CS 412 — Introduction to Machine Learning

Sample Midterm Exam

Department of Computer Science, University of Illinois at Chicago

Time allowed: 75 minutes

Name (print): ____________________________ UIN: _______________

Signature: ____________________________ **I am an undergraduate / graduate.**

(Please circle one)

The total score is **100** for undergraduate students, and **110** for graduate students. It will be multiplied by **8** to contribute to your comprehensive score.

You should now have six (6) sheets of paper.

Pages 1 to 8 contain questions, and pages 9-12 are scratch paper. If you need more scratch paper, just raise your hand. **You can detach pages 9-12 (last two sheets) for your convenience.**

There are seven (7) big questions. Graduate students are to solve all the questions. Q2(c) and Q7(b) are bonus questions for undergraduate students; you can earn an extra of 10 points at most.

Write your answers right below each question. DO NOT write answer on the scratch paper (pages 9-12); it will not be graded.

You are allowed to bring one sheet of letter-sized paper with anything written on both sides. However, you are not allowed to consult any books, notebooks, or friends. You may use a calculator. Fractional number can be used as final results, and you do not need to write out the ground values. But if you like, it's also fine to keep 3-4 significant digits.

At the end of the exam, return all the six (6) sheets (and any additional scratch paper if you requested).

IMPORTANT. **To enable partial grading, you may want to write out the detailed derivations unless the question explicitly requires "no explanation/derivation"**. You are likely to maximize your score if you: read the questions carefully; are concise and clear in your answers; write neatly.

1. [20 pt] True or false. (5 questions, 4 points each)

For each question: write TRUE or FALSE on your answer sheet (2 pt) and give a brief explanation or picture (2 pt). Your explanation must be consistent with the TRUE or FALSE that you selected.

(a) Suppose a questionnaire consists of three questions, and for each question, the answer can be Agree (1) or Disagree (0):

	Agree (1)	Disagree (0)
The lectures are great (X1)		
The workload is too high (X2)		
I like CS 412 (Y)		

It turns out that in all responses, if Y = 1, then X1 = 1 and X2 = 0. And if Y = 0, then X1 = 0 and X2 = 1. This means that X1 is NOT independent of X2 conditioned on Y.

TRUE or FALSE   Explanation:

FALSE. X1 is independent of X2 conditioned on Y. To see it, compute

P(X1=1, X2=1 | Y=1) = 0,   P(X1=1, X2=0 | Y=1) = 1,

P(X1=0, X2=1 | Y=1) = 0,   P(X1=0, X2=0 | Y=1) = 0,

So P(X1=1 | Y=1) = 1, P(X1=0 | Y=1) = 0, P(X2=1 | Y=1) = 0, P(X2=0 | Y=1) = 1.

This means

P(X1, X2 | Y=1) = P(X1 | Y=1) * P(X2 | Y=1)

for all the four combinations of the values of X1 and X2. Similar results hold for Y=0.

(b) The Naïve Bayes classifier that predicts the class Y given the features $X_1, \ldots, X_k$ uses the independence assumption that $X_i \perp X_j$ for all i and j.

TRUE or FALSE   Explanation:

FALSE. The Naïve Bayes classifier assumes $X_i \perp X_j$ | Y.

(c) Adam is developing an app that recommends good restaurants. He knows that users will walk away from the app if a recommended restaurant turns out not so good. That is, he really cannot afford recommending a not-so-good restaurant. As a result of this consideration, a high precision will be more important than a high recall.

TRUE or FALSE   Explanation:

TRUE. High precision means among those we predict as true, a high fraction is indeed true. In contrast, high recall means we do not miss true examples, i.e., if we predict something is false, then it is highly likely to be false.

(d) Bayes estimation takes the mode of the posterior parameter distribution to make predictions for new data.

TRUE or FALSE   Explanation:

False. Bayes estimation uses the expectation (mean) of the posterior parameter distribution.

(e) Suppose our dataset consists of 200 positive examples and 100 negative examples. Our classifier constantly outputs the majority class of the training set (breaking tie arbitrarily). Then the leave-one-out cross validation error on the dataset is 1/3.

TRUE or FALSE   Explanation:

True. In the leave-one-out setting, this classifier will always predict positive. So it's wrong 1/3 time.

Q2. [12 or 17 pt] The scanner A in an airport can detect true terrorists with 60% accuracy. It can detect upstanding citizens with 70% accuracy. There are 100 passengers on your plane, among which one is a terrorist. Now the shifty looking man sitting next to you tests positive.

(a) [6 pt] Let $T \in \{0, 1\}$ denote the variable regarding whether the person is a terrorist (0 for no and 1 for yes), and $S \in \{0, 1\}$ denote the outcome of the scanner A (0 for negative and 1 for positive). Then what do the above 60% and 70% mean in terms of T and S?

Solution:

$$p(S = 1|T = 1) = 0.6; \qquad p(S = 0|T = 0) = 0.7$$

(b) [6 pt] Use Bayes rule to compute the probability that the shifty looking man is a terrorist given that he tests positive.

Solution:

<u>Given</u>:

$p(S = 1|T = 1) = 0.6; \quad p(S = 0|T = 0) = 0.7$

$p(S = 0|T = 1) = 0.4; \quad p(S = 1|T = 0) = 0.3$

$p(T = 1) = \frac{1}{100} = 0.01; \; p(T = 0) = 0.99$

We want to compute:

$p(T = 1|S = 1) = ?$

We can use Bayes rule as:

$$p(T = 1|S = 1) = \frac{p(S = 1|T = 1) \times p(T = 1)}{p(S = 1)} \tag{*}$$

We can compute the marginal $p(S = 1)$ as

$$\begin{aligned} p(S=1) = \sum_{T} p(S=1,T) &= p(S=1,T=0) + p(S=1,T=1) \\ &= p(S=1|T=0) \times p(T=0) + p(S=1|T=1) \times p(T=1) \\ &= 0.3 \times 0.99 + 0.6 \times 0.01 \\ &= 0.303 \end{aligned}$$

Substituting back, we get

$$p(T = 1|S = 1) = \frac{0.6 \times 0.01}{0.303} = \mathbf{0.019}$$

(c) [**Bonus question for undergraduate students, 5 pt**] Suppose there is another scanner B which works independently of A. It can detect true terrorists with 60% accuracy, and can detect upstanding citizens with 50% accuracy. Now the shifty looking man tests negative on scanner B. Does this increase or decrease the probability that he is a terrorist? Explain why.

Solution:

Given:

Let $S' \in \{0,1\}$ be the random variable corresponding to scanner B.

$p(S' = 1|T = 1) = 0.6; p(S' = 0|T = 0) = 0.5$
$p(S' = 0|\mathrm{T} = 1) = 0.4;\ \mathrm{p}(S' = 1|T = 0) = 0.5$

We want to check: If the additional information given by S' increases or decreases the probability that the man is a terrorist.

$p(T = 1|S = 1, S' = 0)\ ?\ p(T = 1|S = 1)$

$$p(T = 1|S = 1, S' = 0) = \frac{p(S = 1, S' = 0|T = 1) \cdot p(T = 1)}{p(S = 1, S' = 0)}$$

$$= \frac{p(S = 1|T = 1) \cdot p(S' = 0|T = 1) \cdot p(T=1)}{p(S=1,S'=0)} \quad \text{(by conditional independence)}$$

So in view of (*), we only need to compare $\frac{p(S' = 0|T = 1)}{p(S=1,S'=0)}$ v.s. $\frac{1}{p(S=1)}$.

Obviously, $p(S' = 0|T = 1)\ p(S = 1)$ = 0.4*0.303= 0.1212 (where 0.303 is from question b).

$$p(S = 1, S' = 0) = p(T = 0)p(S = 1|T = 0)p(S' = 0|T = 0)$$
$$+p(T = 1)p(S = 1|T = 1)p(S' = 0|T = 1)$$
$$= 0.99*0.3*0.5 + \ldots > 0.1485 + \ldots > 0.1212.$$

So $\frac{p(S' = 0|T = 1)}{p(S=1,S'=0)} < \frac{1}{p(S=1)}$. Hence the probability of being a terrorist decreases.

Q3. [12 pt] Suppose we are trying to estimate the mean of a normal distribution $N(\mu, 1)$ based on samples $X_1, X_2, \dots, X_n$. Suppose $\mu \neq 0$. The maximum likelihood estimator (MLE) is $\frac{1}{n}\sum_{i=1}^{n} X_i$.

Now give an example for each of the following questions (4 points each):

(a) Propose an estimator that has higher bias but lower variance.

(b) Propose an estimator that has higher bias and higher variance.

(c) Suppose an estimator just outputs X_1. Does it have higher bias than MLE, or lower, or equal?

Solution:

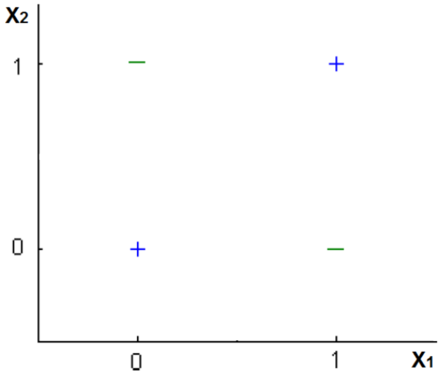
(a) Constantly output 0. Its variance is 0, but has a bias of $|\mu|$.

(b) Sample from a wild range, say [1000, 2000] uniformly at random.

(c) Equal bias. The expectation of X_1 is μ, and so is the expectation of $\frac{1}{n}\sum_{i=1}^{n} X_i$.

Q4. [12 pt] Naïve Bayes

Suppose we learn a Naïve Bayes classifier from the examples in the figure below using maximum likelihood estimator (MLE) as the training rule.



(a) [6 pt] Write down all the parameters and their estimated values. Note: Y, X_1, X_2 are all binary variables.

Solution:

P(Y = 1) = 0.5 = P(Y = 0).

P(X1 = 1 | Y = 0) = P(X1 = 1 | Y = 1) = 0.5 = P(X1 = 0 | Y = 0) = P(X1 = 0 | Y = 1).

P(X2 = 1 | Y = 0) = P(X2 = 1 | Y = 1) = 0.5 = P(X2 = 0 | Y = 0) = P(X2 = 0 | Y = 1).

This is a very poor classifier since for any X1, X2 it will predict P(Y = 1 | X1, X2) = P(Y = 0 | X1, X2) = 0.5. Naturally, it cannot perfectly classify the examples in the figure.

(b) [6 pt] What is the accuracy of this learned naïve Bayes model on the four examples?

Solution: It is not deterministic because P(Y = 1 | X1, X2) = P(Y = 0 | X1, X2) = 0.5. If we break tie uniformly at random, then the accuracy is 50% in expectation. But any number among {0, 25, 50, 75, 100} per cent is possible depending on the tie breaking.

Q5. [20 pt] Consider the geometric distribution, which has $p(X = k|\theta) = (1-\theta)^{k-1}\theta$. Assume in our training data, *X* took on the values 1, 2, and 3.

(a) [6 pt] Write an expression for the log-likelihood of the data as a function of the parameter *θ*.

Solution:

$$L(\theta) = (1-\theta)^0\theta \cdot (1-\theta)^1\theta \cdot (1-\theta)^2\theta$$

Therefore $$\log L(\theta) \ = 3\log(1-\theta) + 3\log\theta.$$

(b) [6 pt] A random variable *θ* ranging in [0, 1] has a Laplace distribution if $p(\theta) \propto e^{-a\theta}$ over $\theta \in [0,1]$, and $p(\theta) = 0$ for other values of *θ*. Here a is a positive parameter (like the α and β in Beta distribution). Then is the posterior distribution of *θ* also a Laplace distribution on [0, 1]?

Solution: No. The posterior distribution of *θ* is:

$$p(\theta|D) \propto e^{-a\theta}(1-\theta)^3\theta^3.$$

This is surely not proportional to $e^{-a\theta}$ (in *θ*), hence not a Laplace distribution.

(c) [8 pt] Suppose *a* = 1 in the prior distribution of $p(\theta)$. Then what is the MAP estimate of *θ*? If your answer boils down to finding a root of a quadratic function $x^2 + bx + c = 0$, it is fine not to solve it.

Solution: Take the log of the $p(\theta|D)$ above (up to a normalization constant), we get the objective function to maximize

$$f(x) = -\theta + 3\log(1-\theta) + 3\log\theta.$$

Take derivative of it and set to 0

$$0 = -1 - \frac{3}{1-\theta} + 3\frac{1}{\theta}$$

and this gives

$$\theta^2 - 7\theta + 3 = 0.$$

The root between 0 and 1 is 0.4586.

Q6. [14 pt] Suppose for each P(F=0|C), P(D=0|C), P(S=0|C), and P(G=0|C), we assign a Beta(2, 2) prior. Recall Beta(α, β) distribution has density function $f(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$. Its mean is $\frac{\alpha}{\alpha+\beta}$ and its mode is $\frac{\alpha-1}{\alpha+\beta-2}$.

Note that the question differs from the assignment, in that we now consider the probability of being 0 instead of 1.

Suppose we observed the following five data points (documents):

F	D	S	G	C
1	0	1	0	1
0	1	0	1	1
1	1	0	0	0
1	0	1	0	0
0	1	0	1	0

(a) [7 pt] What is the MAP estimate of P(D=0|C=0)?

Solution: Given C = 0, there are two cases where D = 1, and one case where D = 0. So the posterior probability of D=0 given C=1 is Beta(3, 4), and so the MAP estimate of P(D=0|C=0) = (3-1)/(3+4-2) = **2/5**.

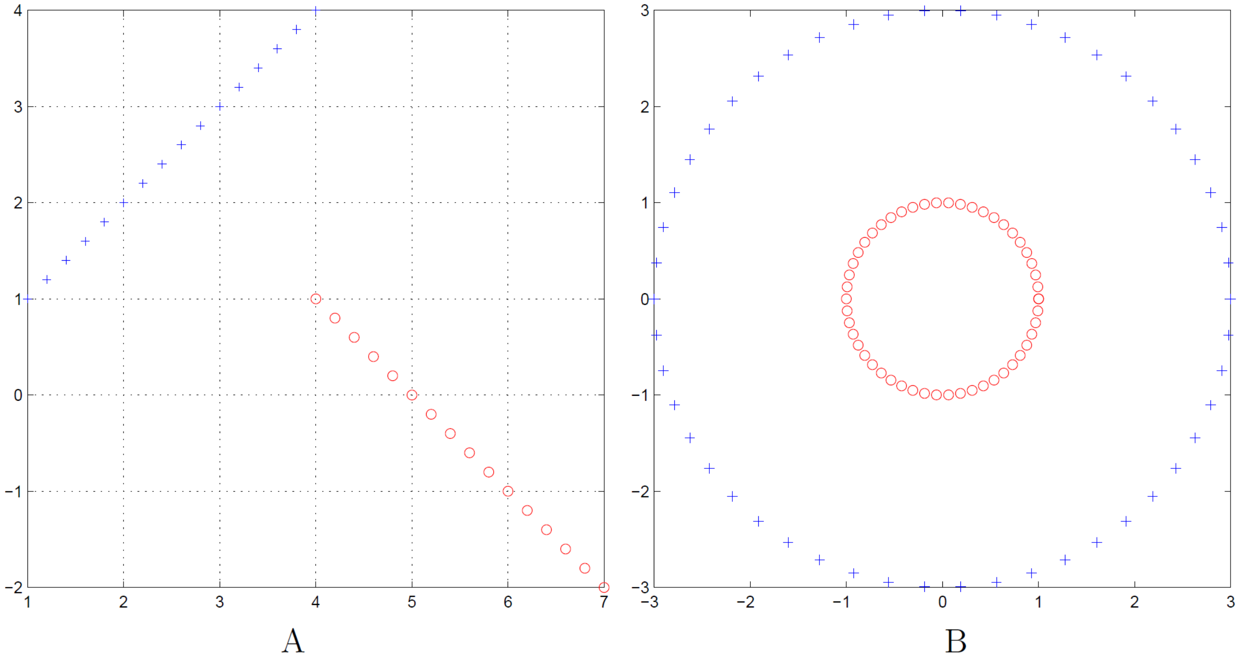
(b) [7 pt] Suppose we also assign a Beta(2, 2) prior on P(C=1). Using the data table in Q3, what is the probability that the next document will **not** be about sports (C=0)?

Solution: The posterior probability of C=1 is Beta(4, 5). So the probability of C = 1 is 4/9, and hence the probability of C = 0 is **5/9**.

Q7. [10 or 15 pt] Consider the two datasets shown in plots (A) and (B) below:

- In each of these datasets there are two classes, ’+’ and ’o’.
- Each class has the same number of points.
- Each data point has two real valued features, the X1 and X2 coordinates.

To answer the following sub-questions (a) and (b), you can write down some intermediate steps to earn **partial credits**. They are not necessary though, provided your final answer is right.



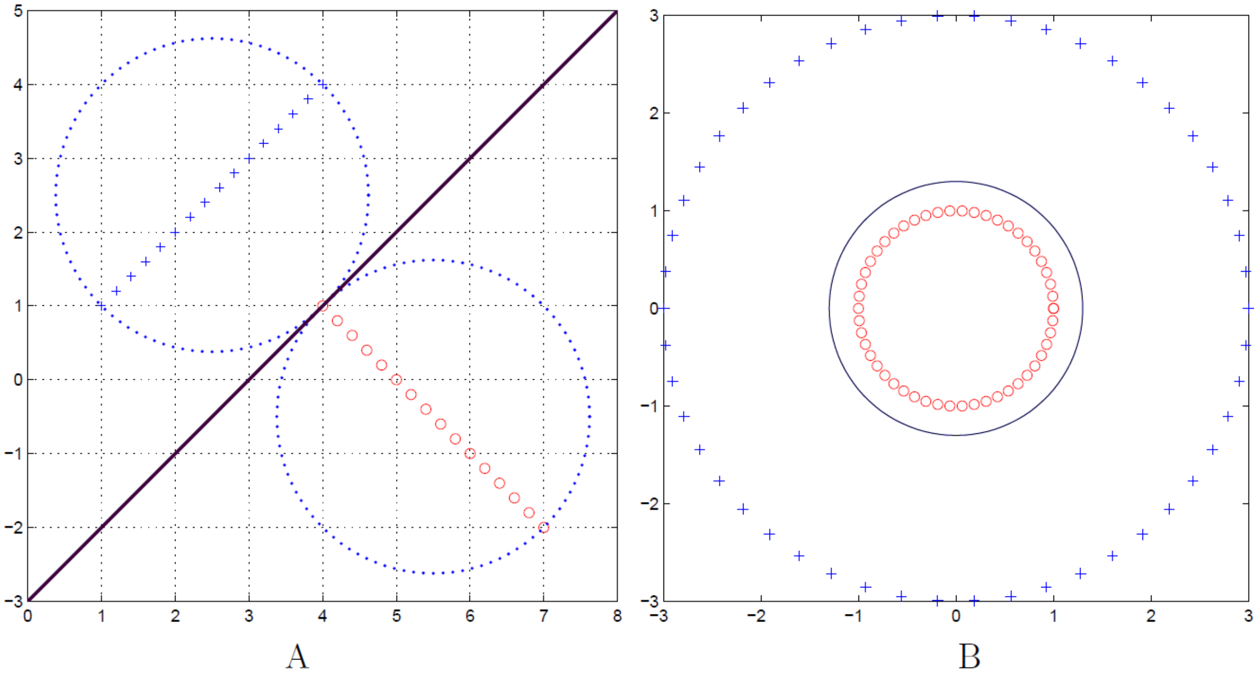
(a) [10 pt] For plot A, let us estimate a Gaussian Naive Bayes likelihood model for both + and o classes. In the left plot **below**, draw a contour of the density P(X1, X2 | Y= +) and of P(X1,X2 | Y = o). You can use any level for the contour as long as it is illustrative. Point to the two contours with an arrow labeled by “contour of +” and “contour of o”, respectively.

Next, do the same on plot B (right plot below). We only look for a qualitative illustration.

(b) [**Bonus question for undergraduate students, 5 pt**] For each of these two datasets, draw, in the plots **below**, the decision boundary that a Gaussian Naive Bayes classifier will learn.

Solution: For A, the crucial detail is that a Gaussian Naive Bayes model learns diagonal covariance matrices yielding axis aligned Gaussians. In the left figure below, the two circles are the Gaussians learned by GNB, and hence the decision surface is the tangent through the point of contact.

For B, a Gaussian Naive Bayes model learns two Gaussians, one for the circle inside with small variance, and one for the circle outside with a much larger variance, and the decision surface is roughly shown on the right. The original red circle (composed of little points) serves as a contour of the density P(X1, X2 | Y= o), and the blue circle (composed of little plus signs) serves as a contour of the density of P(X1,X2 | Y = +). Actually, any circle centered at the origin serves as a contour of both P(X1, X2 | Y= +) and of P(X1,X2 | Y = o).



A

B