CS 412: Fall'23
Introduction To Data Mining

Take-Home Midterm Exam

(Due Wednesday, October 18, 06:00 pm)

- The midterm is due at Wednesday, October 18, 6 pm. We will be using Gradescope for the midterm assignments. Please make sure to tag the appropriate pages for each question on Gradescope. Please do NOT email a copy of your solution. Contact the TAs if you are having technical difficulties in submitting the assignment. We will NOT accept late submissions!
- Please use Slack first if you have questions about the midterm. You can also come to our (zoom) office hours and/or send us e-mails.
- You will have to answer the questions yourself, you cannot consult with other students in class. It is an open book exam, so you can use the textbook and the material shared in class, e.g., slides, lectures, etc.
- You are expected to typeset the solutions. If your solution has any handwritten components, e.g., equations, tables, figures, etc., please make sure they are legible—otherwise you may not get credit.
- For each question, you will NOT get full credit if you only give out a final result. Please show the necessary steps and details, including any formulae you use, what the variables mean in the formulae, any derivation or calculation steps, and explanations as appropriate. Consider the following two examples:

Example 1 **Q:** Given a dataset $\mathcal{X} = \{3.1, 4.2, -1\}$, compute the mean.
A: For any set of n numbers $\mathcal{X} = \{x_1, \ldots, x_n\}$, the mean can be computed as $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$. For the given dataset $\mathcal{X}$, the mean is $\mu = \frac{3.1+4.2-1}{3} = 2.1$

Example 2 **Q:** A coin claimed to be unbiased has been tossed 100 times, with 54 heads and 46 tails. What is the χ^2 statistic?
A: For a categorical variable taking k possible values, if the expected values are $e_i, i = 1, \ldots, k$ and the observed values are $o_i, i = 1, \ldots, k$, then the χ^2 statistic can be computed as: $S = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{o_i}$. For the problem, since the coin is claimed to be unbiased, the expected values are $50, 50$. Further, the observed values are $54, 46$. Then, the chi-squared statistic is given by $S = \frac{(54-50)^2}{50} + \frac{(46-50)^2}{50} = 0.64$.

1. (26 points) Ms. Jones, the fifth-grade teacher at Heartland Elementary School observed that students' grades in her class are influenced by playing sports and music. Moreover, the teacher observed that student grades also influence participation in voluntary work in the community and signing up for advanced classes.

   Based on observations and data collection, the teacher constructed the following Bayes network illustrated in Figure1.

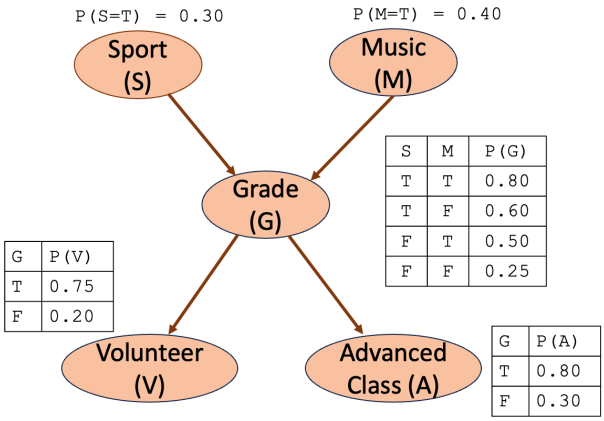

Figure 1: Bayesian network representing features influencing students' grades, volunteering activity, and signing up for advanced classes.

Answer the following questions (*note: show formulas and all intermediate steps*):

(a) (6 points) Compute $P(S = T, M = F, G = T, V = T, A = F)$

(b) (6 points) Compute $P(G = T|S = T)$

(c) (6 points) Compute $P(G = T|S = F)$

(d) (8 points) Suppose you are given a fully observed training set $\mathcal{D} = \{x^1, \ldots, x^n\}$ where $x^i = \langle s^i, m^i, g^i, v^i, a^i \rangle$ is the `i-th` example in the training data (s, m, g, v, and a stand for sport, music, grade, volunteer and advanced class respectively.) Derive the log-likelihood formula that captures the likelihood of observing the training set $\mathcal{D}$ under the given network in Figure 1.

Hint: expand $\log P(\mathcal{D}|\mathcal{T}, \theta_{\mathcal{T}})$ *(where* $\mathcal{D}$ *is the training set,* $\mathcal{T}$ *represents the features and* $\theta_{\mathcal{T}}$ *is the observed value of* $\mathcal{T}$*) and simplify.*

2. (28 points) A local bank has developed two models for credit card fraud detection named $M1$ and $M2$ respectively. The models were evaluated using a dataset of $n = 12000$ where $n = a + b + c + d$ credit card transaction. The following confusion matrix summarizes the results:

		Prediction	
		Fraud	Not Fraud
Truth	Fraud	a	b
	Not Fraud	c	d

The two models had these specific values for the confusion matrix:

(M1) $a = 300, b = 20, c = 10, d = 11670$

(M2) $a = 320, b = 1, c = 2, d = 11677$

Clearly define the following quantities in terms of a, b, c, d in the confusion matrix and compute their numerical values (rounded to 3 places after the decimal) for M1 and M2.

(*Note: show formulas and all intermediate steps*)

(a) (4 points) Sensitivity.

(b) (4 points) Specificity.

(c) (5 points) Accuracy.

(d) (5 points) Precision.

(e) (5 points) Recall.

(f) (5 points) F1 score.

3. (22 points) We consider comparing the performance of two classification algorithms A_1 and B_1 based on k-fold cross-validation. The comparison will be based on a t-test to assess statistical significance with significance level $\alpha = 5\%$. The null hypothesis is that the mean accuracy of the two algorithms A_1 and B_1 are exactly the same.

   (a) (4 points) We will assess the results for $k = 10$-fold cross-validation. What should be the degrees of freedom for the test? Briefly explain your answer.

   (b) (10 points) The accuracies for $k = 10$-fold cross-validation from two algorithms A_1 and B_1 are given in Table 1.

	1	2	3	4	5	6	7	8	9	10
A_1	0.908	0.962	0.878	0.956	0.939	0.955	0.944	0.933	0.881	0.949
B_1	0.449	0.585	0.381	0.433	0.475	0.430	0.520	0.590	0.565	0.443

Table 1: Accuracies on 10-folds for Algorithms A_1 and B_1.

   Is the performance of one of the two algorithms significantly different than the other based a t-test at significance level $\alpha = 5\%$? Clearly explain your answer by showing details of (a) the computation of the t-statistic, and (b) the computation of the p-value. Given the t-statistic `t_stat` and degrees of freedom `df`, you should be able to compute the p-value using the following:[1]

```
from scipy.stats import t
p_val = (1-t.cdf(abs(t_stat), df)) * 2
```

   (c) (8 points) Suppose we have a better version of algorithm B_1 called B_2. The accuracies for $k = 10$-fold cross-validation from algorithms A_1 and B_2 are given in Table 2.

	1	2	3	4	5	6	7	8	9	10
A_1	0.908	0.962	0.878	0.956	0.939	0.955	0.944	0.933	0.881	0.949
B_2	0.968	1.000	0.950	0.994	0.989	0.989	1.000	0.994	0.966	0.966

Table 2: Accuracies on 10-folds for Algorithms A_1 and B_2.

   Is one of the two algorithms significantly better than the other based a t-test at significance level $\alpha = 5\%$? Clearly explain your answer by showing details of (a) computation of the t-statistic, and (b) computation of the p-value.

[1] Alternatively, you can look a table for p-values for t-statistic, similar to how you had done it for the χ^2-statistic earlier in the semester.

4. (24 points) Let $\mathcal{Z} = \{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^n, y^n)\}, \mathbf{x}^i \in \mathbb{R}^d, y^i \in \{0, 1\}, i = 1, \ldots, n$ be a set of n training samples. The input $\mathbf{x}^i, i = 1, \ldots, n$ are d-dimensional features, and x^i_j denotes the j-th feature of the i-th data point $\mathbf{x}^i$. The output $y^i \in \{0, 1\}, i = 1, \ldots, n$, are the class labels. Let $\hat{y}^i = f_\theta(\mathbf{x}^i)$ be the prediction on $\mathbf{x}$ from a predictive model $f_\theta(\mathbf{x})$, and let $\ell(y^i, \hat{y}^i)$ be a suitable loss comparing the true label to the prediction.

   (a) (8 points) Assume that $f_\theta(\mathbf{x}) = \sigma(\theta^\top \mathbf{x})$, where σ denotes the sigmoid function and $\ell(y, \hat{y})$ is the 2-class cross-entropy loss. What is the empirical loss $L(\theta)$ on the training set which is to be minimized? Please write the loss expression explicitly in terms of θ, $\mathbf{x}^i$ and y^i ($i \in \{1, 2, ..., n\}$), and standard functions like $\log, \exp$, etc.

   (b) (8 points) With the same assumption as (a), show that the loss expression in (a) above can be derived from a maximum log likelihood estimation problem based on a conditional Bernoulli model for the true labels y^i given the input $\mathbf{x}^i$.

   (c) (4 points) Since the prediction $\hat{y}^i = \sigma(\theta^\top \mathbf{x}^i) \in [0, 1]$ tries to model $P(1|\mathbf{x}^i)$, we can view the prediction as a Bernoulli distribution $\text{Bern}(\hat{p}^i) := [\hat{y}^i \quad (1 - \hat{y}^i)]$,[2] so that $\text{Bern}(\hat{p}^i)(1) = \hat{y}^i$ and $\text{Bern}(\hat{p}^i)(0) = 1 - \hat{y}^i$. While the true labels $y^i \in \{0, 1\}$ are deterministic, we can also view it as a Bernoulli distribution $\text{Bern}(p^i) = [\mathbb{1}[y^i = 1] \quad (1 - \mathbb{1}[y^i = 1])]$, where $\mathbb{1}$ denote the indicator function. Let $KL(\text{Bern}(p), \text{Bern}(\hat{p}))$ denote the KL-divergence between the Bernoulli distributions $\text{Bern}(p)$ and $\text{Bern}(\hat{p})$ and $\ell(y^i, \hat{y}^i)$ be the 2-class cross-entropy loss, which you have already used in (a) above. What is $KL(\text{Bern}(p^i), \text{Bern}(\hat{p}^i)) - \ell(y^i, \hat{y}^i)$? You can use the fact that the true labels $y^i \in \{0, 1\}$.

   (d) (4 points) Consider the special case when $x^i \in \mathbb{R}$, i.e., the features are one dimensional, and the dataset $(x^i, y^i), i = 1, \ldots, n$ is linearly separable, i.e., $\exists \theta^*$ such that $y^i = \frac{1}{2}(1 + \text{sign}(\theta^* x^i))$. Professor Astral claims that without any additional assumptions the solution θ^* to the optimization problem in (a) above must be unbounded, i.e., $|\theta^*| \to \infty$, where $|\cdot|$ denotes the absolute value. Do you agree with Professor Astral? Clearly justify your answer. [3]

[2] For a Bernoulli distribution $\text{Bern}(p) = [a \quad (1 - a)]$ for $a \in [0, 1]$, we mean the probability of getting a '1' is a, denoted as $\text{Bern}(p)[1] = a$, and the probability of getting a '0' is $(1 - a)$, denoted as $\text{Bern}(p)[0] = 1 - a$.

[3] If you agree, you will have to provide a technical argument why; if you disagree, a non-trivial counter-example will be sufficient. By non-trivial, we mean—you cannot assume all $\mathbf{x}^i$ are the same, etc.