Midterm Exam

CEE 616: Probabilistic Machine Learning

November 24, 2025

time limit: 24 Hours

Instructions

This exam contains **8 problems** worth a total of **81 points** worth **7 points**. You have **24 hrs** to complete the exam from the time you download the PDF.

The following rules apply:

- This is an open-resource examination, so you are free to consult any resource at your disposal.
- Submit your completed examination as a **PDF**. You may type, typeset or write your answers and scan (you can also use the spaces provided in this document and submit an edited version of this PDF).
- Name your file as `LASTNAME_FIRSTNAME_Midterm.pdf`.
- **Organize your work**, in a reasonably neat and coherent way, in the space provided. Work scattered all over the page without a clear ordering will receive very little credit.
- **Show ALL your work where appropriate**. The work you show will be evaluated as well as your final answer. Thus, provide ample justification for each step you take. Indicate when you have used a probability table to obtain a result. In the long response questions, simply putting down an answer without showing your steps will not merit full credit. **EXCEPTION:** For short response or "True/False" questions, *no explanations are required*. Generally, however, the more work you show, the greater your chance of receiving partial credit if your final answer is incorrect.
- If you need more space, use the blank pages at the end, and clearly indicate when and where you have done this.

Problem 1 *True/False questions (10 points)*

Respond "T" (*True*) or "F" (*False*) to the following statements.

(i) ☐ In exemplar-based methods, the "effective" number of parameters remains the same for each test observation.

(ii) ☐ Cost-complexity pruning can mitigate overfitting in decision trees.

(iii) ☐ Mercer kernels are symmetric functions.

(iv) ☐ A softmax activation function can be used in the output layer of an ANN for a regression problem.

(v) ☐ Bagging reduces stability in decision tree predictions.

(vi) ☐ Boosting can be considered an ensemble learning method.

(vii) ☐ Gaussian process regression is a parametric modeling approach.

(viii) ☐ In principal components analysis, the extracted principal components are pairwise orthogonal.

(ix) ☐ Clustering is a supervised learning approach.

(x) ☐ The decision boundary of a fitted support vector classifier is linear.

Problem 2 *Short response: nonparametric methods (12 points)*

(a) Name the [hyper]parameters required for a K-nearest neighbor classification model. *[2]*

(b) Provide two examples of nonparametric modeling methods that use the kernel trick. *[2]*

(c) Name two density kernels. *[2]*

(d) Briefly state Mercer's theorem. *[2]*

(e) Name two Mercer kernels. *[2]*

(f) Briefly explain the key distinction between Gaussian Process estimation with noise-free and noisy observations. *[2]*

Problem 3 *Short answer questions (22 points)*

(a) The correlation coefficient is a measure of the linear dependence between two random variables. Which measure provides a more robust measure of the dependency between two random variables? *[1]*

(b) Briefly explain or show why the determinant of a positive semidefinite matrix can be zero. *[2]*

(c) In the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w}$, we will not have a unique solution if $\boldsymbol{X}^\top\boldsymbol{X}$ is singular. How can this issue be addressed? *[2]*

(d) Write the likelihood $p(y|\boldsymbol{x}, \boldsymbol{\theta})$ of an outcome y which is governed by a Gaussian distribution with parameters $\boldsymbol{\theta} = (\boldsymbol{w}, \sigma^2)$. Note that the mean of the Gaussian is $\boldsymbol{w}^\top\boldsymbol{x}$, its variance is σ^2 and $\boldsymbol{x} = [1, x_1, \ldots, x_D]$. *[2]*

(e) What is a quantity that is used to measure the dissimilarity between two distributions p and q? Write its equation in terms of p and q (either discrete or continuous case is acceptable). *[2]*

(f) Briefly state the fundamental difference between the linear discriminant analysis and the quadratic discrimant analysis models. *[2]*

(g) In ordinary least squares (OLS), what is the *hat matrix*? Write an equation for this in terms of the design/data matrix $\boldsymbol{X}$. *[2]*

(h) List two differences between ridge regression and lasso regression. *[2]*

(i) In a 2D binary logistic regression (i.e. $y \in \{0, 1\}$), the mean response (i.e. probability that $y = 1$) is given by: *[3]*

$$p(y = 1|\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{1 + e^{-(w_0 + w_1 x_1 + w_2 x_2)}} \tag{1}$$

If a threshold of $\tau = 0.5$ is used to assign predictions, derive the equation of the decision boundary. Your final answer should only have x_2 on the LHS.

(j) Write the equation for the ridge regression estimate $\hat{\boldsymbol{w}}^{\text{ridge}}$ in matrix form using the standard symbols $\boldsymbol{X}$, $\boldsymbol{y}$ and λ. Is the intercept term included in the ridge estimate? Why or why not? State the dimensions of $\boldsymbol{X}$ in this case, assuming N observations and D features. *[4]*

Problem 4 *Bayes (7 points)*

The posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$ of parameters $\boldsymbol{\theta}$ in a model estimated on a dataet $\mathcal{D}$, given by Bayes rule as: *[3]*

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})} \tag{2}$$

(a) Write the names/labels for each of the terms on the RHS of the above equation. (For example, the name of the term on the LHS is the **posterior**.

(b) Which term(s) in Equation 2 need(s) to be optimized to find the MLE estimate $\hat{\boldsymbol{\theta}}_{\text{mle}}$. Briefly justify your response. *[2]*

(c) Which term(s) in Equation 2 need(s) to be optimized to find the MAP estimate of $\hat{\boldsymbol{\theta}}_{\text{map}}$. Briefly justify your response. *[2]*

Problem 5 *Short response: neural networks (11 points)*

(a) What kind of RNN is used for sequence classification? *[1]*

(b) What is the basic algorithm used for training RNNs? *[1]*

(c) What are two approaches for handling the vanishing/exploding gradient problem in RNN training? *[2]*

(d) What is semantic segmentation? Name a neural net architecture that can tackle this problem. *[2]*

(e) What is the advantage of the mini-batch approach over the plain stochastic gradient descent? *[2]*

(f) Write the equation for the mini-batch $\mathcal{B}_t$ gradient update $\boldsymbol{\theta}_{\ell,t+1}$. Define all the symbols in your equation. *[3]*

Problem 6 *CNN (6 points)*

Consider a CNN trained for a classification problem (shown in Figure 1).

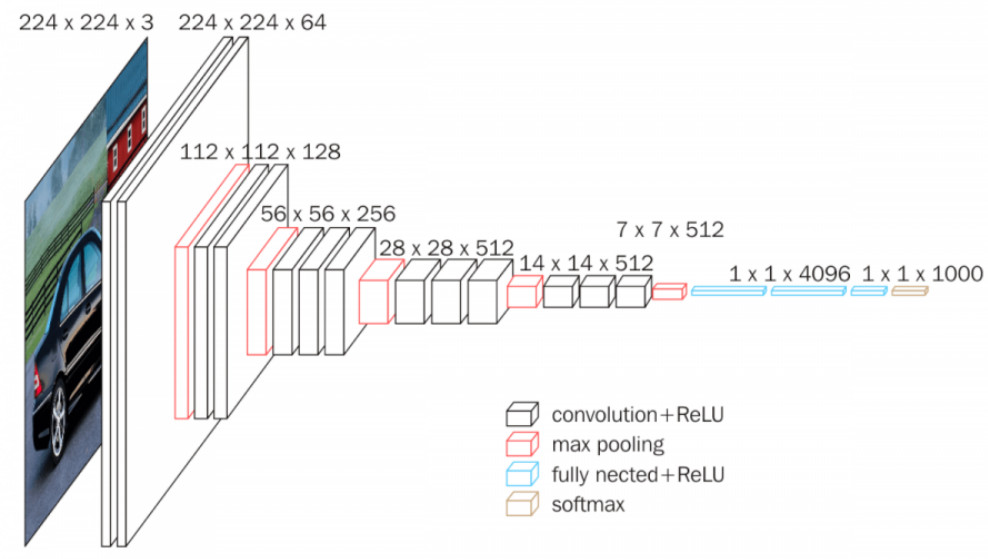


Figure 1: Diagram of a convolutional neural network architecture

(a) What are the dimensions of the input image?

(b) How many filters are specified in the first convolutional layer?

(c) What is the kernel size of the pooling layers used in this model?

(d) How many filters are specified in the third convolutional layer?

(e) What is the kernel size in all the convolutional layers?

(f) How many class labels are in the dataset?

Problem 7 *SVM (8 points)*

(a) In a binary classification problem, a maximal margin classifier can only be fitted if the observations are _____ separable. *[1]*

(b) What is another term for the support vector classifer (SVC)? *[1]*

(c) The SVC optimization problem allows for margin-overlapping observations through the use of _____ variables. *[1]*

(d) In SVC/SVM, the number of margin-overlapping observations is controlled via the _____ parameter. *[1]*

(e) The estimated decision boundary $\hat{f}$ in SVC is given by *[1]*

$$\hat{f}(\boldsymbol{x}) = \hat{w}_0 + \sum_{n=1}^{N} \hat{\alpha}_n \tilde{y}_n \boldsymbol{x}_n^\top \boldsymbol{x} \tag{3}$$

Identify the inner product term in the equation.

(f) Describe the modification to Equation 3 that would result in an SVM decision boundary and write the new equation. (Hint: Your equation should include the symbol $\mathcal{K}$.) *[3]*

Problem 8 *Trees and Ensemble learning (5 points)*

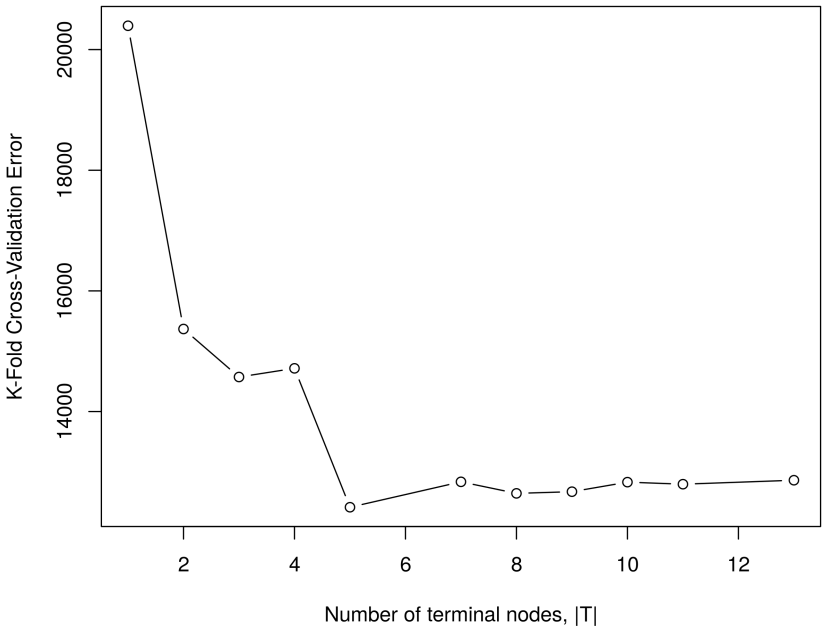
(a) You are tasked with fitting a regression tree to predict the median housing values in Boston suburbs based on two predictors: (1) the proportion of non-retail business acres per town, and (2) per capita crime rate by town. You want to choose the best tree T via cost-complexity pruning, where the cost $C_\alpha(T)$ is given by:

$$C_\alpha(T) = \left[\sum_{m=1}^{|T|} |x_i \in R_m| \sum_{x_i \in R_m} (y_i - \overline{y}_m)^2 \right] + \alpha |T| \tag{4}$$

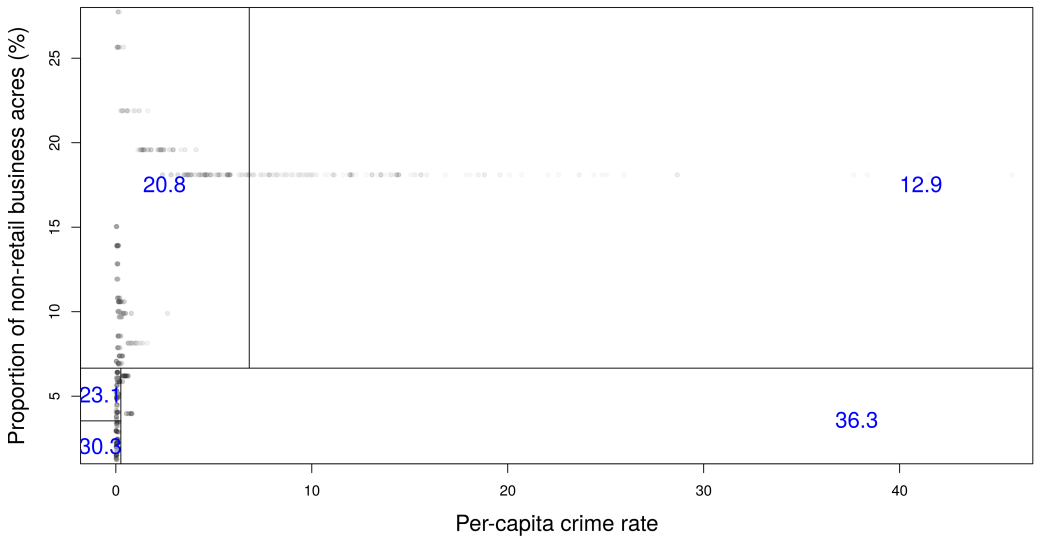
where: α is a hyperparameter, m is the index of terminal nodes (leaves), $|x_i \in R_m|$ is the number of observations in each terminal region/node R_m and $\overline{y}_m$ is the average response in each terminal node.

(i) You use cross-validation to find the cost-complexity-pruned tree with the lowest CV error. The resulting plot is shown below. *[1]*

What is the size (number of terminal nodes) of the best-fitting tree?



(ii) The terminal regions of the best-fit pruned tree are shown in the figure below. The resulting predictions for each node are given in thousands of US$. *[1]*



What is the predicted median housing value (to 2 significant figures) for a suburb with a crime rate of 40 per-capita and proportion of non-retail business acres of 5% ?

(b) Trees are inherently high-variance models. Which approach might you use to address this? *[1]*

(c) The out-of-bag (OOB) error rate in bagging or random forests is an estimate of the _____ error. *[1]*

(d) The expected proportion of observations that not in a bootstrap sample is _____. *[1]*
(An approximate answer is fine here.)

CEE 616 | J. Oke | Fall 2025

Department of Civil and Environmental Engineering
University of Massachusetts Amherst