CS 6375: Machine Learning Midterm Examination

University of Texas at Dallas

11/04/2024

Question	Topic	**Points**
1	Short Answers	50
2	Probabilistic Methods: True/False	20
3	Linear Regression and Loss Functions	30
Total		100

Instructions:

1. You have **ONE (1) hour, Fifteen (15) Minutes** to complete the examination.
2. Please show all the steps clearly of how you came up with the final answer. Just the final answer with no work will be zero points!!
3. Either you can use this paper or a separate set of sheets to fill in your answers. Write clearly so we can understand your handwriting.
4. Please order your questions according to the questions and write in clear handwriting so it is easy for us to grade. Otherwise we will deduct 10 points.
5. Please do not search online for answers to the questions. If the answers are similar to something available online, you will get zero points on this examination.
6. The examination has to be done individually by everyone. If someone copies, the entire group of students involved will get a zero.
7. Work efficiently. Some questions are easier, some more difficult. Be sure to give yourself time to answer all of the easy ones, and avoid getting bogged down in the more difficult ones.
8. All the Best!!

Question 1: Short Answers

[50 pts] Please provide short and clear answers for the questions below. Please explain your answer and show your work. No credit if the explanation is incorrect. Each question below is 5 points each.

(a) Explain the purpose of using regularization in linear regression models. What are the key differences between L1 and L2 regularization? Write down the formulation of the L1 and L2 regularized Linear Regression for a dataset $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(N)}, y^{(N)})\}$. Assume w and b are the weight and bias parameters.

(b) What is the main limitation of the perceptron algorithm? Describe a type of dataset that a single-layer perceptron cannot classify accurately. Under what conditions is the perceptron algorithm guaranteed to converge? What alternative model might you consider if these conditions are not met?

(c) For linearly separable data, can a small slack penalty ("C") hurt the training accuracy when using a linear SVM (with no kernel)? If so, explain how and if not, why not?

(d) Why can the k-nearest neighbors (k-NN) algorithm struggle with high-dimensional data? Say, I have a 100-dimentional problem with a 1000 instance dataset. What happens with nearest neighbor methods? Discuss the potential limitations of using Euclidean distance as a similarity measure in k-NN and how will you fix it.

(e) Explain how the choice of the splitting criteria affects the decision tree performance and what are aspects to consider in coming up with a splitting criteria. Please provide a splitting criteria other that those seen in class (conditional entropy, information gain, gini coeffient, mean square reduction) for classification or regression.

(f) Describe how the choice of prior affects the posterior distribution in Bayesian methods, especially in cases of

small data. What happens to the influence of the prior as the data size increases?

(g) Why is Naive Bayes considered a "naive" algorithm? Provide an example where its core assumption might lead to poor performance. Why does Naive Bayes often perform surprisingly well even if its assumptions are violated?

(h) Describe how logistic regression handles multi-class classification. What are one-vs-rest and softmax approaches?

(i) Explain how the initial choice of centroids can impact the outcome of k-means clustering. How can one mitigate

the issues arising from poor centroid initialization?

(j) A random variable follows an exponential distribution with parameter $\lambda : \lambda > 0$, and has the following density:

$$p(t) = \lambda e^{-\lambda t}, t \in [0, \infty] \tag{1}$$

This distribution models waiting times between events. Given a iid data: $T = (t_1, \cdots, t_n)$, where each t_i is modeled as drawn from the exponential distribution with parameter λ. Then Compute the log-likelihood $p(T|\lambda)$ and Solve for λ_{MLE}

Question 2: Probabilistic Methods: True/False Questions

[20 pts] Below are some conceptual True/False questions on probabilistic methods. **You need to explain your answer in 2-3 sentences and you cannot just say true or false.** Each True/False question is for two points and there are ten questions below.

(a) Naive Bayes classifiers require that all predictors (features) be conditionally independent given the class label; if this assumption is violated, the model cannot be applied.

(b) The decision boundary of a logistic regression model is a Sigmoid Function in the feature space.

(c) A Naive Bayes Model can be used in chat-gpt to generate text because it is a generative model.

(d) In Bayesian statistics, the prior and posterior distributions must always belong to the same family of distributions for computational convenience, a property known as conjugacy.

(e) Naive Bayes classifiers can be used for regression tasks by predicting a continuous value instead of class labels.

(f) Bayesian inference requires larger datasets to be effective compared to frequentist methods because it has to update priors with new data.

(g) A logistic regression model can incorporate polynomial features to model non-linear relationships between the independent variables and the log odds of the dependent variable.

(h) A high prior variance in Bayesian inference reflects strong initial beliefs about the values of the parameters before observing any data.

(i) Regularization techniques, such as L1 and L2 regularization, are not applicable in logistic regression models because these models are inherently resistant to overfitting due to their probabilistic nature.

(j) Given a Logistic Regression Model learnt to distinguish spam and not spam emails, I can generate a spam email.

Question 3: Linear Regression and Loss Functions

[30 pts] Consider a regression task that fits a piece wise linear function of the form: $f(x) = a_1 x + b_1, x < 0$ and $f(x) = a_2 x + b_2$ if $x \geq 0$.

(a) Part 1 (10 points): Given data points $(x^{(1)}, y^{(1)}), \ldots, (x^{(M)}, y^{(M)})$, where $x^{(m)} \in \mathbb{R}$ and $y^{(m)} \in \mathbb{R}$, formulate a regression problem to predict y as a loss minimization problem and explain how to use gradient descent.

(b) Part 2 (8 points) If we want to apply an additional constraint that f must be continuous, then formulate regression under this new constraint as a convex optimization problem.

(c) Part 3 (12 points) Below is a loss function called the Huber Loss (y is the true label and $\hat{y}$ is the predicted label):

$$L_\delta(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2 & \text{if } |y - \hat{y}| \leq \delta, \\ \delta \cdot |y - \hat{y}| - \frac{1}{2}\delta^2 & \text{if } |y - \hat{y}| > \delta. \end{cases}$$

Below are a few questions with respect to this loss:

(a) Would you use this for classification or regression? Is this a valid loss function for classification and regression? Provide justification.

(b) Is this loss function differentiable? Please provide the gradient of this loss function.

(c) What are some of the features and benefits of this loss function compared to MAE and MSE Losses? How will it work if there are outliers in the dataset compared to just using the MSE loss function?