Homework 6
Learning Theory and Generative Models[1]

CMU 10-601: Machine Learning (Fall 2018)

piazza.com/cmu/fall2018/10601bd
OUT: Thursday, Nov 1st, 2018

DUE: Wednesday, Nov 7th, 2018, 11:59pm
TAs: Brynn Edmunds, Zhuoran Liu, Emilio Arroyo-Fang, George Xu

START HERE: Instructions

Homework 6 covers topics on learning theory, MLE/MAP, and Naive Bayes. The homework includes multiple choice, True/False, and short answer questions.

- **Collaboration policy:** Collaboration on solving the homework is allowed, after you have thought about the problems on your own. It is also OK to get clarification (but not solutions) from books or online resources, again after you have thought about the problems on your own. There are two requirements: first, cite your collaborators fully and completely (e.g., "Jane explained to me what is asked in Question 2.1"). Second, write your solution *independently*: close the book and all of your notes, and send collaborators out of the room, so that the solution comes from you only. See the Academic Integrity Section on the course site for more information: http://www.cs.cmu.edu/~mgormley/courses/10601bd-f18/about.html#7-academic-integrity-policies
- **Late Submission Policy:** See the late submission policy here: http://www.cs.cmu.edu/~mgormley/courses/10601bd-f18/about.html#6-general-policies
- **Submitting your work:**
    - **Gradescope:** For written problems such as short answer, multiple choice, derivations, proofs, or plots, we will be using Gradescope (https://gradescope.com/). Please use the provided template. Submissions can be handwritten onto the template, but should be labeled and clearly legible. If your writing is not legible, you will not be awarded marks. Alternatively, submissions can be written in LaTeX. Regrade requests can be made, however this gives the TA the opportunity to regrade your entire paper, meaning if additional mistakes are found then points will be deducted. Each derivation/proof should be completed on a separate page. For short answer questions, you **should not** include your work in your solution. If you include your work in your solutions, your assignment may not be graded

[1] Compiled on Thursday 1$^{\text{st}}$ November, 2018 at 10:23

correctly by our AI assisted grader. In addition, please tag the problems to the corresponding pages when submitting your work.

For multiple choice or select all that apply questions, shade in the box or circle in the template document corresponding to the correct answer(s) for each of the questions. For LaTeXusers, use ■ and ●for shaded boxes and circles, and don’t change anything else.

Instructions for Specific Problem Types

For "Select One" questions, please fill in the appropriate bubble completely:

Select One: Who taught this course?

● Matt Gormley

○ Marie Curie

○ Noam Chomsky

If you need to change your answer, you may cross out the previous answer and bubble in the new answer:

Select One: Who taught this course?

● Matt Gormley

○ Marie Curie

⊗ Noam Chomsky

For "Select all that apply" questions, please fill in all appropriate squares completely:

Select all that apply: Which are scientists?

■ Stephen Hawking

■ Albert Einstein

■ Isaac Newton

□ I don't know

Again, if you need to change your answer, you may cross out the previous answer(s) and bubble in the new answer(s):

Select all that apply: Which are scientists?

■ Stephen Hawking

■ Albert Einstein

■ Isaac Newton

☒ I don't know

For questions where you must fill in a blank, please make sure your final answer is fully included in the given space. You may cross out answers or parts of answers, but the final answer must still be within the given space.

Fill in the blank: What is the course number?

10-601

10-~~7~~601

1 Learning Theory [22 pts]

1. [**3 pt**] Let $\delta = |H|e^{-\epsilon m}$. According to the PAC theorems discussed in class, which of the following is correct? Select one.

   Select one:

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

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

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

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

2. [**3 pt**] 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?

   Fill in the blank:

   □

3. [**4 pt**] 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.

   Select one:

   ○ 2

   ○ 3

   ○ 4

   ○ 5

   ○ 6

4. [**3 pt**] Alex is given a classification task to solve. He has no idea where to start, so he decided to try out a decision tree learner with 2 binary features X_1 and X_2. He recently learned about PAC learning, and would like to know what is the minimum number (N) of data points that would suffice for the PAC criterion with $\epsilon = 0.1$ and $\delta = 0.01$.

   Notice that a valid decision tree may or may not be full, meaning it doesn't have to split on all features.

   Fill in the blank:

5. [**3 pt**] Sally thinks Alex shouldn't have used a decision tree with 2 binary features. Instead, she thinks it would be best to use logistic regression with 16 real-valued features in addition to a bias term. Sally overherd Alex talking about this cool concept called PAC learning and she too would like to use it to analyze her method. She first trains her logistic regression model on N examples to obtain a training error $\hat{R}$. What is the the upper bound on the true error R in terms of $\hat{R}$, δ, and N. You may use big-$\mathcal{O}$ notation.

   Fill in the blank:

6. [**3 pt**] Sally wants to argue her method has lower bound on the true error. Assuming Sally has obtained enough data points to satisfy PAC criterion with $\epsilon = 0.1$ and $\delta = 0.01$. Which of the following is true?

   Select one:

   ○ Sally is wrong. Alex's method will always classify unseen data more accurately since it is simpler as it only needs 2 binary features.

   ○ She must first regularize her model by removing 14 features to make any comparison at all.

   ○ It is sufficient to show that the VC Dimension of her classifier is higher than Alex's, therefore having lower bound for the true error.

   ○ It is necessary to show that the training error she achieves is lower than the training error Alex achieves.

7. [**3 pt**] Write an English description of VC Dimension.

2 MLE/MAP [34 pts]

1. [**3 pt**] **True or False:** 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.

   Select One:

   ○ True

   ○ False

2. [**3 pt**] Let θ be a random variable with the following probability density function (pdf):

$$f(\theta) = \begin{cases} 2\theta & \text{if } 0 \leq \theta \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

   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:

$$f_{exp}(y) = \begin{cases} \lambda e^{-\lambda y} & \text{if } y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

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

   Select one:

   ○ 0

   ○ 1/3

   ○ 1

   ○ 2

3. **[3 pt]** 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?

   Select one:

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

   ○ $y^{(i)} \sim N(0, \sigma^2)$

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

   ○ None of the above

4. **[4 pt]** 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?

   Select one:

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

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

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

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

5. **[4 pt]** Then, the MLE of the parameters is just $\text{argmax}_{\mathbf{w}}\, \ell(\mathbf{w})$. Among the following expressions, select ALL that can yield the correct MLE.

   Select all that apply:

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

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

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

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

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

6. **[3 pt]** 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?

   Select one:

   ○ 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.

7. **[3 pt]** 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.)

   Select one:

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

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

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

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

8. **[3 pt]** 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?

Select one:

○ Increase variance σ^2

○ Decrease variance σ^2

○ Try MLE estimator instead

○ None of the above

9. **[4 pt]** 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 L2 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 L2 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?

Select one:

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

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

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

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

10. **[4 pt]** MAP estimation with what prior is equivalent to L1 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 $x \in \mathbb{R}$.

Select one:

○ 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]

3 Naive Bayes [44 pts]

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

   Select one:

   ○ 1/2

   ○ 2/3

   ○ 1/3

   ○ Not enough information

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

   Select one:

   ○ 1/2

   ○ 2/3

   ○ 1/3

   ○ Not enough information

3. [**4 pt**] 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 $P(K = 1 \mid A = 1, B = 1, C = 0)$?

If your answer is in decimals, answer with precision 4, e.g. (6.051, 0.1230, 1.234e+7).

Fill in the blank:

4. [**4 pt**] Using the same table from the previous question, according to the Naive Bayes classifier, what is $(K = 0 \mid A = 1, B = 1)$?

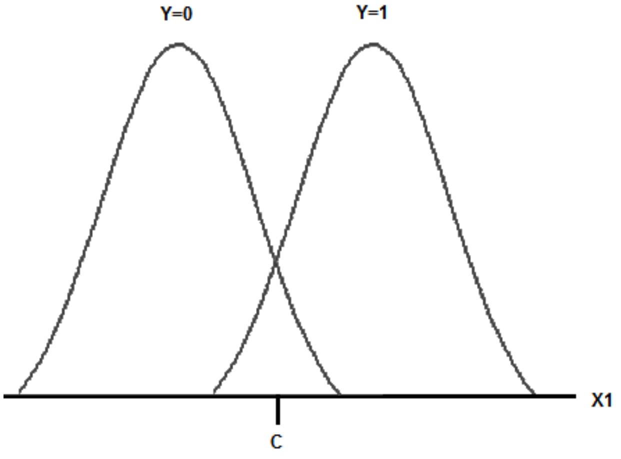
If your answer is in decimals, answer with precision 4, e.g. (6.051, 0.1230, 1.234e+7).

Fill in the blank:

5. [**4 pt**] 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

Select one:

○ 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.

6. [**4 pt**] 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?

   If your answer is in decimals, answer with precision 4, e.g. (6.051, 0.1230, 1.234e+7)

   Fill in the blank:

7. [**4 pt**] 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?

   Select all that apply:

   □ 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

8. [**3 pt**] Which of the following machine learning algorithms are probabilistic generative models?

   Select one:

   ○ Decision Tree

   ○ K-nearest neighbors

   ○ Perceptron

   ○ Naive Bayes

   ○ Logistic Regression

   ○ Feed-forward neural network

9. **[15 pt]** Logistic Regression and Naive Bayes.

When Y is Boolean and $\mathbf{X} = \langle X_1 ... X_n \rangle$ is a vector of continuous variables, then the assumptions of the Gaussian Naive Bayes classifier imply that $P(Y \mid \mathbf{X})$ is given by the logistic function with appropriate parameters W. In particular:

$$P(Y = 1 \mid \mathbf{X}) = \frac{1}{1 + \exp(b + \sum_{i=1}^{n} w_i X_i)}$$

and

$$P(Y = 0 \mid \mathbf{X}) = \frac{\exp(b + \sum_{i=1}^{n} w_i X_i)}{1 + \exp(b + \sum_{i=1}^{n} w_i X_i)}$$

Consider instead the case where Y is Boolean and $\mathbf{X} = \langle X_1 ... X_n \rangle$ is a vector of Boolean variables. Prove for this case also that $P(Y \mid \mathbf{X})$ follows this same form (and hence that Logistic Regression is also the discriminative counterpart to a Naive Bayes generative classifier over Boolean features).

Hints

(a) Simple notation will help. Since the X_i are Boolean variables, you need only one parameter to define $P(X_i \mid Y = y_k)$. Define $\phi_{i1} \equiv P(X_i = 1 \mid Y = 1)$, in which case $P(X_i = 0 \mid Y = 1) = (1 - \phi_{i1})$. Similarly, use ϕ_{i0} to denote $P(X_i = 1 | Y = 0)$.

(b) Notice with the above notation you can represent $P(X_i \mid Y = 1)$ as follows

$$P(X_i \mid Y = 1) = \phi_{i1}^{(X_i)} (1 - \phi_{i1})^{(1 - X_i)}$$

Note when $X_i = 1$ the second term is equal to 1 because its exponent is zero. Similarly, when $X_i = 0$ the first term is equal to 1 because its exponent is zero.

Write your solution on the following page.

Collaboration Questions Please answer the following:

After you have completed all other components of this assignment, report your answers to the collaboration policy questions detailed in the Academic Integrity Policies found here.

1. Did you receive any help whatsoever from anyone in solving this assignment? Is so, include full details.
2. Did you give any help whatsoever to anyone in solving this assignment? Is so, include full details.
3. Did you find or come across code that implements any part of this assignment ? If so, include full details.

Solution