University of Massachusetts Dartmouth
Department of Computer and Information Science
CIS 530: Advanced Data Mining (Spring 2022)

March 16, 2022

Printed Full Name: ____

Student ID: ____

DO NOT TURN THE PAGE OVER UNTIL YOU ARE INSTRUCTED TO DO SO

Please read the following instructions:

1. **You have 3 hours to complete the part of this examination.**
2. **This examination is OPEN BOOK.**
3. **Type your answer in space provided on the examination sheets, any work not on the examination sheets will not be graded.**
4. **You may use the empty sheet and the back of answer sheet as scratch paper.**
5. **Type your answers legibly.**
6. **Submit your answer by the end of the examination.**
7. **DO NOT communicate any of your classmates during the examination.**

Honor Policy: copying in whole or in part of the examination will be considered to be an act of scholastic dishonesty. Students who violate university rules on scholastic dishonesty are subject to disciplinary penalties, including the possibility of failure in the course and/or dismissal from the university. Since such dishonesty harms individuals, all students, and the integrity of the university, policies on scholastic dishonesty will be strictly enforced.

I have read the above instructions and I will act in accordance with all of them.

______________________________ ______________________________

Student Signature **Date**

Type your name and date to agree the policy before you start!

Name: ________________________ **UMass ID:**____________________

Part 1: Multiple Choice (24 pts total, 3 pts each. Note: only one option is correct.)

1. For a $m \times n$ $(m \neq n)$ matrix A, which of the following statement is true:

   (A) A is a singular matrix

   (B) A must be invertible

   (C) If A is full rank, then its determinant > 0

   (D) The maximum rank of A is minimum (m, n)

2. Which of following matrix operations is correct?

   (A) $(AB)^T = A^T B^T$

   (B) $(AB)^{-1} = A^{-1}B^{-1}$

   (C) $[1,1]^T \times [1,1] = 2$

   (D) $\det\begin{bmatrix}1 & 2\\ 3 & 5\end{bmatrix} = -1$

3. Which of the following is NOT a numeric attribute in Data Mining?

   (A) Nominal

   (B) Discrete

   (C) Continuous

   (D) Binary

4. Please select the correct claim about TF-IDF feature:

   (A) A more common word w in different documents will give a smaller $\mathrm{DF}(w)$

   (B) A more frequent word w within a document D will give a smaller $\mathrm{TF}(w)$

   (C) If w is unique to document D, then IDF(w) is minimized

   (D) TF-IDF will help remove stop words.

5. Which of the following about probability is Wrong?

   (A) Maximum likelihood estimation of μ of a Gaussian is unbiased

   (B) Maximum likelihood estimation of σ^2 of a Gaussian is biased

   (C) Posterior = Prior × Likelihood function

   (D) If we toss a coin three times and get three heads, from frequentists' view, $\Pr(x = \text{head}) = 1$

6. Which of the followings is **NOT** a reason for overfitting?

   (A) Insufficient training data

   (B) Noisy training data

   (C) Model is over complex

   (D) The test dataset only contains data from one class

7. A larger Node Impurity of a node in a decision tree means:

   (A) Smaller Entropy

   (B) Smaller Gini function value

   (C) Larger Classification error

   (D) Larger Information gain

8. Which of the followings about similarity and distance metric is Wrong?

   (A) Often falls in range [0, 1]

   (B) Sim(p, q) =0, if and only if p$\neq q$

   (C) L_1 distance is also called Manhattan distance

   (D) Editing distance (basic one, with insert and delete only) between **x = abcde** and **y = bcduve** is 3

Part III: True&False (20 pts total, 2 pts each)

1. Outer product of two 3D vectors will yield a 3×3 matrix. (T)
2. All square matrices are invertible. (F)
3. With more training data, models tend to provide low bias and high variance. (F)
4. In S-fold cross-validation, if S=N where N is the number of samples in training dataset, then it is also called “leave-one-out”. (T)
5. Hamming distance is the same as L_1 distance for two binary vectors. (T)
6. Uniform distribution will provide the smallest entropy for discrete variables. (F)
7. There might be more than one decision trees that fit the same data. (T)
8. When building a decision tree, more nodes than necessary will lead to overfitting and a lower training error. (F)
9. Decision boundary of decision tree is usually NOT parallel to the coordinate axis. (F)
10. “Recall” has the same meaning as “False Positive Rate”. (F)

Part III: Completion (16 pts total)

1. We have 100 people. 51 are women, namely, $p(W) = 0.51$, and 49 are men, $p(M) = 0.49$. If we select **three persons one by one,** what is the probability of $p(W, W, W)$ that all are women?

   (A) Sampling with replacement: ______________________

   (B) Sampling without replacement: ______________________

Sampling with replacement:
$p(W, W, W) = p(W) \times p(W) \times p(W) = 0.51 \times 0.51 \times 0.51 \approx 0.1326$

Sampling without replacement:
$p(W, W, W) = p(W) \times p(W|W1) \times p(W|W1,W2) = 0.51 \times (50/99) \times (49/98) \approx 0.1244$

2. Assume integers are coming in a stream and you do not know the size of stream (N) in advance. You want to sample each of them uniformly at random. What is your sampling strategy? Assume the index of integers is n $(n \leq N)$.

One possible sampling strategy for this scenario is called "Reservoir Sampling". The idea is to keep a reservoir (or a sample) of a fixed size k that will be updated as the stream of integers is processed. Here are the steps of the algorithm:

1. Initialize an empty reservoir of size k.
2. For each incoming integer with index n:
    * If n ≤ k, add the integer to the reservoir.
    * If n > k, generate a random number between 1 and n inclusive.
        * If the random number is less than or equal to k, replace a random integer in the reservoir with the incoming integer.

At the end of the stream, the reservoir will contain k integers that are uniformly sampled from the stream.
The intuition behind this algorithm is that the first k integers are added to the reservoir with probability 1, as they are the only candidates. For integers with index n > k, they have a probability of k/n of being selected, as we replace a random integer in the reservoir with probability k/n. This ensures that each integer in the stream has an equal probability of being selected for the reservoir.

3. If you toss a fair coin ten times, what is the probability of **at least** one head? Please also provide the complete deduction as well as probability.

The probability of getting at least one head in ten tosses of a fair coin can be found using the complement rule. That is, we first find the probability of getting no heads (all tails), and then subtract this from 1 to get the probability of getting at least one head.

The probability of getting tails on one toss of a fair coin is 1/2. Therefore, the probability of getting tails on all ten tosses is (1/2)^10 = 1/1024.

To find the probability of getting at least one head, we subtract the probability of getting no heads from 1:

P(at least one head) = 1 - P(no heads) = 1 - (1/2)^10 = 1 - 1/1024 = 1023/1024 ≈ 0.999

Therefore, the probability of getting at least one head in ten tosses of a fair coin is approximately 0.999, or 99.9%.

4. The training error of a decision tree with ONE leaf node is 10/40 and 40 is the total number of training samples. What is the **pessimistic error** of this decision tree: _____________? After the splitting of the leaf node, there are four leaf nodes, and the training error becomes 9/40. Then what is the **pessimistic error** of the new decision tree: _______________? Should you prune the sub-tree after the splitting (yes/no): ___NO_____?

Blank 1:
Pessimistic Error = 10/40 + 2*0.137
= 0.284

Blank 2:
Pessimistic Error = 9/40 + 2*0.135
= 0.279

Part IV: Free Response (40 pts total)

1. (10 pts) Please prove that $\text{trace}(A^T A) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_r^2$, where σ_i is the singular value of matrix A. Hint: (1) apply SVD to A; (2) use $\text{trace}(B^{-1}AB) = \text{trace}(A)$ if B is invertible; (3) definition of trace operation.

Given a matrix A, we can perform Singular Value Decomposition (SVD) on it as follows:

$A = U\Sigma V^T$

where U and V are orthogonal matrices, and Σ is a diagonal matrix containing the singular values $\sigma_1, \sigma_2, \ldots, \sigma_r$ of A. Here, r is the rank of A.

Using this SVD, we can express the matrix A^TA as follows:

$A^TA = (U\Sigma V^T)^T(U\Sigma V^T) = V\Sigma^TU^TU\Sigma V^T = V\Sigma^2V^T$

where we have used the fact that $U^TU = I$ since U is an orthogonal matrix, and $\Sigma^T = \Sigma$ since Σ is a diagonal matrix.

Now, using the definition of the trace operation, we have:

$trace(A^TA) = trace(V\Sigma^2V^T) = trace(\Sigma^2V^TV) = trace(\Sigma^2)$

since V is an orthogonal matrix, and hence $V^TV = I$.

Finally, using the definition of singular values, we have:

$\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_r^2 = ||A||_F^2$

where $||A||_F$ is the Frobenius norm of A. Therefore, we can write:

$trace(A^TA) = trace(\Sigma^2) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_r^2 = ||A||_F^2$

Thus, we have proven that $trace(A^TA) = ||A||_F^2$, which is equivalent to $trace(A!A) = \sigma_1 + \sigma_2 + \cdots + \sigma_r$.

2. (10 pts) Assume you have 3 different classes when modeling decision tree.

   (A) At node **a**, the data distribution for three different classes are (1/2,1/4,1/4). Please compute the: (1) entropy; (2) Gini; (3) Classification error.

```
For the data distribution (1/2,1/4,1/4), we have:

$$H(a) = -\frac{1}{2} \log_2 \frac{1}{2} -\frac{1}{4} \log_2 \frac{1}{4} -\frac{1}{4} \log_2 \frac{1}{4}$$
$$H(a) \approx 1.5$$

So, the entropy of the node is approximately 1.5.

Gini:

For the data distribution (1/2,1/4,1/4), we have:

$$G(a) = \frac{1}{2} (1-\frac{1}{2}) + \frac{1}{4} (1-\frac{1}{4}) + \frac{1}{4} (1-\frac{1}{4})$$
$$G(a) = \frac{3}{8}$$

So, the Gini impurity of the node is 3/8.

Classification error:

For the data distribution (1/2,1/4,1/4), we have:

$$E(a) = 1 - \frac{1}{2}$$
$$E(a) = \frac{1}{2}$$

So, the classification error of the node is 1/2.
```

   (B) We split the node **a** into nodes **b** and **c**, and the data distribution for different classes in **b** is (1/3, 1/3, 1/3) and in **c** is (3/5, 1/5, 1/5). In addition, we know that the data ratio between nodes **b** and **c** is 3:5. Please compute the **Information Gain** after splitting node **a** into nodes **b** and **c**.

```
Impurity measure of node a:
We have already calculated the entropy and Gini impurity of node a in part (A):

Entropy: $H(a) \approx 1.5$
Gini impurity: $G(a) = \frac{3}{8}$
Impurity measures of child nodes b and c:
For node b, the data distribution is (1/3, 1/3, 1/3), so:

Entropy: $H(b) = -\frac{1}{3}\log_2(\frac{1}{3}) -\frac{1}{3}\log_2(\frac{1}{3}) -\frac{1}{3}\log_2(\frac{1}{3}) = 1.585$
Gini impurity: $G(b) = \frac{1}{3}(1-\frac{1}{3}) + \frac{1}{3}(1-\frac{1}{3}) + \frac{1}{3}(1-\frac{1}{3}) = \frac{2}{3}$
For node c, the data distribution is (3/5, 1/5, 1/5), so:

Entropy: $H(c) = -\frac{3}{5}\log_2(\frac{3}{5}) -\frac{1}{5}\log_2(\frac{1}{5}) -\frac{1}{5}\log_2(\frac{1}{5}) \approx
0.972$
Gini impurity: $G(c) = \frac{3}{5}(1-\frac{3}{5}) + \frac{1}{5}(1-\frac{1}{5}) + \frac{1}{5}(1-\frac{1}{5}) = \frac{12}
{25}$

Information gain:
The number of instances in node b is 3/8 of the total instances in node a, and the number of instances in node c is 5/8 of
the total instances in node a.

Information gain using entropy:
IG_{\text{entropy}}(a,b,c) &= H(a) - (\frac{3}{8}H(b) + \frac{5}{8}H(c)) \\
&= 1.5 - (\frac{3}{8} \times 1.585 + \frac{5}{8} \times 0.972) \\
&\approx 0.094
\end{aligned}$$
- Information gain using Gini impurity:
$$\begin{aligned}
IG_{\text{Gini}}(a,b,c) &= G(a) - (\frac{3}{8}G(b) + \frac{5}{8}G(c)) \\
&= \frac{3}{8} - (\frac{3}{8} \times \frac{2}{3} + \frac{5}{8} \times \frac{12}{25}) \\
&\approx 0.034
\end{aligned}$$
So, the information gain for splitting node a into nodes b and c is approximately 0.094 using
```

3. (10 pts) Recall the fruit and buckets problem. The probability for blue bucket is $p(B = b)$ = 0.6, and that for red bucket $p(B = r)$ = 0.4. We have 2 apples 6 oranges in blue buckets, and 3 apples 1 orange in red buckets. Could you compute the following probability, $p(B = r|F = a)$?

To compute the probability $p(B = r|F = a)$, we need to use Bayes' theorem, which relates the conditional probability of an event given some observed evidence to the prior probability of the event and the likelihood of the evidence given the event and its complement.

$p(B = r|F = a) = p(F = a|B = r)\ p(B = r)\ /\ p(F = a)$

To calculate this probability, we need to compute three quantities:

$p(F = a|B = r)$, the probability of observing an apple given that the fruit is in the red bucket. There are 3 apples and 1 orange in the red bucket, so the probability of observing an apple given that the fruit is in the red bucket is:
$p(F = a|B = r)$ = 3 / 4 = 0.75

$p(B = r)$, the prior probability of the red bucket. We are given that $p(B = r)$ = 0.4.
$p(F = a)$, the marginal probability of observing an apple. This can be calculated using the law of total probability, which states that the probability of observing an event is the sum of the probabilities of that event given each possible value of the conditioning variable:
$p(F = a) = p(F = a|B = b)\ p(B = b) + p(F = a|B = r)\ p(B = r)$

We can calculate $p(F = a|B = b)$ using the same reasoning as in step 1:

$p(F = a|B = b)$ = 2 / 8 = 0.25

Therefore, $p(F = a)$ = (0.25 * 0.6) + (0.75 * 0.4) = 0.45

Putting all the pieces together, we get:

$p(B = r|F = a)$ = (0.75 * 0.4) / 0.45 = 0.667

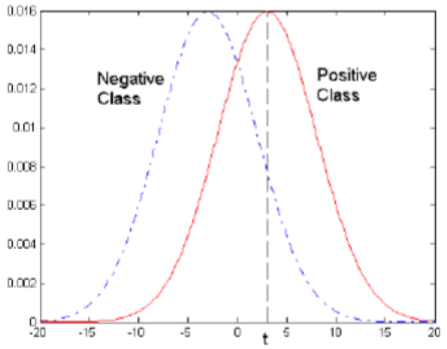
So the probability of the fruit being in the red bucket given that it is an apple is approximately 0.667 or 2/3.

4. (10 pts) Given M and N independent and identically distributed (i.i.d.) observations: $(x_{11}, x_{12}, \dots x_{1M}) \in \mathbb{R}^{1\times M}$ and $(x_{21}, x_{22}, \dots x_{2N}) \in \mathbb{R}^{1\times N}$ that follow two Gaussian distributions $N_1(\mu_1, \sigma_1^2)$ and $N_2(\mu_2, \sigma_2^2)$, respectively, please answer the following questions:

(A) Maximum likelihood solution to $N_1(\mu_1, \sigma_1^2)$ and $N_2(\mu_2, \sigma_2^2)$.

```
L(μ,σ^2)=∏i=1N 1/(√2πσ^2)exp(−(xi−μ)2/2σ^2)
Taking the natural logarithm of the likelihood function and simplifying, we obtain:
lnL(μ,σ^2)=−N/2 ln(2πσ^2)−∑i=1N(xi−μ)2/2σ^2
To obtain the maximum likelihood estimates of μ and σ^2, we need to differentiate lnL with respect to these parameters and
set the resulting expressions to zero. This yields the following equations:
μ̂ = (1/N) ∗ ∑(xi)
σ̂^2 = (1/N) ∗ ∑(xi − μ̂)^2
where μ̂ and σ̂^2 are the maximum likelihood estimates of μ and σ^2, respectively.
```

(B) Assume the N_1 and N_2 represent the negative and positive classes shown on the right, respectively. Please shadow the area for "false positive". The threshold has been given by t.



(C) Given the **confusion and cost matrix** below, please compute the **cost** of this classification problem:

Cost Matrix	PREDICTED CLASS	
ACTUALCLASS	+	-
+	-1	100
-	1	0

Confusion Matrix	PREDICTED CLASS	
ACTUALCLASS	+	-
+	150	140
-	160	250