# Final Exam
# STA9890
# Statistical Learning for Data Mining

**Assignment Parameters:**
Date Assigned: 2024-05-17
Date Due: 2024-05-22 @ 11:55pm

**This is a closed-note, closed-book exam.**
**You may not use any external resources.**

## Instructions

This exam will be graded out of **200 points**.

You will have 3 hours (180 minutes) to complete this exam.

You will have 190 minutes from the time you download this exam on Blackboard to upload your solutions: that is, you have 180 minutes to complete the exam and 10 minutes to scan and upload any written answers. *You are responsible for uploading your exam in time, so plan accordingly.*

**Late exams will not be accepted.**

This exam is divided into five parts:

- Multiple Choice (50 points)
- Short Answer (50 points)
- Issue Spotter (25 points)
- Design of ML Pipeline (25 points)
- Mathematics of Machine Learning (50 points)

*Not all sections are worth the same amount of points nor are they equally difficult. Individual questions within a section also vary in difficulty. You need to use your time wisely. Skip questions that are not easy to answer quickly and return to them later.*

For the multiple choice and short answer questions, please answer only in the space provided on the exam sheet. I have provided some space for the issue spotter, pipeline design, and mathematical questions. If you need more space for these three sections, use the additional pages at the end of the PDF.

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

I *will not* answer questions during the exam period. If a question is ambiguous, do your best to answer it. If you need to make additional assumptions to answer a question, please state them in full.

**This is a closed-note, closed-book exam. You may not use any external resources.**

# Generalized Multiple Choice (50 points; 2 points each)

For each question, select zero or more answers, per question instructions.

**MC1.)** ☐ True or ☐ False (Select 1): After using cross-validation to select a model, the cross-validation error provides an unbiased estimate of predictive accuracy

**MC2.)** ☐ True or ☐ False (Select 1): In classification a *false positive* refers to an observation which is a member of the positive class but which is falsely labeled as negative by a classifier

**MC3.)** ☐ True or ☐ False (Select 1): When fitting the lasso, the optimal value of the penalty parameter $\lambda$ is higher for variable selection than it is for predictive accuracy

**MC4.)** ☐ True or ☐ False (Select 1): The singular value decomposition of a symmetric matrix is equal to its eigendecomposition.

**MC5.)** ☐ True or ☐ False (Select 1): Boosting is the practice of building an ensemble by sub-sampling observations.

**MC6.)** ☐ True or ☐ False (Select 1): The $K$-Means algorithm converges to a local - but not necessarily global - optimum.

**MC7.)** ☐ True or ☐ False (Select 1): $\ell_p$ norms define convex loss functions and penalties for $p > 0$.

**MC8.)** ☐ True or ☐ False (Select 1): Under a suitable generative model (i.e., $y \sim \text{Bern}(\mathsf{Logit}^{-1}(\boldsymbol{x}^\top \boldsymbol{\beta}_*))$), logistic regression is BLUE.

**MC9.)** ☐ True or ☐ False (Select 1): A sufficiently deep neural network can approximate any (continous) function arbitrarily well.

**MC10.)** ☐ True or ☐ False (Select 1): Poisson - or log-linear - regression is suitable for data with a positive continuous response, like insurance loss claims.

**MC11.)** ☐ True or ☐ False (Select 1): The decision boundaries of a $K$-nearest neighbor classifier become more jagged with larger $K$.

**MC12.)** ☐ True or ☐ False (Select 1): Suppose data follows the regression model $Y = f(X) + \epsilon$ for some mean-zero noise term $\epsilon$. The *regression function* $f(X) = E[Y|X]$ minimizes test error.

**MC13.)** ☐ True or ☐ False (Select 1): A linear model with $p$ features will always have worse training error than a linear model with $p + 2$ features fit to the same data.

**MC14.)** ☐ True or ☐ False (Select 1): The sample covariance matrix is always strictly positive definite.

**MC15.)** ☐ True or ☐ False (Select 1): Best subsets provides smaller training error than the lasso.

**MC16.)** Suppose $Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$ for some mean-zero $\epsilon$. If a misspecified linear model is fit with the features $(X_1, X_2, X_3)$ – note the extra $X_3$ – the resulting model will have ☐ higher or ☐ lower bias than the correctly specified model and ☐ higher or ☐ lower variance than the correctly specified model. (Select 2 boxes)

**MC17.)** Which choice of linkage produces the most compact clusters in hierarchical clustering: ☐ single; ☐ complete; ☐ average; ☐ Ward's.

**MC18.)** PCA guarantees which of the following properties: ☐ orthogonal loading vectors; ☐ orthogonal singular values; ☐ orthogonal score vectors; ☐ mean zero loading vectors.

**MC19.)** Which of the following kernels correspond to a finite-dimensional feature expansion: ☐ linear; ☐ polynomial; ☐ exponential; ☐ radial basis

**MC20.)** Which of the following are properties of support vector classifiers: ☐ use of hinge-loss; ☐ automatic identification of kernel points; ☐ lack of tuning parameter; ☐ probabilistic output.

**MC21.)** Which of the following conditions are *necessary* for boosting to improve predictive performance: ☐ non-zero predictive ability of base learning method; ☐ independence of base learning method; ☐ convexity of base learning method.

**MC22.)** Which of the following conditions are *necessary* for stacking to improve predictive performance: ☐ non-zero predictive ability of base learning methods; ☐ independence of base learning methods; ☐ convexity of base learning method.

**MC23.)** MAD regression estimates the conditional ☐ mean; ☐ median; ☐ mode; ☐ variance of the data distribution.

**MC24.)** Small decision trees (stumps) are immune to overfitting because: ☐ they only use randomly chosen features; ☐ they optimize Gini impurity; ☐ they use only a single split.

**MC25.)** Which principle(s) can be used to design pipelines to provide reproducible intepretations and discoveries: ☐ randomization principle; ☐ convexity principle; ☐ stability principle; ☐ cross-validation; ☐ robutness principle.

## Short Answer (50 points; 5 points each)

SA1. Suppose a data analyst provides you with the results of clustering on a training data set. Give one method that could be used to *validate* this clustering on a new data set.

SA2. Compare and contrast *best subsets* regression and the *lasso*

SA3. Describe how single-linkage clustering can be used to identify outliers in data.

SA4. Given a binary classifier, how might we extend it to multi-class classification? (*Hint: "One-vs-rest" approaches may be easier to describe.*)

SA5. Consider the bias, variance, training error, and test error of ridge regression as a function of $\lambda$. Sketch curves depicting the effect of $\lambda$ on each of these parameters. Label each curve clearly.

SA6. Recall that the SVD version of PCA can be implemented using the following algorithm:

---

(a) Initialize $\boldsymbol{u} \in \mathbb{R}^n$ and $\boldsymbol{v} \in \mathbb{R}^p$ randomly.
(b) Repeat until convergence:
- $\boldsymbol{u} \leftarrow \boldsymbol{X}\boldsymbol{v}/\|\boldsymbol{X}\boldsymbol{v}\|$
- $\boldsymbol{v} \leftarrow \boldsymbol{X}^\top\boldsymbol{u}/\|\boldsymbol{X}^\top\boldsymbol{u}\|$

---

Modify this code to add a non-negativity constraint on $\boldsymbol{u}$.

(For full credit, remember that the sign of $\boldsymbol{u}$ is not uniquely defined.)

SA7. Compare and contrast *bagging* and *boosting*.

SA8. Write out pseudo-code for $K$-means clustering (the Lloyd algirithm we discussed in class).

SA9. Compare and contrast *additive (spline) models* and *kernel regression*.

SA10. Under what general conditions should linear methods be preferred to non-linear statistical machine learning methods and why?

# Issue Spotter (25 points)

This is an *issue spotter* question. Below, I describe (in words) a hypothetical application of the ML techniques we have discussed in this class. Your task is to find **5 mistakes** in the ML pipeline and to:

i) describe the problem (2 points each); and

ii) say how that step could have been performed more effectively / accurately. (3 points)

---

Geneticists have collected *high-throughput* genetic data on a variety of kidney tissue samples, half of which are cancerous and half of which are not. For each of $p = 10,000$ genes, they have recorded the level of gene expression (how 'turned on' that gene is) in the tissue sample; their data set is equally balanced with 200 healthy samples and 200 cancerous samples. The goal of the study is to identify which genes are associated with higher probabilities of renal cancer.

Because $p \gg n$ for this problem, the scientists first perform PCA to reduce the dimensionality of the data. After doing so, they retain the top 20 principal components and perform $K$-means clustering in this reduced space with $K = 2$. They examine the results of the clusters and find that Cluster 1 is almost entirely cancerous samples and that PC3 separates the two clusters well. They scientists identify the 5 genes with the highest loadings in PC3 and report them as important candidates for future study.

After performing this analysis, they scientists speak to a data scientist who recommends they approach this problem instead as a supervised learning problem. The data scientist creates a response vector $\boldsymbol{y}$ where $y_i = 1$ if the cell is cancerous and $y_i = 0$ otherwise. The data scientist then fits a support vector classifier (no kernel) to the 400 samples.

In order to tune the SVC, the data scientist picks the value of the tuning parameter that minimizes the false positive rate, recognizing that it is important not to falsely tell patients they have cancer. The data scientist then picks the genes with the most positive values of the coefficients to identify the most important genes for cancer prediction. Because the SVC is over 95% accurate, the data scientist informs the scientists that the selected variables are statistically significant at 95% confidence ($p < 0.05$).

**Mistake #1:**

**Fix #1:**

---

**Mistake #2:**

**Fix #2:**

---

**Mistake #3:**

**Fix #3:**

---

**Mistake #4:**

**Fix #4:**

---

**Mistake #5:**

**Fix #5:**

# Design of ML Pipeline (25 points)

Below, I describe a hypothetical data analytic challenge. Describe how you would approach this problem, taking care to describe practical considerations like data splitting, selection of tuning parameters, estimation of predictive power, model validation *etc.* as appropriate.

> You are a data analyst working on the advertising team of a presidential campaign. Your team's goal is to identify target demographics who can be persuaded to vote for your candidate and to design ads that highlight your candidate's appeal to that demographic.
>
> Your data team has provided you with a detailed *voter data file* containing the following information for all voting age adults in the US:
>
> - Age
> - Ethnicity
> - Gender
> - Estimated Household Income
> - Level of Education
> - State of Residence
> - Current voter registration status (is this person already registered to vote in the upcoming election?)
> - Did this person vote in the previous election?
> - Did this person vote in both of the previous elections?
> - Did this person vote in the past three previous elections?
> - Is this person a registered member of your party?
> - What issue (of a provided list) is most important to this person?
> - What issue (of a provided list) is second most important to this person?
>
> (Note that not all of these fields are necessarily accurate, but the data team has *imputed* this data so there are no missing values.)
>
> Your polling team reports that the race is particularly tight in North Carolina and campaign leadership has asked you to identify the target demographic for a new advertising campaign. You have been given the budget to create 3 possible ads and to convene focus groups on them before beginning the full-scale advertising campaign.
>
> **How would you approach this problem?**

*(Note that this question does not have a single right answer. Grading will be based on creativity and suitability of approach, adherence to ML best practices, justification of decisions, etc.)*

# Mathematics of ML (50 points total; 25 points each)

**MA1.) Properties of Linear Regression Under Orthogonal Design.**

Suppose $\boldsymbol{X}$ is an $n \times n$ *orthogonal* matrix satisfying $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{I}$. Suppose further that the response is generated as $y = \boldsymbol{x}^\top \boldsymbol{\beta}_* + \epsilon$ where $\boldsymbol{\beta}_*$ is a fixed (unknown) vector and $\epsilon$ is a (scalar) Gaussian random variable with mean 0 and standard deviation $\sigma$. ($\boldsymbol{x}_i, \epsilon_j$ are mutually IID.)

**MA1.a)** What is the solution to the following *ridge regression* problem?

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{\beta}\|_2^2$$

**MA1.b)** What is the bias of ridge regression in this scenario? What is the bias of OLS?

**MA1.c)** What is the variance of ridge regression in this scenario? What is the bias of OLS?

**MA1.d)** What is the *prediction MSE* of ridge regression in this scenario?

**MA1.e)** Is it possible to *correct for* the bias of ridge regression in this scearnio? If so, provide a formula for bias-corrected ridge regression and give its MSE. Does it improve upon (biased) ridge regression?

**MA2.) Derivation of a Generative Classifier**.

Suppose you have data points generated from a two class mixture as:

$$\boldsymbol{X}_i \sim \begin{cases} \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) & \text{if } i \text{ is in class 1} \\ \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}) & \text{if } i \text{ is in class 2} \end{cases}$$

Here, the mean vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ but the common covariance $\boldsymbol{\Sigma}$ is known. State and derive the decision boundary of an appropriate *generate* classifier for this problem: show your work.

You may choose to use the following steps, but it is not necessary.

(a) State the *conditional* PDF for $\boldsymbol{X}_i$.

(b) State Bayes' rule for estimating the class membership of a test data point.

(c) Manipulate Bayes' rule to get a linear expression for the decision boundary (the points where the probabilites of the two classes are equal).

(d) Simplify all expressions.

(Extra page for longer answers)

(Extra page for longer answers)

(Extra page for longer answers)

(Extra page for longer answers)

(Blank page for scratch work - not graded)

(Blank page for scratch work - not graded)

(Blank page for scratch work - not graded)

(Blank page for scratch work - not graded)