STA 9890 - Statistical Learning for Data Mining

In-Class Test 3

This is a closed-note, closed-book exam.
You may not use any external resources other than a (non-phone) calculator.

Name: _____

Instructions

This exam will be graded out of **100 points**.

This exam is divided into three sections:

- Multiple Choice (30 points; 10 questions at three points each)
- Mathematics of Machine Learning (20 points)
- Practice of Machine Learning *Supervised* (25 points)
- Practice of Machine Learning *Unsupervised* (25 points)

You have one hour to complete this exam from the time the instructor says to begin. The instructor will give time warnings at: 30 minutes, 15 minutes, 5 minutes, and 1 minute.

When the instructor announces the end of the exam, you must stop **immediately**. Continuing to work past the time limit may be considered an academic integrity violation.

Write your name on the line above *now* before the exam begins.

Each question has a dedicated answer space. Place all answers in the relevant spot. Answers that are not clearly marked in the correct location **will not** receive full credit. Partial credit may be given at the instructor's discretion.

Mark and write your answers clearly: if I cannot easily identify and read your intended answer, you may not get credit for it.

Additional pages for scratch work are included at the end of the exam packet.

This is a closed-note, closed-book exam.

You may not use any external resources other than a (non-phone) calculator.

Multiple Choice: 30 points total at 3 points each

For each question, **CIRCLE** or **CHECK** your answer(s) as appropriate.

MC1. True/False: PCA cannot be applied without centering the data first.
☐ True ☐ False

MC2. True/False: The singular values of a (real-valued) matrix are always positive.
☐ True ☐ False

MC3. True/False: Estimated principal components are orthogonal to the data.
☐ True ☐ False

MC4. True/False: The sample covariance matrix of a data set is always *strictly positive definite* (all eigenvalues strictly greater than zero).
☐ True ☐ False

MC5. Select One: Which of the following relates the singular values of $\boldsymbol{X}$ (σ_i) and the eigenvalues of $\boldsymbol{X}^\top \boldsymbol{X}$ (λ_i) for all $i = 1, \ldots, \mathsf{Rank}(\boldsymbol{X})$?
☐ $\sigma_i = \lambda_i$ ☐ $\sigma_i^2 = \lambda_i$ ☐ $\sigma_i = \lambda_i^2$ ☐ $\sigma_i \geq \lambda_i$ ☐ $\sigma_i \leq \lambda_i$

MC6. Select One: Which of the following **is not** a common rule for selecting the PCA rank?
☐ Elbow Rule 80% Rule ☐ Boulder Statistic ☐ Hypothesis Testing

MC7. Select One: Which of the following pairs of vectors are orthogonal?
☐ Loading Vector i and Loading Vector $i + 2$ ☐ Score Vector j and Score Vector $j - 1$ ☐ Loading Vector k and Score Vector $k + 1$

MC8. Select *All* That Apply: Which of the following are properties of (generic/standard) PCA decomposition?
☐ Additive ☐ Ordered ☐ Orthogonal ☐ Regularized ☐ Sparse ☐ Non-Negative ☐ Nested ☐ Easily-Computed to Global Optimality ☐ Convex

MC9. Select *All* That Apply: Which of the following are properties of Non-Negative Matrix Factorization?
☐ Additive ☐ Ordered ☐ Orthogonal ☐ Regularized ☐ Sparse ☐ Non-Negative ☐ Nested ☐ Easily-Computed to Global Optimality ☐ Convex

(If a property holds for some NMF variants, but not others, check it anyways. I will allow answers for that hold for any NMF formulation I know.)

MC10. Select *All* That Apply: Which of the following are properties of Sparse PCA?
☐ Additive ☐ Ordered ☐ Orthogonal ☐ Regularized ☐ Sparse ☐ Non-Negative ☐ Nested ☐ Easily-Computed to Global Optimality ☐ Convex

(If a property holds for some Sparse PCA variants, but not others, check it anyways. I will allow answers for that hold for any Sparse PCA formulation I know.)

Mathematics of Machine Learning: 20 points total

In this problem, you will derive a K-Means-type clustering algorithm for *bivariate Poisson* data. Specifically, you will use the EM-Framework to cluster data that is generated from the following statistical model:

> For each data point, $\boldsymbol{x} \in \mathbb{N}^2$, both coordinates $(x^{(1)}, x^{(2)})$ are sampled IID from a Poisson distribution with (common) mean λ. The value of λ depends on the underlying (but unknown) class: formally,

$$\boldsymbol{x}_i | i \in \text{Cluster-}j \overset{\text{IID}}{\sim} \text{Poisson}^{\otimes 2}(\lambda_j)$$

You can think of this is a model for clustering genetic data where i indexes patients, j indexes disease subtypes, and each patient has their gene expression values processed by two different IID labs for reliability.

Recall that the Poisson distribution with parameter λ has a PMF given by:

$$\mathbb{P}(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

MML1. M-Step: To build the M-Step of our clustering algorithm, we need to determine the *maximum likelihood estimator* (MLE) of the Poisson parameters λ_j. Recall that, within a cluster, our data are IID so we will begin by deriving the MLE for IID Poisson samples. (10 points total)

M(a) Suppose a single point z is observed from a Poisson distribution with unknown parameter θ. What is the *log-likelihood* associated with this observation? (2 points)

Recall: The likelihood is a function of the parameter, θ, holding the data z constant.

M(b) Now, suppose we have n IID samples $(z_1, z_2, \ldots, z_n)$ from the same distribution. What is the *joint* log-likelihood associated with these observations? (3 points)

M(c) What is the *maximum likelihood estimator* associated with this observation? Recall that the MLE is the value of the parameter that maximizes the likelihood, or equivalently maximizes the log-likelihood (or equivalently minimizes the negative log-likelihood). (3 points)

M(d) For our *bivariate* model, we have multiple data values $z_i^{(1)}, z_i^{(2)}$ for each observation instead. How should we modify our MLE for this data? What is the resulting MLE? (2 points)

Recall that the two values per observation are IID.

MML2. E-Step: To build the E-Step of our clustering algorithm, we need a rule for guessing the most probable class label for each observation. That is, assuming we know $\lambda_1, \lambda_2, \ldots, \lambda_K$, we build a *classification rule* for each observation $\boldsymbol{x}_i$. For this analysis, we will fix $K = 2$. (5 points total)

E(a) Given a new observation $\boldsymbol{x}_i$, compute its PMF under class 1 ($\lambda = \lambda_1$) and class 2 ($\lambda = \lambda_2$). (2 points)

E(b) Determine the relevant classification rule as a (simple) function of $x_i^{(1)}, x_i^{(2)}, \lambda_1, \lambda_2$. You may assume (without loss of generality) that $\lambda_1 > \lambda_2$. (3 points)

MML3. Finally, we are ready to put these steps together into a single unified clustering algorithm. Fill out the skeleton below using the M- and E-Steps you derived above (5 points)

Algorithm 1 An EM-Clustering Algorithm for IID-Bivariate Poisson Data

---

Input: n samples: $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$

Initialize:: Set cluster labels

$C_i =$ for $i = 1, \ldots n$

Repeat:

(a) M-Step: For $k = 1, 2$

(b) E-Step: For $i = 1, \ldots, n$

Until:

Return: Cluster labels $C_1, C_2, \ldots, C_n$

Practice of ML - Supervised: 25 points total

In this *issue spotter* question, I will describe a theoretical application of supervised machine learning. As described, this application falls short of "best practices" in several regards. Identify *five* places where this pipeline could be improved and recommend alternative (superior) practices. For each issue, you will receive:

- 1 point for identifying a valid issue
- (Up to) 2 points for explaining what is *wrong* with the described practice and *what impact / bias / error* would be induced
- (Up to) 2 points for accurately describing an alternative approach

Scenario:

> You have been hired to build a extreme weather prediction system for the City of New York. For the first stage of your project, the client has asked you to focus on predicting periods of extreme rainfall.
>
> To train your model, you have collected hourly rainfall data for the previous 25 years, giving a total of $n = 219,250$ training points. With each training point, you have also collected several predictors from weather stations surrounding NYC:
>
> - Temperature
> - Time of Day
> - Wind speed
> - Wind direction
> - Cloud cover percentage
> - Relative Humidity
> - Barometric Pressure
> - Visibility
> - Rate of Precipitation
>
> You have each of these measured contemporaneously, at one hour lag, at a three hour lag, and at a 24 hour lag, giving a total of $p = 36$ features.
>
> Given the dimensions of your data, you prefer to use a linear model and you begin by fitting a ridge regression model to all $n = 219,250$ training points. You observe an error of approximately 0.5 inches RMSE. In order to make your model more interpretable, you fit a lasso model. In order to tune λ, you pick λ so that your MSE matches that which you got from ridge regression. This gives you a model with 5 important features. Because the difference between 'extreme' rain

and 'average' rain is more than 0.5in, you consider your model to be sufficiently accurate to deliver to the customer.

After you send your results to your customer, your boss calls you into the office and provides some *rather emphatic* feedback. Based on this feedback, you scrap your model building process and start from scratch. You first transform the response variable to 0/1 indicating extreme rain (1) or or not (0). You then fit a xgboost classifier on this data using all the features. The longer you run xgboost, the smaller your True Negative Rate (TNR) becomes, so you run over $100,000$ iterations to get a TNR less than 0.01%. You show these results to your boss, who is now happy with them, and you then share them with the customer.

Your client is interested in better understanding the resulting model. To help your client, you provide feature importance scores which show current rate of precipitation and one-hour delayed rate of precipitation as the two most important features. Because these features cannot be known 3 or more hours in advance, your client believes the model is useless and wants to terminate the project without paying your company. To save the deal, you argue that the model also uses other features and that you can simply replace the two troublesome measures with 3 and 6 hour delayed values and the model won't suffer much loss in accuracy because those features are highly correlated. Satisfied, the client prepares to deploy your model into production.

Please note that there are more than five possible issues in this scenario.

Mistake #1:

Potential Adverse Impact #1:

Fix #1:

Mistake #2:

Potential Adverse Impact #2:

Fix #2:

Mistake #3:

Potential Adverse Impact #3:

Fix #3:

Mistake #4:

Potential Adverse Impact #4:

Fix #4:

Mistake #5:

Potential Adverse Impact #5:

Fix #5:

Practice of ML - Unsupervised: 25 points total

For this problem, I will describe a scenario in which unsupervised learning may be useful to achieve one or more scientific aims. You should describe, in reasonable detail, how you would approach this problem. Be sure to justify each step (say *why* you are making particular choices).

Note that this section will likely take longer to answer fully than any previous section, so budget your time wisely.

Scenario:

> You have been hired as a data analyst at marketing firm. Because you do not have any client relationships, you have been assigned to a *business development* (BD) team, which attempts to identify potential customers (companies) that could benefit from a new marketing campaign. After a productive strategy session with representatives from across the company, the BD manager asks you to identify several 'factors' that could be used to identify potential new customers. You are given access to the company's existing client data files which contain the following features:
>
> - Company Details: Company Size (Number of Customers); Average Revenue per Customer; Industry / Sector; Years with Current Ad Agency; Total Promotional Budget; Average Monthly Repeat Customers (pre-Advertising); Average Monthly New Customers (pre-Advertising); Average Monthly Spend per Existing Customer (pre-Advertising); Average First Month Spend per New Customer (pre-Advertising)
> - Current Marketing Campaign Details: Total Campaign Budget; Budget spent on Facebook Ads; Budget spent on YouTube Ads; Budget spent on Google AdSense Ads; Target Gender Breakdown (% Female); Target Age Breakdown (% 18-25); Target Age Breakdown (% 25-49); Target Age Breakdown (% 50-64); Target Age Breakdown (% 65+); Celebrity Endorsement (TRUE/FALSE);
> - Estimated Campaign Benefit: Increase in New Customers; Increase in Customer Retention; Increase in First Month Spend per New Customer; Increase in Monthly Spend per Existing Customer
>
> You have data for over 10,000 companies in your data base. Of these, only 2000 are your current clients for which you have marketing campaign details; of those, only 1000 have been engaged in their current campaign long enough to have estimated benefits.
>
> Before you start this project, your manager warns you that the database is not particularly well-curated and has occasional large errors.
>
> Your manager gives you a week to work with this data and asks you to develop an interpretable ML pipeline, resulting in an interpretable 'BD Index' that can

be passed to the sales team. For your meeting with the sales team in a week, you are asked to provide:

1. The names of the top 10 companies on the BD index
2. An explanation of the BD index that the sales team can provide feedback upon
3. Estimates of the potential benefits your company could provide the potential new clients identified by the BD index
4. An 'action plan' for addressing data quality issues in the current database
5. A strategy for 'validating' the BD index

While this project is "unsupervised" in that you are not provided a single canonical response variable, you may used supervised techniques wherever helpful.

Answer Below (up to 2.5 pages)

Pratice of ML - Unsupervised (continued):

Pratice of ML - Unsupervised (continued):

(Blank page for scratch work - not graded)

(Blank page for scratch work - not graded)