Name:	Perm:

CS 165B Machine Learning Practice Final Exam

12:00 PM - 3:00 PM Thursday March 23rd, 2023

This is a **closed-book** test, with one exception: you may use two A4 (8.5 inch x 11 inch) pieces of paper with any notes you wish to write on it (both sides, hand-written or typed).

You may use a scientific or graphing calculator, if it is only used for numerical calculations (i.e., no pre-stored information). No other electronic devices are allowed. Be sure to read each question carefully and **provide all the information requested**.

Show your work and explain your answer where necessary.

Exams must be turned in by 3:00pm sharp. You must show us your access card or another form of ID for identification. Before leaving the room, you must make sure we have your exam. Otherwise, you will receive a zero score on the exam and fail the class.

To keep track of time, aim to spend no more than 1.5 minutes per point value of the question. (i.e., if a question is worth 10 points, try not to spend more than 15 minutes.) Raise your hand to ask any clarifying questions. Take a deep breath and good luck! Thanks for an enjoyable class! Have a great spring break!

Q1	[8]	Q2	[6]	Q3	[6]	Q4	[6]
Q5	[14]	Q6	[10]	Q7	[4]	Q8	[8]
Q9	[12]	Q10	[12]	Q11	[4]	Q12	[20]
Q13	[8]	Q14	[12]	Q15	[10]	Q16	[10]
EC	[???]	The practice final has 150 points, but the actual final will only have 100 points.					

Q1. Ryan goes on a Santa Barbara food crawl and rates the restaurants he visits a unique numerical score between 15 and 115. He classifies a restaurant as satisfactorily good if the score is above 100.

(a) [4'] Explain how the precision and recall of Ryan classification ability to identify all the Michelin star restaurants in Santa Barbara would change if Ryan decreases the threshold for a satisfactory good restaurant?

(b) [4'] If Ryan's goal is to maximize the F1-score for his classification problem, please explain how Ryan should change his decision threshold based on observations of precision and recall.

Q2. Before Danny graduates, he aspires to reach a number of citations on his research papers. For each of the following measures, qualitatively state in terms of papers and citations what Danny's primary goal is.

(a) [2'] L0 norm

(b) [2'] L1 norm

(c) [2'] L-infinity norm

Q3. The XOR operation returns a positive classification if exactly one of the inputs is nonzero and returns false otherwise. Explain whether each of the following models can achieve perfect accuracy.

(a) [2'] a basic perceptron without any activation function

(b) [2'] a multi-layer perceptron without any activation functions

(c) [2'] a soft-margin support vector machine

Q4. The training data for a machine learning algorithm comprises these {x, y} points:

{(2, 3), +1}, {(-1, 4), +1}, {(3, 0), −1}, {(2, -1), −1}

(a) [3'] Compute the gram matrix for this dataset.

(b) [3'] Compute the scatter matrix for this dataset.

Q5. Suppose we are given a feature vector **x** containing **d** features and wish to predict its real-valued regression label **y**.

(a) [4'] The R^2 metric measures the goodness of fit of a linear model. Andy trained a regression model that has a negative R^2 value during training. Explain whether Andy's model underfitting, overfitting, or fitting well? (Recall if the model outputs the mean label value for all inputs, the R^2 value will be 0. If the model correctly outputs all label values for all inputs, the R^2 value will be 1.)

(b) [2'] State the ordinary least squares loss function **L(w)** using matrix and norm notation.

(c) [6'] Derive the optimal weights for the least squares regressor by computing $\nabla_w L(w)$.

Hint 1: $\nabla_w(w^T A w) = (A + A^T)w$

Hint 2: $\nabla_w(w^T b) = b$

(d) [2'] What does the existence of a closed-form solution imply about the convergence of a linear regression model optimized using gradient descent?

Q6. Suppose we are given a feature vector **x** with **n** dimensions for the classification problem with **C** classes.

(a) [2'] Provide a mathematical formulation for the Naive Bayes classifier.

(b) [2'] Rewrite the formulation in (a) to use LaPlacian correction.

(c) [2'] Explain why the Naive Bayes simplifying assumption is necessary.

(d) [2'] Explain whether conditional independence implies independence.

(e) [2'] Explain whether independence implies conditional independence.

Q7 [4'] The continuous bag of words neural network model takes as input a **1 x C** representation of a word and maps it to a hidden layer of size **M**; from this hidden layer, this model outputs a **1 x V** dense representation of the input word. How many learnable parameters does this model architecture contain? (Recall a learnable parameter is a weight or bias.)

Q8. Consider an unweighted k-nearest neighbor classifier using the Euclidean distance metric on a binary classification task.

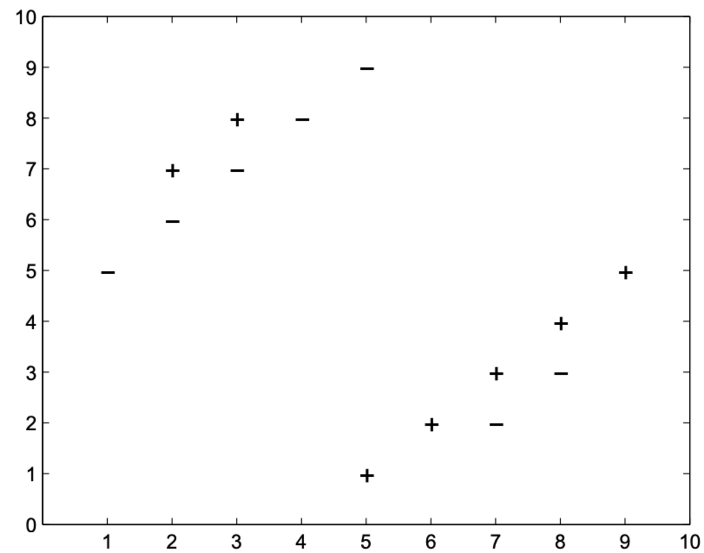


Figure 1: Dataset for KNN binary classification task.

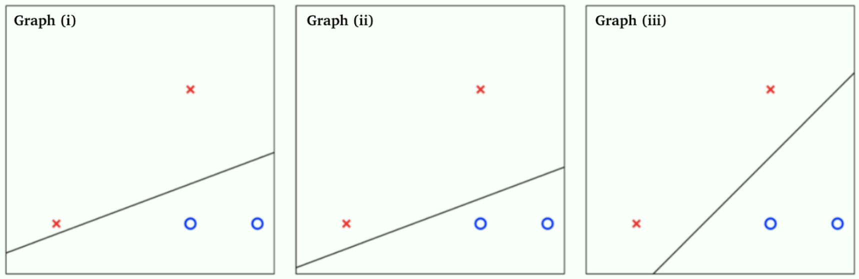
(a) [2'] What value of k minimizes the training error for this dataset?

(b) [2'] Why might using too large values of k be bad for this dataset?

(c) [2'] Why might using too small values of k be bad for this dataset?

(d) [2'] In the figure above, sketch the decision boundary for the 1-nearest neighbor for this dataset.

Q9. Suppose we want to linearly separate four data points, with negative instances denoted with a red x and positive instances denoted with a blue circle. The graphs below show 3 possible decision boundaries that linearly separate these four data points.



(a) [3'] For each graph, would the decision boundary be valid for a converged perceptron? What about a converged hard-margin SVM? (Select all that apply, if any.)

(b) [6'] Maximum Margin Classifier and Support Vector Machine are two names for the same model. Explain how said model gets its name?

(c) [3'] In a soft-margin SVM, how does increasing the tuning constant C affect the chance of having misclassifications in your training set?

Q10 Vaish is performing experiments with the WikiHow dataset containing 40000 training articles and 6000 testing articles.

(a) [2'] In 5-fold cross validation, how many times is one particular article used as a training example and validation example?

(b) [2'] In the boosting ensemble method with 20 time steps, how many times is one particular article used as a training example?

(c) [4'] In the bagging approach to build 25 classifiers using 2000 articles each, what is the expected number of times one particular article will be used to build the classifier?

(d) [4'] In the gradient descent algorithm, how many examples are used during one update step? What about the stochastic gradient descent algorithm?

Q11. [4'] David has a feature that is perfectly uniform and linearly separable. If he builds a decision tree, prove that the information gain is maximized if he splits upon this feature.

Q12. Consider the setting for the following neural network below.

Let's consider a simple two layer neural network for binary classification.

$$z_1 = W_1 x^{(i)} + b_1$$
$$a_1 = ReLU(z_1)$$
$$z_2 = W_2 a_1 + b_2$$
$$\hat{p}^{(i)} = \sigma(z_2)$$
$$L^{(i)} = y^{(i)} * \log(\hat{p}^{(i)}) + (1 - y^{(i)}) * \log(1 - \hat{p}^{(i)})$$
$$J = -\frac{1}{m}\sum_{i=1}^{m} L^{(i)}$$

It has input size 2×1, one hidden layer size 3×1, and output size 1. $\sigma(\cdot)$ is the Sigmoid function.

The weights of the network are: $W_1 = \begin{bmatrix} -0.2 & 0.1 \\ 0.5 & 0.3 \\ 0.35 & 0.4 \end{bmatrix}$, $W_2 = \begin{bmatrix} 1.5 & 0.6 & -0.4 \end{bmatrix}$,

and biases $b_1 = \begin{bmatrix} 1.2 \\ 0 \\ -0.7 \end{bmatrix}$, $b_2 = \begin{bmatrix} 0.5 \end{bmatrix}$,

Assuming the network takes an input of $x = \begin{bmatrix} 1.5 \\ -0.8 \end{bmatrix}$,

and the ground-truth label $y = 0$.

(a) [2'] Draw the architecture of this neural network.

(b) [4'] What is the neural network prediction p-hat?

(c) [14’] Compute the partial derivative of **J** with respect to $\mathbf{W_1}$. Be careful with the shapes.

Q13. Suppose we have one batch of 100 input images each of size 3×64×64. Consider a convolutional layer with 2 output channels, kernel size 5×5, no padding, and stride of 2. Answer the following questions and give brief explanations for your answers.

(a) [2'] What is the shape of the weight parameters for the convolutional layer?

(b) [4'] What is the output size after we feed the batch of input images through the convolutional layer?

(c) [2'] We decide to add a linear layer after the convolutional layer to make a prediction of whether the image contains a pedestrian. What would be the input dimension for the linear layer?

Q14. These are some classic questions you may get asked during a machine learning interview.

(a) [2'] What is the vanishing gradient issue within RNNs?

(b) [2'] What is the primary purpose of a convolution in a convolution neural net using kernel size > 1?

(c) [2'] What is the purpose of a batch normalization layer in a convolutional neural network?

(d) [2'] How does the attention model help improve parallelism over RNNs/LSTMs?

(e) [2'] What is the autoregressive property?

(f) [2'] What is the primary difference between these two groups of activation functions?

Group 1: ReLU, SeLU, GeLU, CeLU; Group 2: tanh, sigmoid

Q15. Consider a specific data point **x(t)** with **d** dimensions and a binary label **y(t)** that is +1 when positive and -1 when negative as a misclassified example by the model **w(t)** with **d** dimensions at iteration **t** in the perceptron learning algorithm.

(a) [6'] Show that $\mathbf{y(t)\ w(t)^T\ x(t) < 0}$.

(b) [4'] Explain the intuition of the perceptron update step geometrically.

Q16. Consider an ensemble method that uses the bagging approach to build several decision tree classifiers, also known as a random forest. In this version, we modify the bagging algorithm to additionally select a subset of the features in addition to selecting a subset of the data. For each of the statements below, state whether they are **sometimes**, **always**, or **never** true and provide a brief justification of your choice.

(a) [2'] Random forests achieve higher accuracy than a single decision tree.

(b) [2'] The hypothesis space of random forests is equivalent to the hypothesis space of a decision tree.

(c) [2'] One of the classifiers in the random forest is equivalent to the singular decision tree built by the ID3 best split algorithm.

(d) [2'] Random forests are as interpretable as a single decision tree.

(e) [2'] Random forests overfit as easily as a single decision tree.

Just for Fun! If you care enough to answer some or all of these questions, you should consider AI/ML research!

(a) Describe a process that leverages the knapsack algorithm to filter low quality sentences for the text summarization task.

(b) Provide precise definitions and clear examples to highlight the differences between overtly unsafe, covertly unsafe, and indirectly unsafe language.

(c) In the context of natural language, what is a foveation and how can foveations be used for data augmentation?

(d) Explain how a novelty and consistency driven approach to generating multimodal infillings can intuitively guarantee that these infillings can help bridge logical gaps.

(e) Discuss how using deception as a means can help a data owner achieve their desired end of demonstrating regulatory compliance and how this impacts the ability for users and stakeholders to rationally, autonomously, and confidently decide whether a particular AI system is trustworthy.

This page is intentionally left blank for additional space.

This page is intentionally left blank for additional space.

This page is intentionally left blank for additional space.