $x \in [a, b]$ and 0 otherwise.
- The pdf of an exponential distribution with rate parameter a is $f(x) = a\exp(-ax)$ for $x > 0$.
- The pdf of a Laplace distribution with location parameter a and scale parameter b is $f(x) = \frac{1}{2b}\exp\left(-\frac{|x-a|}{b}\right)$ for all $x \in \mathbb{R}$

○ Uniform distribution over $[-\mathbf{w}^T\mathbf{x}^{(i)}, \mathbf{w}^T\mathbf{x}^{(i)}]$

○ Exponential distribution with rate parameter $a = \frac{1}{2}$

○ Exponential distribution with rate parameter $a = \mathbf{w}^T \mathbf{x}^{(i)}$

● Laplace prior with location parameter $a = 0$

○ Laplace prior with location parameter $a = \mathbf{w}^T \mathbf{x}^{(i)}$

○ Uniform distribution over [-1, 1]

Naive Bayes

Question 14

4 / 4 pts

I give you the following fact: for events A and B, P(A|B) = 2/3 and P(A|~B) = 1/3, where ~B denotes the complement of B. Do you have enough information to calculate P(B|A)? If not, choose "not enough information", if so, compute the value of P(B|A).

○ 1/2

○ 2/3

○ 1/3

● Not enough information

Question 15

4 / 4 pts

Instead if I give you for events A and B, P(A|B) = 2/3, P(A|~B) = 1/3 and P(B) = 1/3 and P(A) = 4/9, where ~B denotes the complement of B. Do you have information to calculate P(B|A)? If not, choose "not enough information", if so, compute the value of P(B|A).

● 1/2

○ 1/3

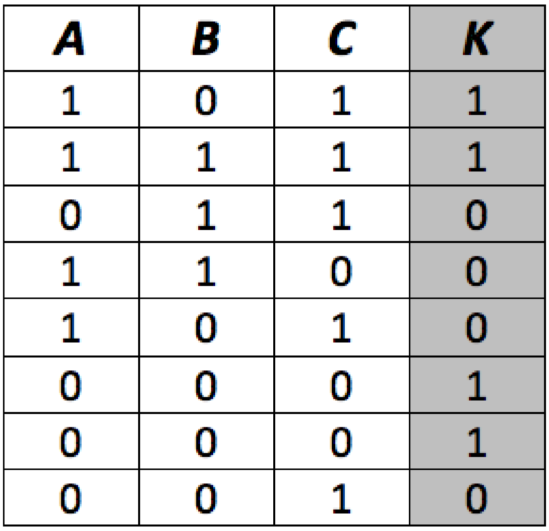
○ 2/3

○ Not enough information

Question 16

4 / 4 pts

Suppose you are given the following set of data with three Boolean input variables A, B, and C, and a single Boolean output variable K.



A	B	C	K
1	0	1	1
1	1	1	1
0	1	1	0
1	1	0	0
1	0	1	0
0	0	0	1
0	0	0	1
0	0	1	0

Suppose you train a Naive Bayes classifier without any priors. According to the Naive Bayes classifier, what is P(K = 1 | A = 1, B = 1, C = 0)? Please leave your answer in 4 decimal places.

0.5000

Question 17

4 / 4 pts

Using the same table from the previous question, according to the Naive Bayes classifier, what is P(K = 0|A = 1, B = 1)? Please leave your answer in 4 decimal places.

0.6667

Question 18

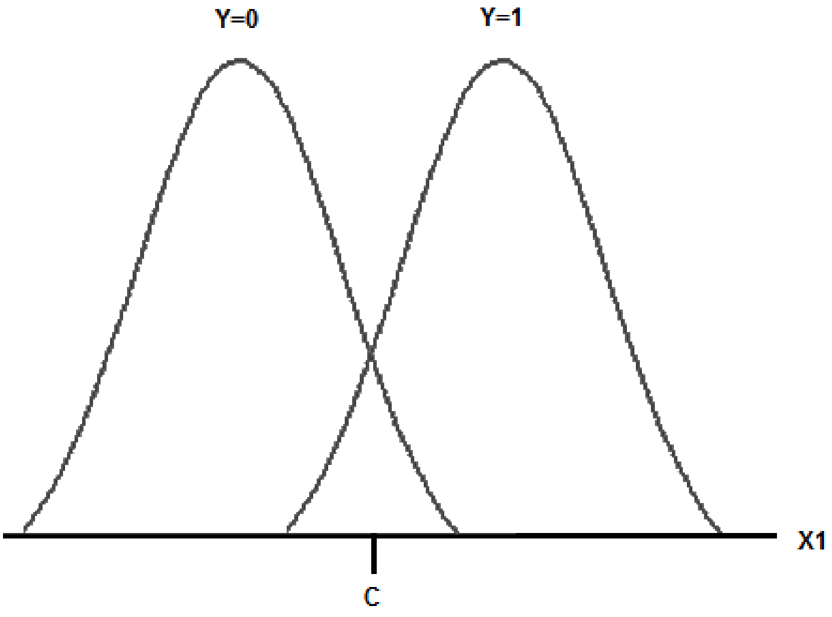
4 / 4 pts

Gaussian Naive Bayes in general can learn non-linear decision boundaries. Consider the simple case where we have just one real-valued feature $X_1 \in \mathbb{R}$ from which we wish to infer the value of label $Y \in \{0, 1\}$. The corresponding generative story would be:

$$Y \sim \text{Bernoulli}(\phi)$$
$$X_1 \sim \text{Gaussian}(\mu_y, \sigma_y^2)$$

where the parameters are the Bernoulli parameter ϕ and the class-conditional Gaussian parameters μ_0, σ_0^2 and μ_1, σ_1^2 corresponding to $Y = 0$ and $Y = 1$, respectively.

A linear decision boundary in one dimension, of course, can be described by a rule of the form "if $X_1 > c$ then $Y = 1$, else $Y = 0$", where c is a real-valued threshold (see diagram provided). Is it possible in this simple one-dimensional case to construct a Gaussian Naive Bayes classifier with a decision boundary that cannot be expressed by a rule in the above form)?



If X1>C then Y=1, else Y=0

○ Yes, this can occur if the Gaussians are of equal means and equal variances.

● Yes, this can occur if the Gaussians are of equal means and unequal variances.

○ Yes, this can occur if the Gaussians are of unequal means and equal variances.

○ No, this cannot occur regardless of the relationship of the means or variances.

Incorrect

Question 19

0 / 4 pts

Suppose that 0.3% people have cancer. Someone decided to take a medical test for cancer. The outcome of the test can either be positive (cancer) or negative (no cancer). The test is not perfect - among people who have cancer, the test comes back positive 97% of the time. Among people who don’t have cancer, the test comes back positive 4% of the time. For this question, you should assume that the test results are independent of each other, given the true state (cancer or no cancer). What is the probability of a test subject having cancer, given that the subject’s test result is positive? (leave your answer in 3 decimal places)

0.0630

Incorrect

Question 20

0 / 4 pts

In a Naive Bayes problem, suppose we are trying to compute $P(Y \mid X_1, X_2, X_3, X_4)$. Furthermore, suppose X_2 and X_3 are identical (i.e., X_3 is just a copy of X_2). Which of the following are true in this case?

☑

Naive Bayes will learn identical parameter values for $P(X_2|Y)$ and $P(X_3|Y)$.

□

Naive Bayes will output probabilities $P(Y|X_1, X_2, X_3, X_4)$ that are closer to 0 and 1 than they would be if we removed the feature corresponding to X_3.

☑

This will not raise a problem in the output $P(Y|X_1, X_2, X_3, X_4)$ because the conditional independence assumption will correctly treat this situation.

□ None of the above

Question 21

3 / 3 pts

Which of the following machine learning algorithms are probabilistic generative models?

□ Decision tree

□ K-nearest neighbors

□ Perceptron

☑ Naive Bayes

□ Logistic regression

□ Feed-forward neural network

Guest Lectures

Question 22

3 / 3 pts

Which of the following are properties of a max-pooling layer in a Convolutional Neural Network (CNN)?

□ A max-pooling layer can normalize the inputs to produce a probability distribution

☑ A max-pooling layer can produce an output that has smaller size than its input

□ A max-pooling layer can create recurrent connections among the layers

□ None of the above

Question 23

3 / 3 pts

Convolutional neural networks often consist of convolutional layers, max-pooling layers, and fully-connected layers. Select all the layer types below that have parameters (i.e. weights) which can be learned by gradient descent / backpropagation.

☑ convolutional layer

□ max-pooling layer

☑ fully-connected layer

Question 24

3 / 3 pts

Consider the black-and-white 5 pixel by 5 pixel image shown below. Black pixels are represented by the value 0 and white pixels by 1.

1	1	0	0	0
1	1	1	0	0

0	1	1	1	0
0	0	1	1	1
0	0	0	1	1

Next consider the 3 by 3 convolution with weights shown below.

-1	0	1
-1	0	1
-1	0	1

Suppose we apply the above convolution (as one would in a CNN convolutional layer) to the above image to produce a new output image. Assume that we do **not** permit any padding to be used and the stride of the convolution is 1. What is the value of the pixel in the upper-left corner of the output image?

(**Important Note**: Convolution is sometimes defined *differently* in machine learning than in other fields, such as signal processing. So be sure to follow the method of convolution shown in the guest lecture on CNNs.)

0.0000

Question 25

3 / 3 pts

For the same output image produced in the previous question, what is the value of the pixel in the upper-right corner of the output image?

-2.0000

Question 26

3 / 3 pts

Long Short Term Memory (LSTM) networks partially address the vanishing gradient problem by incorporating input, output, and forget gates.

● True

○ False

Question 27

3 / 3 pts

Recurrent neural networks (RNNs) have more representational power than standard feed-forward neural networks.

● True

○ False

Question 28

3 / 3 pts

Recurrent neural networks (RNNs) can accept sequence data as input, but can only output a single classification decision for that input sequence.

○ True

● False

Question 29

Not yet graded / 0 pts

Collaboration Policy:

Please answer the following questions:

1. Did you receive any help whatsoever from anyone in solving this assignment? Yes / No.
   - If you answered ‘yes’, give full details: ____________

- (e.g. “Jane Doe explained to me what is asked in Question 3.4”)

2. Did you give any help whatsoever to anyone in solving this assignment? Yes / No.
   - If you answered ‘yes’, give full details: _____________
   - (e.g. “I pointed Joe Smith to section 2.3 since he didn’t know how to proceed with Question 2”)
3. Did you find or come across code that implements any part of this assignment ? Yes / No. (See below policy on “found code”)
   - If you answered ‘yes’, give full details: _____________
   - (book & page, URL & location within the page, etc.).

Your Answer:

1. No.

2. No.

3. No.

Quiz Score: **92** out of 100