CS 189/289A Introduction to Machine Learning
Fall 2024 Jennifer Listgarten, Saeed Saremi Final

- Please do not open the exam before you are instructed to do so.
- **Electronic devices are forbidden on your person**, including cell phones, tablets, headphones, and laptops. Leave your cell phone off and in a bag; it should not be visible during the exam.
- The exam is closed book and closed notes except for your two 8.5 × 11 inch cheat sheets.
- You have 2 hours and 50 minutes (unless you are in the DSP program and have a larger time allowance).
- Please write your initials at the top right of each page after this one (e.g., write "JD" if you are John Doe). Finish this by the end of your 2 hours and 50 minutes.
- Mark your answers on the exam itself in the space provided. Do **not** attach any extra sheets.
- For multiple choice questions, fill in the bubble for the single best choice.
- For short and long answer questions, write within the boxes provided. If you run out of space, you may use the last four pages to continue showing your work.
- **The last question is for CS289A students only**. Students enrolled in CS189 will **not** receive any credit for answering this question.

Your Name	
Your SID	
Name and SID of student to your left	
Name and SID of student to your right	

○ CS 189

○ CS 289A

This page intentionally left blank.

□

initial here

1 Multiple Choice

For the following questions, select the **single best response**. Each question is worth **1.5 points**.

1. Which of the following statements about logistic regression is **false**?
   - ○ Logistic regression models the log-odds as a linear function of the features.
   - ○ The output of a logistic regression model is always between 0 and 1.
   - ○ The "steep" part of the sigmoid function (inputs around -0.1 to 0.1) can cause vanishing gradients.
   - ○ Logistic regression can be used for classification.

2. Which of the following statements about Decision Trees is **true**?
   - ○ The predictions of two Decision Trees in a Random Forest are completely independent of each other.
   - ○ A single Decision Tree in a Random Forest always overfits on the random subset of the data it is trained on.
   - ○ A common Decision Tree training objective is minimizing the misclassification rate.
   - ○ Decision Trees do not benefit from input normalization, as they are invariant to the scale of the features.

3. Which of the following statements about k-means clustering is **true**?
   - ○ k-means requires fewer parameters than Mixture of Gaussians.
   - ○ k-means labels new points using the majority label of the k closest points.
   - ○ k-means can accurately identify clusters of arbitrary shapes and densities.
   - ○ k-means assigns each data point to a cluster by minimizing the pairwise distances between all data points.

4. Bob is running Langevin MCMC to sample from a distribution $p(x) \sim N(\mu, \Sigma)$ using its score function, where $x \in \mathbb{R}^2$. To improve the algorithm's convergence, he hyperparameter tunes the step size $\eta = [\eta_x, \eta_y]^T$ such that $\eta_x > \eta_y$. What is most likely to be **true** regarding the shape of the target distribution $p(x)$?
   - ○ The mean $\mu = [1, 2]^T$ and covariance $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
   - ○ The mean $\mu = [2, 1]^T$ and covariance $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
   - ○ The mean $\mu = [0, 0]^T$ and covariance $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$
   - ○ The mean $\mu = [0, 0]^T$ and covariance $\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$

initial here

5. Two binary classifiers, f_1 and f_2, are trained on the same dataset but evaluated on separate datasets. f_1 is evaluated on dataset $\mathcal{D}_1$, where 30% of the instances are from the positive class. f_2 is evaluated on dataset $\mathcal{D}_2$, where 70% of the instances are from the positive class. Both models achieve an Area Under the Receiver Operating Characteristic (AUROC) of 0.8 on their respective datasets. Which of the following statements is **true**?

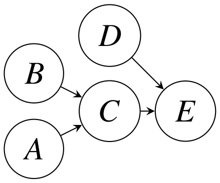
   ○ Compared to f_1, f_2 has a higher probability of correctly ranking a randomly chosen positive instance above a randomly chosen negative instance.

   ○ Both classifiers have identical true positive rates and false positive rates at all classification thresholds.

   ○ The AUROC of the two classifiers cannot be compared because the two evaluation datasets have different proportions of positive instances.

   ○ Both classifiers have the same capability to discriminate between positive and negative examples.

6. Consider the graphical model displayed below



   Which of the following independence statements **does not always hold true**?

   ○ A and B are marginally independent.

   ○ B and E are conditionally independent given C.

   ○ A and D are marginally independent.

   ○ A and B are conditionally independent given C.

7. Consider a Hidden Markov Model with hidden states $X_1, \ldots, X_T$ and corresponding observations $Y_1, \ldots, Y_T$ for each time step t in 1 to T. Assume there are N possible hidden states.

   The Viterbi algorithm uses dynamic programming to determine the most probable sequence of hidden states given the observations. Which of the following statements about the Viterbi algorithm is **false**?

   ○ The Viterbi algorithm initializes the dynamic programming table using the initial state distribution $P(X_1)$ and the emission probability $P(Y_1 \mid X_1)$.

   ○ The most probable hidden state at time t depends only on the observations from Y_1 to Y_t and not the observations from Y_{t+1} to Y_T.

   ○ The Viterbi algorithm maintains backpointers at each time step to facilitate the reconstruction of the most probable state sequence after processing all observations.

   ○ The Viterbi algorithm has a computational complexity of $O(N^2 \cdot T)$, where N is the number of hidden states and T is the length of the observation sequence.

8. Which of the following statements about the discounting factor γ in reinforcement learning is **true**?

   ○ A discount rate $\gamma > 1$ can be used to emphasize future rewards more than immediate rewards.

   ○ Lowering the discount rate can encourage an agent to prioritize short-term rewards over long-term rewards.

   ○ The optimal discount rate can be chosen via Maximum Likelihood Estimation.

   ○ When computing returns, the discount rate only applies to terminal rewards (attained by landing in a terminal state) and not any intermediate rewards.

9. Which of the following statements about optimal policies for finite MDPs is **false**?

   ○ All optimal policies achieve the same state-value function $v(s)$.

   ○ All optimal policies achieve the same action-value function $q(s, a)$.

   ○ If all rewards are doubled, the set of optimal policies changes.

   ○ There always exists an optimal policy for a finite MDP that is completely deterministic (a deterministic policy picks exactly one action in each state).

10. Which of the following statements is **false** regarding kernel methods:

    ○ Kernel methods can only be applied to supervised learning tasks.

    ○ Kernel methods implicitly embed data points in potentially infinite-dimensional spaces.

    ○ Kernel methods can be applied to non-conventional input spaces, such as graphs or strings.

    ○ Kernel methods scale at least quadratically with the number of training samples.

11. Consider the ridge regression objective $J(w) = ||Xw - y||_2^2 + \lambda||w||_2^2$ for some $\lambda > 0$. Let w^* denote the minimizer of the above expression. Which of the following is **true**?

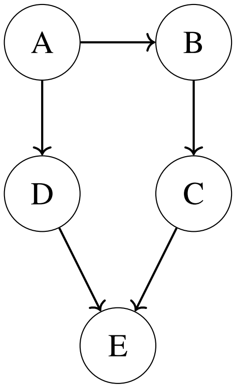
    ○ $Xw^* = y$.

    ○ w^* exists if and only if $X^T X$ is invertible.

    ○ $w^* = X^\dagger y$, where $X^\dagger$ is the pseudo-inverse of X.

    ○ The minimizer w^* is unique.

12. Which of the following is **true** of the following causal graph (graph corresponding to an SEM)?



○ E is a mediator between D and C

○ D and C are confounded by E.

○ A is confounded by E.

○ B and D are confounded by A.

13. Which of the following is **false** about graphical neural networks (GNNs)?

○ Choice of AGGREGATE and COMBINE functions determine the GNN.

○ A valid AGGREGATE function applied with a normalization based on the number of neighbors is still permutation invariant.

○ The node-level predictions of a GNN are permutation equivariant.

○ GNNs can be trained to distinguish non-isomorphic graphs that the Weisfeiler-Lehman (WL) test cannot.

14. You are training a convolutional neural network with five layers to classify different plant species. The model is initialized with random weights. Which of the following is most likely to show the strongest clustering of images by their class when t-SNE is applied?

○ The output of the **first** layer of the network **prior** to training.

○ The output of the **first** layer of the network **after** the model has been trained.

○ The output of the **fourth** layer of the network **prior** to training.

○ The output of the **fourth** layer of the network **after** the model has been trained.

15. In convolutional neural networks, max pooling is used to downsample feature maps. However, max pooling can also introduce challenges during training. Which of the following drawbacks of max pooling is **true**?

    ○ Max pooling increases the number of model parameters, thereby making the model more prone to overfitting.

    ○ Max pooling can disproportionately emphasize noisy activations, potentially leading to unstable feature representations.

    ○ Max pooling changes the activation functions of neurons from linear to non-linear, complicating the training process.

    ○ Max pooling is a non-differentiable operation, preventing the use of gradient-based methods for optimization.

16. Which of the following is **true** of regularization methods?

    ○ Ridge regression induces sparser solutions than LASSO.

    ○ Transforming data by using *all* the principal components of X instead of X itself is a form of regularization.

    ○ Regularization aims to reduce both the bias^2 and variance of a model.

    ○ Adding Gaussian noise to data can be interpreted as a form of regularization.

initial here

2 Short Answer

1. (3 points) Suppose we have n i.i.d. samples $X_1, X_2, \ldots, X_n$ drawn from the same distribution $X \sim \text{Poisson}(\theta)$. Our goal is to estimate the mean θ with the estimator $\hat{X}$. We define $\hat{X}$ to be:

$$\hat{X} = \frac{\alpha + \sum_{i=1}^{n} X_i}{\beta + n}$$

where α and β are constants greater than 0. What is the bias and variance of the estimator $\hat{X}$? *Hint:* The mean and variance of the Poisson distribution are the same, i.e., $\mathbb{E}[X] = \text{Var}[X] = \theta$.

(a) (1.5 points) What is the **bias** of the estimator $\hat{X}$?
Hint: The expectation of a Poisson(θ) random variable is θ.

(b) (1.5 points) What is the **variance** of the estimator $\hat{X}$?
Hint: The variance of a Poisson(θ) random variable is θ.

initial here

2. (2 points) Suppose you have three data points in $\mathbb{R}$:

$$\{x_1 = 0, x_2 = 6, x_3 = 12\}$$

Suppose you run k-means clustering with $k = 2$ and at a particular iteration observe the following clustering:

$$C_1 = \{x_2\}, C_2 = \{x_1, x_3\}$$

What are the centroids, and what is the total sum of squared errors? If this is the current cluster assignment, does the k-means algorithm terminate or keep going?

3. (2 points) Suppose a convolutional layer has the following specifications:

- Input dimensions: [height = 20, width = 20, channels = 3]
- Output dimensions: [height = 10, width = 10, channels = 10]
- Kernel size: [height = 5, width = 5]
- Each kernel also contains an additional bias term.

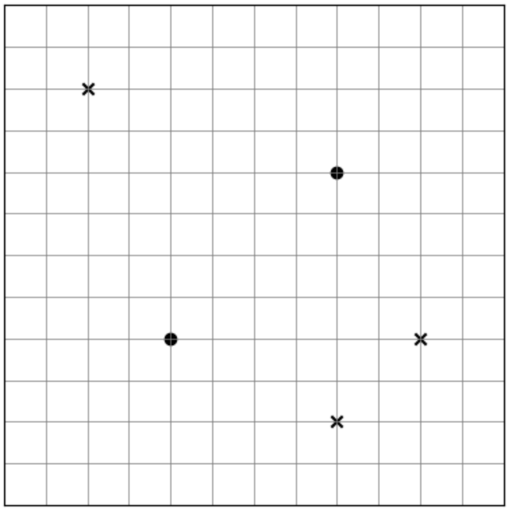
How many trainable parameters are there in this convolutional layer?

4. (2 points) Suppose that we have a dataset of points in $\mathbb{R}$ where each point comes from either class A or class B, both of which are one-dimensional Gaussians with equal prior probabilities:

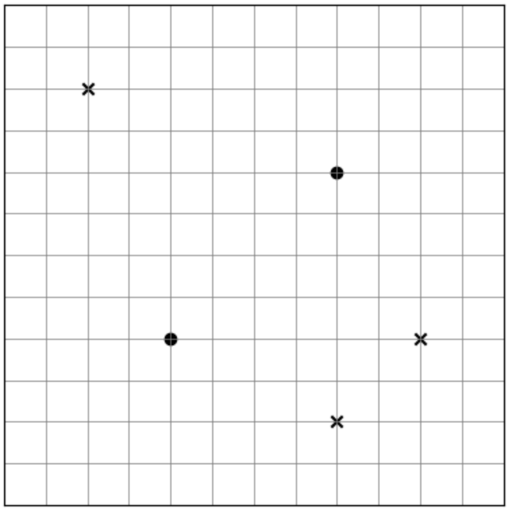
   - Class A: $\mu_A = 0$, $\sigma_A^2 = 1$
   - Class B: $\mu_B = 2$, $\sigma_B^2 = 4$.

   Determine the decision boundary $x \in \mathbb{R}$ where a point is equally likely to belong to class A or class B. You do not need to solve for x exactly. Instead express it as the solution of a quadratic equation $ax^2 + bx + c = 0$ for some $a, b, c \in \mathbb{R}$.

5. (2 points) Consider the following training dataset of points from two classes: dots and crosses. Draw the decision boundary that will be returned by a 1 Nearest Neighbor classifier. Assume that the distance metric being used here is regular Euclidean (ℓ^2) distance.



6. (2 points) Consider the same training dataset as above. Draw a plausible decision boundary that will be returned by a Decision Tree classifier with no bound on its max depth. You may assume that trees are trained using the same greedy procedure shown in lecture, that you also implemented in homework 5.



7. (2 points) Consider a dataset of two features f_1 and f_2 consisting of the six sample points

$$\begin{bmatrix} 4 & 6 & 9 & 1 & 7 & 5 \\ 1 & 6 & 5 & 2 & 3 & 4 \end{bmatrix}^\top$$

with labels

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}^\top .$$

We build a decision tree of depth 1 by choosing the single split that maximizes the information gain.

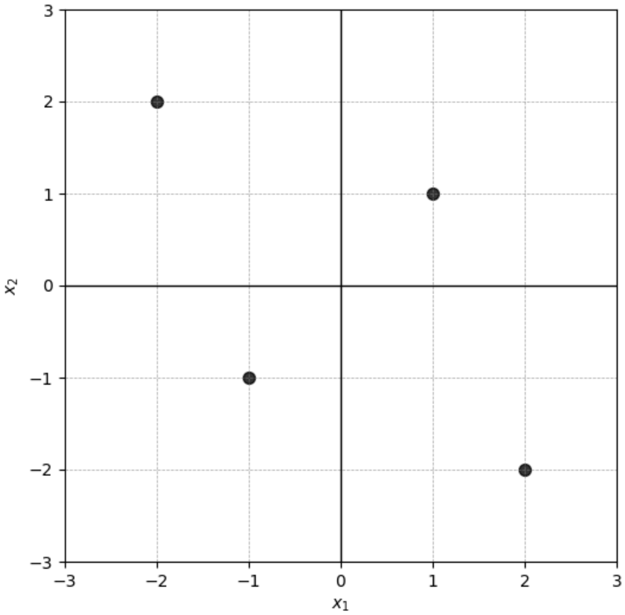
Define j as the feature and v as the value to split on at the root node (so if a value is $\geq v$, then we put it in the right tree, and if it is $< v$, then we put it in the left tree). For the optimal first (and only) split, what is the corresponding j and v?

8. (3 points) Consider an MDP with 3 states C (center), L (left) and R (right), and two actions l (left) and r (right). Here are the transition dynamics described qualitatively:

   - In state C, picking the left action l will transition you to the left state L, and will receive a reward of $+1$.
   - In state C, picking the right action r will transition you to the right state R, and will receive a reward of 0.
   - In state L, picking either action will transition you back to state C, and will receive a reward of 0.
   - In state R, picking either action will transition you back to state C, and will receive a reward of $+2$.

   Describe an optimal policy for this MDP when the discount rate is: (a) $\gamma = 0.0$, (b) $\gamma = 0.9$ and (c) $\gamma = 0.5$. You can reason about the policies intuitively (in fact, it might be helpful to draw this MDP) and you should not have to compute them algorithmically. Assume that an episode always starts in state C.

9. (2 points) Consider the following four points in $\mathbb{R}^2$.



(a) (1 point) We perform PCA on the data. What is one possible first principal component?

(b) (1 point) Suppose we wish to reconstruct the point

$$\begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

using the first principal component. That is, we want to compute the vector projection of our point onto the first PC. What is the resulting vector? (Your answer should be in $\mathbb{R}^2$.)

initial here

10. (3 points) Consider a hidden Markov model of the weather, with states $x_t \in \{\text{sunny}, \text{rainy}\}$ and observations $o_t \in \{\text{dry}, \text{wet}, \text{drenched}\}$, with the prior, transition, and emission probabilities:

x_1	$P(x_1)$
sunny	0.5
rainy	0.5

x_t	x_{t+1}	$P(x_{t+1} \vert x_t)$
sunny	sunny	0.8
sunny	rainy	0.2
rainy	sunny	0.4
rainy	rainy	0.6

x_t	o_t	$P(o_t \vert x_t)$
sunny	dry	0.9
sunny	wet	0.1
rainy	dry	0.2
rainy	wet	0.5
rainy	drenched	0.3

(a) (2 points) What is the probability of the sequence of observations (drenched, drenched, dry)? You may leave your answer as a product of numbers.

(b) (1 point) What is the most likely sequence of states given the sequence of observations (drenched, drenched, dry)?

initial here

11. (2 points) Suppose you are training a graph neural network (GNN) to predict whether a molecule is soluble or not. You represent each molecule as a graph where each atom is a node, and edges exist between nodes if there is a chemical bond between them. Each molecule in your dataset contains at most 100 atoms, and there always exists a path between any pair of nodes (i.e. the graph is connected).

    (a) (1 point) Determine the maximum number of GNN layers required so that information is shared between all pairs of nodes (even if there is not a direct edge between them) for any molecule in your dataset.

    (b) (1 point) **Note: this part is independent of the previous part.** In part (a), you found that your GNN requires a large number of layers to allow information to be shared between all pairs of nodes. If you constructed a GNN with this many layers and used a sigmoid activation function, what potential issue could arise during training? Name one strategy to prevent this without adding new edges between nodes.

initial here

3 MLE and MAP of the exponential distribution (6 points)

The probability density function of an exponential distribution is

$$f_{\text{exponential}}(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Here, $\lambda > 0$ is the parameter of the distribution.

(a) (2 points) Suppose that we observe n independent and identically distributed examples, denoted as $(x_1, x_2, \ldots, x_n)$, from an exponential distribution with parameter λ, where $x_i \geq 0$ for all i. Derive the log-likelihood function for the parameter λ given these observations.

(b) (2 points) Compute the maximum likelihood estimate of λ. You can assume without proof that the log likelihood is concave.

(c) (2 points) Suppose we have a prior on λ that it comes from a Gamma(α,β) distribution for given values of α and β. The probability density function of a Gamma distribution is

$$f_{\text{gamma}}(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}.$$

Show that the posterior distribution $p(\lambda \mid x_1, \ldots, x_n)$ follows a Gamma($n + \alpha$,$\beta + \sum_{i=1}^n x_i$) distribution.

Hint: You may show that the posterior is proportional to the density of the desired Gamma distribution.

initial here

4 Re-Learning Self-Attention (7 points)

OpenCorp is a technology company with a proprietary model that consists of a single self-attention layer with a single head. The company provides an API that returns the intermediate attention weights generated by the model for any given input. However, the model's original weight matrices remain undisclosed, and your goal is to re-learn the key and query weight matrices, W_k and W_q, solely by calling the API.

We will initialize the matrices W_k and W_q randomly and re-learn them using gradient ascent. We call the API n times to create a dataset, $D = \{X^{(i)}, A^{(i)}\}_{i=1}^{n}$, where i denotes the sample index. Each input $X^{(i)}$ is a sequence of L tokens, represented as a matrix whose jth row corresponds to the jth token. Each $A^{(i)} \in \mathbb{R}^{L\times L}$ is a matrix containing the corresponding ground-truth attention weights.

To train our weight matrices, we characterize the self-attention layer forward pass as follows

$$K^{(i)} = X^{(i)} W_k^{\top}$$
$$Q^{(i)} = X^{(i)} W_q^{\top}$$
$$\hat{A}^{(i)} = \text{Softmax}(Q^{(i)} K^{(i)\top})$$

where $\hat{A}^{(i)} \in \mathbb{R}^{L\times L}$ is the matrix of predicted attention weights computed using our current estimates for W_k and W_q, and the softmax is applied row-wise. We will interpret the attention weights $\hat{A}^{(i)}$ and $A^{(i)}$ as probabilities because they normalize to 1.

(a) (1 point) A naive way to learn W_k and W_q is by minimizing squared error between $A^{(i)}$ and $\hat{A}^{(i)}$, which are to be interpreted as probabilities. Briefly explain why minimizing mean squared error (MSE) might be a poor choice for an objective function in this scenario.

initial here

(b) (2 points) Write down a better objective function $\mathcal{L}(W_k, W_q)$ that we can maximize (rather than minimize) to recover W_k and W_q from D.
Hint: once again, attention weights $A_{pq}^{(i)}$ and $\hat{A}_{pq}^{(i)}$ can be interpreted as probabilities.

(c) (2 points) How would you modify your objective function if we want to explicitly encourage the queries for each token in a sequence to be orthogonal to each other?
Hint: formulate and add a penalty term to your objective from part (b). Denote λ to be the hyperparameter associated with the weight of this penalty term.

initial here

(d) (2 points) **Note: this subpart is independent of the previous subparts.**

After closer inspection, we realize that the learned attention weights are not causally masked such that $\hat{A}^{(i)}_{tl} = 0$ for $l > t$. Your friend from Stanford points out that one way to enforce causal masking is to multiply $\hat{A}^{(i)}$ with a mask matrix M, where $M_{tl} = 1$ for $l \leq t$ and 0 otherwise. Is this a valid strategy? If not, suggest one fix to make this masking strategy valid.

initial here

5 Naive Bayes classification (8 points)

The naive Bayes model is a generative model used for classification. We let X denote observed features, and X_i denote the i-th feature (out of d features in total). We let $C \in \{1, 2, \ldots, K\}$ denote the class label. The classifier can be represented using the following directed acyclic graph:

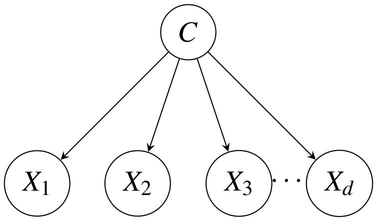


Figure 1: Naive Bayes Graph

(a) (1 point) Write the joint distribution $P(X_1, X_2, \ldots, X_d, C)$ of the graph in simplest form.

(b) (1.5 points) Let i, j denote two distinct feature indices. Are X_i and X_j necessarily conditionally independent given C? You may assume that all the features are discrete in this question. Correct answers with incomplete reasoning will not receive full credit.

initial here

(c) (1.5 points) Are X_i and X_j necessarily independent? You can again assume that all the features are discrete in this question.

(d) (3 points) Assume now that $P(C = k) = \frac{1}{K}$ for all k. We also assume that conditional on $C = k$, each X_i is i.i.d. from a Bernoulli distribution with parameter $\theta_k \in (0, 1)$ that is already known. Our goal is to classify a new observation based on its features, $X_1, X_2, \ldots, X_d$.

Express $P(C = k \mid X_1, X_2, \ldots, X_d)$ in terms of $\theta_1, \ldots, \theta_K$ and $X_1, \ldots, X_d$.

(e) (1 point) How would you classify a new observation based on its features? You can either write an optimization problem or explain your strategy (a single sentence should suffice).

initial here

6 Langevin MCMC for Logistic Regression (11 points)

Consider a logistic regression model, for $x,\ \beta \in \mathbb{R}^d$ and $y \in \{-1, 1\}$, where the likelihood is

$$p(y \mid x, \beta) = \sigma(y \cdot \beta^\top x)$$

where $\sigma(\cdot)$ represents the sigmoid function and the prior on the weights is Gaussian

$$\beta \sim N(0, \sigma^2 I)$$

In this question, we will explore why sampling from the posterior $p(\beta \mid x, y)$ is challenging. We will consider how Langevin Markov Chain Monte Carlo (MCMC) sampling offers a potential solution, using the score function $\nabla_\beta \log p(\beta \mid x, y)$.

(a) (2 points) Show that the posterior can be written as follows:

$$p(\beta \mid x, y) = \frac{p(y \mid x, \beta) p(\beta)}{p(y \mid x)}$$

Assume that β, x are independent.

initial here

(b) (2 points) Write the normalization term $p(y \mid x)$ in terms of $p(y \mid x, \beta)$ and $p(\beta)$. Explain why computing this normalization constant can be computationally intractable.

(c) (3 points) Compute the following gradients. (These will be useful to us in the following parts.)

i. $\nabla_\beta \log p(y \mid x, \beta)$

ii. $\nabla_\beta \log p(\beta)$

iii. $\nabla_\beta \log p(y \mid x)$

(d) (2 points) Using the previous part, write an expression for the score function $\nabla_\beta \log p(\beta \mid x, y)$. Why might the score function be easier to compute than the posterior?

(e) (2 points) Now let us use this score function $\nabla_\beta \log p(\beta \mid x, y)$ for sampling in Langevin MCMC.

i. Assume $x = 0$ and $y = 1$. Once the Langevin MCMC chain reaches its stationary distribution, what are the mean and variance of the samples?

ii. Assume $x = 1$ and $y = 1$. How does each term in the score function affect β?

initial here

7 (CS289A Only) Kernelized Nearest Neighbors (7 points)

Consider a dataset where each point X_i is a d-dimensional vector:

$$X_i = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}.$$

We would like to classify a query point X' using k-Nearest Neighbors (k-NN). However, instead of performing k-NN directly on the original features, we first transform the data using the function Φ that produces a quadratic embedding:

$$\Phi(X_i) = \begin{bmatrix} x_1^2 \\ \vdots \\ x_d^2 \\ x_1 x_2 \sqrt{2} \\ x_1 x_3 \sqrt{2} \\ \vdots \\ x_2 x_3 \sqrt{2} \\ \vdots \\ x_{d-1} x_d \sqrt{2} \\ x_1 \\ \vdots \\ x_d \end{bmatrix}.$$

(a) (2 points) In k-NN, we need to compute squared Euclidean distances between our query point X' and each point in our dataset. Fortunately, your friend Ruchir has found a kernel function k (not to be confused with the hyperparameter k in k-NN!) that works for the transformation Φ:

$$k(X_i, X_j) = (X_i^T X_j) + (X_i^T X_j)^2.$$

Write an expression for

$$||\Phi(X') - \Phi(X_i)||_2^2$$

only in terms of the kernel function k.

initial here

(b) (2 points) Recall that X' and X_i are both d-dimensional. What is the big-O complexity of computing $\|\Phi(X') - \Phi(X_i)\|_2^2$ without kernelization? What is the complexity using the kernel Ruchir found? You do not have to show work.

(c) (3 points) The RBF kernel is defined by

$$k(X_i, X_j) = \exp\left(-\frac{\|X_i - X_j\|_2^2}{2\sigma^2}\right).$$

For simplicity, we let $\sigma^2 = 1$ and $X_i \in \mathbb{R}$. Find a feature map Φ over an appropriate inner product space corresponding to this kernel. *Hint:* Recall the Taylor expansion:

$$e^X = \sum_{k=0}^{\infty} \frac{X^k}{k!}.$$

□

initial here

You may use this page to show extra work. Clearly mark your work with the problem number here, and also mention in the problem-specific box that your work is continued here.

initial here

You may use this page to show extra work. Clearly mark your work with the problem number here, and also mention in the problem-specific box that your work is continued here.

initial here

You may use this page to show extra work. Clearly mark your work with the problem number here, and also mention in the problem-specific box that your work is continued here.

initial here

You may use this page to show extra work. Clearly mark your work with the problem number here, and also mention in the problem-specific box that your work is continued here.