University of Waterloo
Cheriton School of Computer Science

CS 480/680– SAMPLE Midterm Examination
Introduction to Machine Learning, Fall 2024

Instructor: Kathryn Simone

29 October 2024

Name: ______________________________

ID Number: ______________________________

This exam contains 10 pages (including this cover page) and 5 questions. Total of points is 100.

Distribution of Marks

Question	Points	Score
1	40	
2	10	
3	10	
4	10	
5	30	
Total:	100	

Part I: Multiple Choice

1. (a) (5 points) Select all that apply: A Gaussian Process is completely specified by its _____

    ○ Mean and Covariance

    ○ Covariance, assuming an isotropic Gaussian prior for the weights

    ○ Covariance, assuming a Multivariate Gaussian prior for the weights centered at the origin

    ○ Covariance, assuming a Gaussian prior for the weights with a mean vector $\boldsymbol{\mu} = \mathbf{0}$

   (b) (5 points) After solving the dual problem of the hard-margin Support Vector Machine (SVM), we need to compute the parameters of the decision boundary. Assume the primal form of the SVM problem involves maximizing the margin by finding the weight vector $\mathbf{w}$ and the bias term b. Which of the following combinations provides the *minimum necessary* information to compute both the weight vector $\mathbf{w}$ and the bias b for the hard-margin SVM?

    ○ The Lagrange multipliers λ_i, the data labels y_i, and the feature vectors x_i of all training examples.

    ○ The subset of nonzero Lagrange multipliers λ_i, the corresponding data labels y_i, and the feature vectors x_i of the support vectors only.

    ○ The Lagrange multipliers λ_i, the feature vectors x_i of the support vectors, and the distance from the decision boundary to the support vectors.

    ○ The subset of zero Lagrange multipliers λ_i, and the data points that are not support vectors.

   (c) (5 points) Least-squares regression can be derived from _____ under the assumption of _____.

    ○ Maximum a posteriori estimation, an isotropic Gaussian for the prior distribution of the weights

    ○ Ridge regression, regularization hyperparameter $\lambda = 1.0$

    ○ Maximum likelihood estimation, normally-distributed errors

    ○ Maximum likelihood estimation, normally-distributed weights

   (d) (5 points) Which of the following is/are strong arguments for the interpretability of decision trees? Select all that apply.

    ○ Because they mimic human reasoning.

    ○ Because the decision path for each input can be easily traced and understood.

    ○ Because each decision is based directly on the input features of the data.

    ○ Because small changes in the training data result in large changes in the structure of the tree.

   (e) (5 points) Which of the following is/are true about gradient descent? Select all that apply.

    ○ It updates parameters in the opposite direction of the gradient of the loss function.

    ○ It finds the global minimum of non-convex functions.

    ○ It does not require learning rate tuning.

○ It performs well even with large batch sizes.

(f) (5 points) What is the purpose of the kernel trick in machine learning?

○ To compute inner products between vectors in a higher-dimensional space without explicitly mapping the data.

○ To apply dimensionality reduction to the data.

○ To regularize the decision boundary.

○ To minimize the loss function.

(g) (5 points) What happens to the KNN model as k increases?

○ The model becomes less sensitive to noise but may underfit.

○ The model becomes more sensitive to noise and may overfit.

○ The computational complexity decreases.

○ The decision boundaries become more flexible.

(h) (5 points) Which of the following are linear models for binary classification?

○ Logistic regression

○ k-Nearest Neighbors

○ Kernel SVM

○ Perceptron

○ Decision tree

Part II: Long-Answer

2. *Loss Functions and Optimization.* The log-cosh loss function is a smooth approximation of the absolute loss, defined as:

$$l(y, \hat{y}) = \log(\cosh(\hat{y} - y))$$

Where y is the true response, $\hat{y} = w^T x + b$ is the predicted response for some feature vector x given model parameters w and b, and $\log(\cosh(x))$ is a differentiable function.

(a) (3 points) Derive the gradient of the loss with respect to the parameters w and b. The following identities may be useful to you:

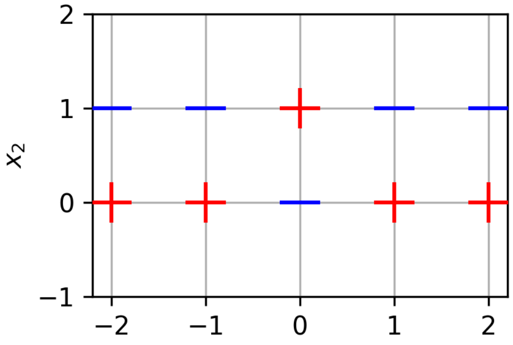
$$\frac{d}{dx}\cosh(x) = \sinh(x),\ \frac{\sinh(x)}{\cosh(x)} = \tanh(x),\ \frac{d}{dx}\log(x) = \frac{1}{x},\ \frac{\partial y}{\partial x} = \frac{\partial y}{\partial u}\frac{\partial u}{\partial x}$$

(b) (4 points) Formulate a gradient-descent-based algorithm to solve this optimization problem for the parameters. Refer to the gradient of the loss with respect to a parameter θ as $\nabla_\theta l$.

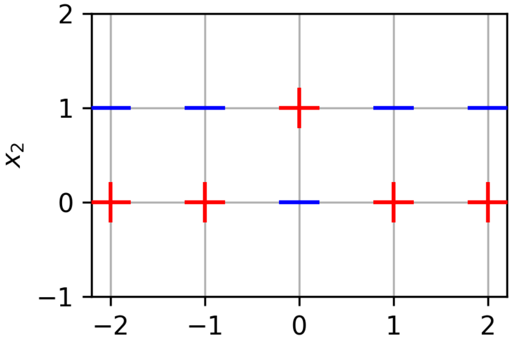
(c) (3 points) What advantage(s) would you anticipate for using the log-cosh loss, as compared to the mean-squared error studied in class?

3. *Classifiers* Draw the decision boundaries learned by each of the following classifiers, where relevant. Mark each region as labeled + or -, and assume ties are broken arbitrarily.

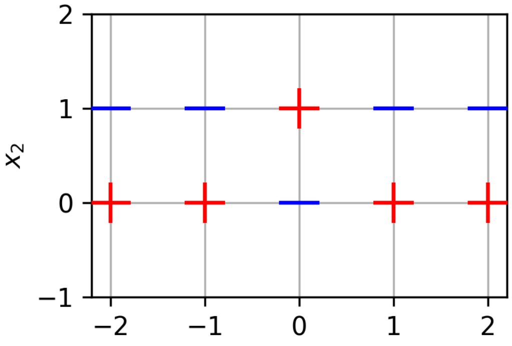
(a) (1 point) Perceptron:



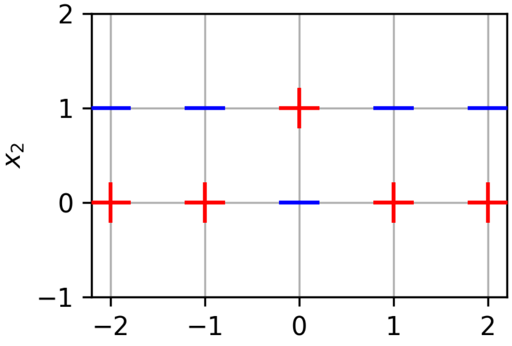
(b) (2 points) Logistic Regression:



(c) (3 points) 1-Nearest Neighbors:

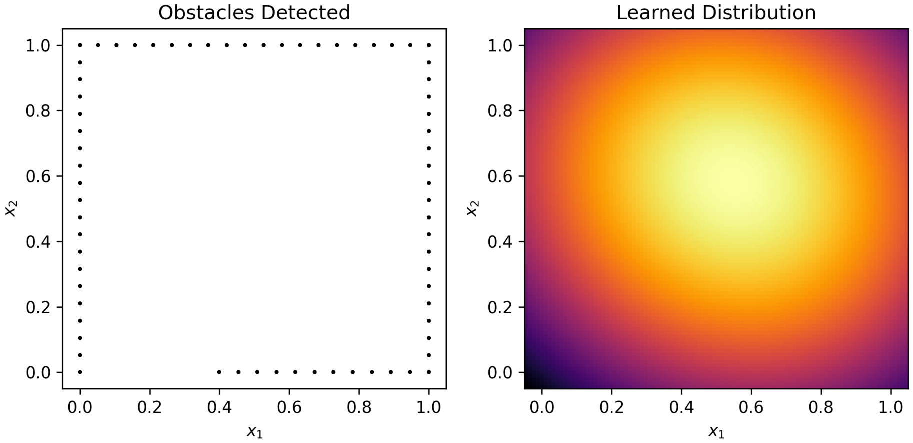


(d) (4 points) 3-Nearest Neighbors:



4. *Kernel Density Estimation*

(a) (4 points) The figures below depict the position of obstacles detected by a robot navigating in 2D space (left) and the learned distribution of obstacles (right).



A Gaussian kernel,

$$k(x, x') = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-x')^2}{2\sigma^2}}$$

was used, where σ^2 is the variance and σ is the scale or bandwidth parameter. Comment on the learned distribution and provide a recommendation to improve the approximation, if possible.

(b) (3 points) The normalized sinc function,

$$\text{sinc}(\|x - x'\|) = \frac{\sin(\pi(\|x - x'\|))}{\pi(\|x - x'\|)}$$

is significant across many fields of science and engineering. Is this a valid density kernel? Explain why or why not, with appropriate reference to the necessary properties of a density kernel.

(c) (3 points) Formulate an approach to transform the sinc function into a valid kernel.

Part III: True or False

5. For each question, answer whether it is true or false, and provide a brief (1 sentence) justification for your answer.

   (a) (5 points) Increasing the regularization hyperparameter in SVM increases the size of the margin.

   (b) (5 points) Unlike logistic regression, one can always derive a closed-form solution for the optimal parameters for linear regression.

   (c) (5 points) The SVM algorithm is guaranteed to find the global optimum for both linear and non-linear classification problems.

   (d) (5 points) K-Means can only find spherical clusters.

   (e) (5 points) Maximum likelihood estimation minimizes the distance between the estimated and true parameter values.

   (f) (5 points) Any local minimum of a convex function is also a global minimum.

This page is intentionally left blank to accommodate work that wouldn't fit elsewhere and/or scratch work.