STA 9890 - Statistical Learning for Data Mining

In-Class Test 2: Supervised Learning

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

Name: _____

Instructions

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

This exam is divided into three sections:

- True/False (24 points; 12 questions at two points each)
- Short Answer (52 points; 13 questions at 4 points each)
- Mathematics of Machine Learning (24 points; one long question in 6 parts)

You have 90 minutes 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 will 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.

True/False: 24 points total at 2 points each

For each question, **CIRCLE** your answer(s).

TF1. True/False: A maximum likelihood estimator is one which sets the unknown parameters in order to minimize the negative log PDF/PMF of the observed data.

TRUE FALSE

TF2. True/False: Multi-class classification with K classes can be performed by combining K binary classifiers using a "one-vs-rest" strategy.

TRUE FALSE

TF3. True/False: At the same level of sparsity, the lasso has smaller training error than best subsets regression.

TRUE FALSE

TF4. True/False: *Bagging* is the practice of building an ensemble by sub-sampling features.

TRUE FALSE

TF5. True/False: Linear models are preferred in (relatively) small data scenarios because they have (relatively) lower variance.

TRUE FALSE

TF6. True/False: The elastic net penalty combines the ridge and best subset penalties.

TRUE FALSE

TF7. True/False: Because kernel methods are more flexible than purely linear models, they always provide out-of-sample (test) error improvements.

TRUE FALSE

TF8. True/False: In an SVM, *support vectors* are training points closest to the decision boundary.

TRUE FALSE

TF9. True/False: Ridge regression has the closed form solution $\hat{\boldsymbol{\beta}}_{\text{Ridge}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$

TRUE FALSE

TF10. True/False: Spline methods can only be applied in regression settings, not classification

TRUE FALSE

TF11. True/False: Best subset regression can be solved efficiently using backwards stepwise approaches, but not forward stepwise.

TRUE FALSE

TF12. True/False: Logistic regression can be considered a *soft* or *probabilistic* classifier.

TRUE FALSE

Short Answer: 52 points total at 4 each

SA1. Which of the following are *discriminative classifiers*? Check the box for all that apply.

a. ☐ Logistic Regression
b. ☐ Naive Bayes
c. ☐ Decision Trees
d. ☐ Linear Discriminant Analysis
e. ☐ Random Forests
f. ☐ Bayes' Rule
g. ☐ `xgboost`
h. ☐ Support Vector Machines
i. ☐ Kernel Logistic Regression
j. ☐ Deep Neural Network
k. ☐ Quadratic Discriminant Analysis
l. ☐ Latent Dirichlet Allocation

SA2. Define *shrinkage* in the context of regression and explain why it may be helpful.

SA3. Give three reasons we may want to use a *sparse* regression or classification model.

SA4. In the context of *ML Fairness*, "equality of opportunity" is often defined as two groups having the same *true positive rate* where the TPR is defined as

$$\text{TPR} = \frac{\text{TP}}{\text{P}} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \text{Sensitivity / Recall}$$

Given the following two confusion matrices, is equality of opportunity satisfied? If no, how far off is it?

Group 1		Ground Truth +	Ground Truth −
Prediction	+	50	20
	−	10	1000

Group 2		Ground Truth +	Ground Truth −
Prediction	+	25	15
	−	5	100

SA5. Success of Lasso regression can be measured on two criteria: i) does it make accurate predictions? (or, at least, predictions as accurate as any sparse linear model could make) and ii) does it accurately select the right features? For each criterion, describe the conditions needed for the lasso to perform well on each criterion. Which is more commonly satisfied?

Prediction Accuracy:

Selection Accuracy:

SA6. In your own words, write out a 5-fold cross-validation scheme for tuning ridge logistic regression to maximize classification accuracy. You must also include steps that guarantee an *unbiased* estimator of the test-set classification accuracy.

SA7. Describe the three parts of a *generalized linear model*, noting their general role in GLM specification and precisely identifying them in Poisson (log-linear or count) regression:

I. Name: _____

- Purpose: _____
- In Poisson regression:_____

II. Name: _____

- Purpose:_____
- In Poisson regression:_____

III. Name: _____

- Purpose:_____
- In Poisson regression:_____

SA8. Some statistical software implements logistic regression with a small ridge penalty that cannot be disabled. Give two reasons why this is a reasonable design choice.

SA9. Why is there minimal risk of overfitting from excessive bootstrapping in a *bagging* procedure?

SA10. Compare and contrast spline and kernel models. Give 2 key similarities ("compare") and 2 key differences ("contrast").

SA11. Give three scenarios, other than simply combining different base learners to improve performance, where *model stacking* could solve a practical problem.

SA12. Rank the following methods in terms of statistical complexity (*i.e.* potential wiggliness) with 1 being the lowest and 5 being the highest:

- OLS _____
- 1-Nearest Neighbor Regression _____
- Piecewise Cubic Spline Regression _____
- Piecewise Cubic Polynomial Regression _____
- Ridge Regression _____

You may assume the two piecewise methods have the same number of pieces.

SA13. Modern multi-layer (deep) neural networks are formed by *composing* a series of linear and non-linear ('activation') terms. *E.g.*,

$$\hat{y} = f_0(f_{1,1}(f_{1,1,1}(\boldsymbol{X}), f_{1,1,2}(\boldsymbol{X})), f_{1,2}(f_{1,2,1}(\boldsymbol{X}), f_{1,2,2}(\boldsymbol{X})))$$

where each $f(\cdot)$ is of the form $(\boldsymbol{X}\boldsymbol{w})_+$ - that is, multiply $\boldsymbol{X}$ by some vector of weights $\boldsymbol{w}$ and set negative elements to zero. (Real networks typically have many more layers and inputs to each layer.) Using the principles discussed in this course, what can you infer about the behavior of this type of system? When will it be successful and when will it struggle? What does this tell you about the inputs and scenarios required to build and successfully deploy this type of model?

Hint: Recall that non-linear functions of linear inputs are non-linear.

Mathematics of Machine Learning: 24 points total

You are a researcher trying to use a new *single-cell sequencing technique* as a diagnostic tool for a moderately rare disease that occurs in about 5% of the population. (You may assume this is an 'on/off' disease and that there are no gradations or different severities.) The details of this technique are irrelevant to this problem, but in essence, it can count the number of proteins of a given type in a cell. Your hypothesis is that patients with the disease have a different amount of this protein in their cells than 'healthy' (control) patients.

Because this protein is relatively rare in both patient groups, you have chosen to model the count of the protein as a *Poisson* random variable with different means for each group. Based on prior research, healthy patients have an average of 3 copies of this protein per cell, while diseased patients have 6 copies per cell. The sequencing technique you use produces statistically independent counts for 5 cells from each patient; that is, you do not need to worry about correlation among the counts. In this problem, you will develop a custom *generative* classifier to identify this disease.

Recall: if a variable X follows a Poisson distribution with mean λ it has a probability mass function given by

$$\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \text{ for } k = 0, 1, 2, \ldots$$

Finally, note that MML4 and MML6 do not depend on solving MML1-3, so attempt those even if the earlier parts are too difficult.

MML1) You apply your sequencing technique to a patient and it reports the following protein counts: $(5, 4, 10, 1, 7)$. What is the probability of observing this set of observations if the patient is from the diseased group? (You do not need to evaluate this fully; just simplify it as far as is reasonable.)

MML2) Using Bayes' rule as the basis of a generative classifier, what is the probability the patient from the previous part has the disease?

Hint: The denominator of the Poisson PMF doesn't depend on λ, so many factorial terms will simplify and cancel.

MML3) Generalizing your results, develop a simple *decision rule* for this problem. That is, give a simple set of conditions that indicate whether the patient is predicted to have the disease or not. Note that your decision rule must be *simple* enough to be implemented using a six-function calculator (add, subtract, multiply, divide, powers, logs), and may not just be a general aphorism like "do Bayes' rule."

Hint: Independent Poisson random variables add. That is, if X, Y are independent with $X \sim \text{Poisson}(\lambda_X)$ and $Y \sim \text{Poisson}(\lambda_Y)$ then $X + Y \sim \text{Poisson}(\lambda_X + \lambda_Y)$

MML4) After some early initial successes, you want to develop a clinical protocol using your diagnostic system. You have settled on the following structure:

- If $\mathbb{P}(\text{disease}|X) = \hat{p} > \overline{p}$, provide additional 'gold-standard' testing at a cost of $\$4,000$. You may assume the 'gold-standard' test never makes a mistake.
  - If the gold-standard test confirms disease, treat at a cost of an *additional* $\$10,000$
  - If the gold-standard test confirms no disease, no additional treatments are required
- If patient is diseased but this is missed in initial screening, complications ensue that cost $\$50,000$ to treat

Here $\overline{p}$ is a cost-sensitive threshold that is chosen to balance the various *costs* associated with both false positives and false negatives in order to ensure minimum expected cost.

Find the optimal $\overline{p}$ for this problem. You may assume your classifier is well-calibrated: that is, if you have $\mathbb{P}(\text{disease}|X) = \hat{p}$, then the patient *truly* has a probability $\hat{p}$ of having the disease and (hence) the gold-standard method will detect disease with probability $\hat{p}$.

Hint: This calculation does not depend on the classification rule you developed in the previous part. You simply need to find the $\overline{p}$ that balances the costs associated with sending the patient for 'gold-standard' testing vs not sending them and use that to set your boundary.

MML5) Combining your answers from the previous two sections, what is the clinical decision rule for this protocol? *I.e.*, fill in the blank for "If _____, send the patient for gold-standard testing."

MML6) At an *intuitive* and *qualitative* level, explain what happens to your protocol in the following scenarios:

- The cost of gold-standard testing *decreases*
- A new pill is developed so now the costs of missing a diagnosis and leaving the disease untreated *decreases* to $\$25,000$.
- You adapt your protocol for a different country in which the disease is *more common*
- A new study suggests that protein levels are *higher* in patients with the disease

A brief answer (one or two sentences) will suffice for each part. You do not need to provide exact numbers - just indicate and explain the *direction* of changes.

(Blank page for scratch work - not graded)

(Blank page for scratch work - not graded)

STA9890 - Test 1 - Formula Sheet

Linear Algebra:

- A *n-vector* is an ordered set of n (real) numbers: $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, with addition $\boldsymbol{x}+\boldsymbol{y} = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)$ and vector (inner / dot) product: $\boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{x}^\top \boldsymbol{y} = \sum_{i=1}^n x_i y_i$
- Vector norms: $\|\boldsymbol{x}\|_p = \sqrt[p]{\sum_{i=1}^n |x_i|^p}$ with $\|\boldsymbol{x}\|_\infty = \max_i\{|x_i|\}$ and $\|\boldsymbol{x}\|_0 =$ Number of non-zero elements o
- An $m \times n$ matrix is a 2D array of real numbers with m rows and n columns:

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix}$$

- A matrix-vector product takes an n-vector as input and gives an m-vector as output:

$$\boldsymbol{A}\boldsymbol{x} = (\boldsymbol{A}_{1\cdot} \cdot \boldsymbol{x}, \boldsymbol{A}_{2\cdot}\boldsymbol{x}, \ldots, \boldsymbol{A}_{m\cdot} \cdot \boldsymbol{x}) \in \mathbb{R}^m$$

- We can multiply an $m \times n$ matrix with an $n \times p$ matrix - note that the 'inner' dimensions must match:

$$\boldsymbol{A}\boldsymbol{B} = \begin{pmatrix} \boldsymbol{A}_{1\cdot} \cdot \boldsymbol{B}_{\cdot 1} & \boldsymbol{A}_{1\cdot} \cdot \boldsymbol{B}_{\cdot 2} & \ldots & \boldsymbol{A}_{1\cdot} \cdot \boldsymbol{B}_{\cdot n} \\ \boldsymbol{A}_{2\cdot} \cdot \boldsymbol{B}_{\cdot 1} & \boldsymbol{A}_{2\cdot} \cdot \boldsymbol{B}_{\cdot 2} & \ldots & \boldsymbol{A}_{2\cdot} \cdot \boldsymbol{B}_{\cdot n} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{A}_{m\cdot} \cdot \boldsymbol{B}_{\cdot 1} & \boldsymbol{A}_{m\cdot} \cdot \boldsymbol{B}_{\cdot 2} & \ldots & \boldsymbol{A}_{m\cdot} \cdot \boldsymbol{B}_{\cdot n} \end{pmatrix} \in \mathbb{R}^{m \times p}$$

  Consider n-vectors as *one-column* matrices to make all of these definitions consistent. Requiring the dimensions in multiplication to align is a good way to verify linear algebra claims. (*E.g.*, $\boldsymbol{A}\boldsymbol{A}$ does not work for non-square $\boldsymbol{A}$)
- A matrix inverse satisfies $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$. Only full-rank square matrices have inverses
- An (square) orthogonal matrix $\boldsymbol{Q}$ satisfies $\boldsymbol{Q}^\top = \boldsymbol{Q}^{-1}$. If we take the first $n' \leq n$ columns (rows) of an orthogonal matrix we have $\boldsymbol{Q}_{1:n'\cdot}\boldsymbol{Q}_{1:n'\cdot}^\top = \boldsymbol{I}_{n' \times n'}$ so it's transpose-inverse along the 'short-side'
- Any matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ has a *singular value decomposition*: $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ where $r = \min\{m, n\}$, $\boldsymbol{D}$ is a non-negative diagonal $r \times r$ matrix, $\boldsymbol{U} \in \mathbb{R}^{m \times r}$ is the first r columns of an orthogonal $m \times m$-matrix, and $\boldsymbol{V} \in \mathbb{R}^{n \times r}$ is the first r columns of an orthogonal $n \times n$ matrix
- Distributive rules: $(\boldsymbol{A}\boldsymbol{B})^\top = \boldsymbol{B}^\top\boldsymbol{A}^\top$ and $(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$ (if all defined)

Matrix Calculus:

- Quadratics: $f(\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} \implies \nabla f(\boldsymbol{x}) = 2\boldsymbol{A}\boldsymbol{x}$; $f(\boldsymbol{x}) = \|\boldsymbol{x}\|^2 = 2\boldsymbol{x}$
- Chain rule: $g(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{x}) \implies \nabla g(\boldsymbol{x}) = \boldsymbol{A}^\top(\nabla f)(\boldsymbol{A}\boldsymbol{x})$

Convexity:

- A function $f : \mathbb{R}^p \to \mathbb{R}$ is *convex* if

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}) \text{ for all } \lambda \in [0,1], \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$$

  If f is convex and second-differentiable at a point, its second derivative matrix is *positive semi-definite*

- A set $\mathcal{C} \in \mathbb{R}^p$ is convex if

$$\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C} \implies \lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y} \in \mathcal{C} \text{ for all } \lambda \in [0,1], \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$$

- If $\nabla f(\boldsymbol{x}_*) = 0$ for convex $f(\cdot)$, then $\boldsymbol{x}_*$ is a global minimizer of $f(\cdot)$

Gradient Methods:

- Given an optimization problem $\min_{\boldsymbol{x} \in \mathcal{C}} f(\boldsymbol{x})$, gradient descent works by repeating the following update:

$$\boldsymbol{x}^{(k+1)} \leftarrow \boldsymbol{x}^{(k)} - c \nabla f(\boldsymbol{x}^{(k)})$$

  If $c > 0$ is sufficiently small and $\mathcal{C} = \mathbb{R}^p$, $\boldsymbol{x}^{(k)}$ will converge to a minimizer of f

STA9890 - Test 2 - Formula Sheet

Ordinary Least Squares:

- $\hat{\beta} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$

Penalties:

- Best Subsets: $\|\boldsymbol{\beta}\|_0$. Number of non-zero elements (non-convex), induces sparsity in $\boldsymbol{\beta}$
- Lasso: $\|\boldsymbol{\beta}\|_1$. Tightest convex relaxation of best subsets
- Ridge: $\frac{1}{2}\|\boldsymbol{\beta}\|_2^2$. Very nice to work with (differentiable). $\frac{\partial}{\partial \boldsymbol{\beta}} \|\boldsymbol{\beta}\|_2^2 = 2\boldsymbol{\beta}$.
- Elastic Net: α-weighted combination of ridge and lasso $\alpha\|\boldsymbol{\beta}\|_1 + \frac{(1-\alpha)}{2}\|\boldsymbol{\beta}\|_2^2$

Classification:

- (True/False) (Positive/Negative) = (Correct/Incorrect) Prediction
- Generative: $p(X|Y) \implies p(Y|X)$ via prior and Bayes' Rule. Discriminative: model $p(Y|X)$ directly.
- Bayes' Rule:

$$p(A|B) = \frac{p(B|A) * p(A)}{P(B)} = \frac{p(B|A) * p(A)}{P(B|A) * P(A) + P(B|A^c) + P(A^c)}$$

Non-Linearity:

- Feature expansion and engineering: fit linear models to non-linear parts
- Splines: piecewise polynomial models with additional smoothness constraints
- Kernel methods: feature expansion made 'easy'. Replace inner product with a 'kernel function'.

Ensembles:

- Stacking: Linear combination of base learners. Typically non-negative and sum-to-one constrained
- Bagging (Bootstrap Aggregation): building an ensemble by averaging bootstrapped based learners
- Boosting: building an ensemble by adding new ensemble members to correct past mistakes. Fit slowly for 'gradient descent' on functions

Distributions:

- Bernoulli Distribution: $X \sim \text{Bernoulli}(p) \in \{0,1\} \implies \mathbb{P}(X = x) = p^x(1-p)^{1-x}$
- Binomial Distribution: $X \sim \text{Binomial}(n,p) \in \{0,\dots,n\} \implies \mathbb{P}(X = x) = \binom{n}{x} p^x(1-p)^{n-x}$
- Poisson Distribution: $X \sim \text{Poisson}(\lambda) \in \{0,1,\dots\} \implies \mathbb{P}(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$
- Standard normal distribution. $Z \sim \mathcal{N}(0,1)$. Mean Zero + Variance 1
- Standard normal PDF - $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$. Standard normal CDF $\Phi(z) = \int_{-\infty}^{z} \phi(x)\,\mathrm{d}x$ - no closed form.
- General normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$ - generated by scale+shift of standard normal $X \overset{d}{=} \mu + \sigma Z$.
- Normal PDF via standardization (z-score): $f_X(x) = \phi(\frac{x-\mu}{\sigma}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$. CDF: $\Phi(\frac{x-\mu}{\sigma})$.
- Multivariate normal parameterized by mean vector and (co)variance matrix: $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Standard multi-normal: $\boldsymbol{Z} \sim \mathcal{N}_n(\boldsymbol{0}_n, \boldsymbol{I}_n)$. PDF $f_{\boldsymbol{Z}}(\boldsymbol{z}) = (2\pi)^{-n/2} e^{-\|\boldsymbol{z}\|^2/2}$.
- General multi-normal $\boldsymbol{X} \overset{d}{=} \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}$ where $\boldsymbol{\Sigma}^{1/2}$ is a matrix square root (Cholesky or symmetric).
- Bivariate normal PDF

$$f_{(X,Y)}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2\left[1-\rho^2\right]}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right)$$

- Multivariate normal: any linear combination (weighted sum) of X_i is normal.
- If $\mathbb{C}[X_i, X_j] = 0$, then $X_i \perp\!\!\!\perp X_j$ (for multi-normal, uncorrelated implies independent)
- If $\boldsymbol{Z}$ is a standard normal n-vector, $\|\boldsymbol{Z}\|^2 = \sum_{i=1}^{n} Z_i^2$ has a χ^2 distribution with n degrees of freedom