COMP 5212 Machine Learning (2024 Fall)

Homework 4:

Hand out: Nov 25, 2024

Due: Dec 2, 2024, 11:59 PM

Total Points: 72

Your solution should contain below information at the top of its first page.

1. Your name
2. Your student id number

Some Notes:

- Homeworks will not be easy, please start early.
- For this homework, please submit a single PDF file to Canvas. Make sure to prepare your solution to each problem on a separate page.
- The total points of this homework is 100, and a score of 100 already gives you full grades on this homework. There are possibly bonus questions, you can optionally work on them if you are interested and have time.
- You can choose either using LaTeX by inserting your solutions to the problem pdf, or manually write your solutions on clean white papers and scan it as a pdf file – in the case of handwriting, please write clearly and briefly, we are not responsible to extract information from unclear handwriting. **We highly recommend you use LaTeX for the sake of any misunderstandings about the handwriting.** If your submission is a scan of a handwritten solution, make sure that it is of high enough resolution to be easily read. At least 300dpi and possibly denser.
- We encourage students to work in groups for homeworks, but the students need to write down the homework solutions or the code independently. In case that you work with others on the homework, please write down the names of people with whom you've discussed the homework. You are not allowed to copy, refer to, or look at the exact solutions from previous years, online, or other resources.
- **Late Policy:** 3 free late days in total across the semester, for additional late days, 20% penalization applied for each day late. **No assignment will be accepted more than 3 days late**.
- Please refer to the Course Logistics page for more details on the honor code and logisitcs. We have zero tolerance — in the case of honor code violation for a single time, you will fail this course directly.

1 A Reflection of COMP 5212

Hi everyone, thank you for being with us for the whole semester. As the very last homework of this semester, let's revisit the stuff we have had fun with. This section contains 36 simple conceptual level multiple choice/True False questions (2 points each). Each one is directly taken from the materials used for lectures and homework. Each question has only one correct answer.

Please answer your questions here

Question	Choice	Question	Choice	Question	Choice	Question	Choice
1	D	2	D	3	B	4	B
5	C	6	B	7	A	8	B
9	B	10	A	11	C	12	B
13	A	14	B	15	A	16	A
17	C	18	D	19	D	20	B
21	A	22	A	23	C	24	D
25	B	26	B	27	D	28	B
29	A	30	D	31	B	32	B
33	A	34	A	35	B	36	C

Feel free to use other latex sources or submit hand-written versions, but please organize your solution sheet as our suggestion (i.e. put every four in a row). Thanks.

1. (Lecture 2) For linear regression, which of the following will never increase the least squares loss? *Hint: The bias term is the component of the weight vector associated with a constant feature for all samples.* $\sum (y - \theta^T x)^2$

   [A] Setting the bias term to zero or not fitting a bias term.

   [B] Augmenting the set of features used for the regression.

   [C] Projecting all samples onto a lower dimensional feature space with PCA before performing regression on the projected samples.

   [D] Subtracting the empirical mean from the data before performing regression on the centered samples.

2. (Lecture 3) For an exponential family distribution over random variable x, what is the variance of x?

   [A] The derivative of the sufficient statistic

   [B] The second-order derivative of the sufficient statistic

   [C] The derivative of the log partition function

   [D] The second-order derivative of the log partition function

3. (Lecture 3) In numerical optimization, Newton's Method is used to:

   [A] Find the derivative of a function at a given point.

   [B] Approximate the roots of a nonlinear equation.

   [C] Minimize a function by iteratively updating the parameter estimates.

   [D] Solve systems of linear equations.

4. (Lecture 4) Consider a dataset with the response variable Y and predictors X_1, X_2, and X_3. Y denotes the number of people passed to a bus stop in a fixed time frame. You are using a Generalized Linear Model (GLM) to analyze this data. Which of the following distributions would be most appropriate for modeling this scenario?

   [A] Normal distribution.

   [B] Poisson distribution.

   [C] Multinomial distribution.

   [D] Dirichlet distribution.

5. (HW1) When applying kernel methods in machine learning, such as the kernel SVM, why is the choice of kernel important and what does it determine in the context of model training?

   [A] The kernel determines the learning rate of the model and influences the speed of convergence during training.

[B] The kernel specifies the type of regularization applied to the model to prevent overfitting during training.

[C] The kernel maps the original data into a higher-dimensional space to potentially make it easier to find a linear separation.

[D] The kernel directly controls the size of the margin in the SVM model and dictates the spacing between the data points.

6. (Lecture 5) Which of the following is NOT the necessary condition for a valid kernel function K?

   [A] $K(x, x) \geq 0$

   [B] The kernel matrix is positive definite.

   [C] $K(x, y)$ can be written in the form as $\phi(x)^T \phi(z)$

   [D] The kernel matrix is symmetric

7. (Lecture 5) In SVMs, the geometric margin of a classifier to a data point is a critical concept in understanding the robustness of the model. For a separatable dataset, which of the following best describes the geometric margin in the context of SVMs?

   [A] The geometric margin is the distance between the nearest data point in the training set and the decision boundary, measured in the feature space.

   [B] It is equal to functional margin

   [C] The geometric margin is equivalent to the sum of the distances from all support vectors to the decision boundary, normalized by the norm of the weight vector.

   [D] It is defined as the distance from the decision boundary to the closest data point, normalized by the norm of the weight vector.

8. (Lecture 6 7) Which of the following is NOT true about the Generalized Lagrangian method?

   [A] Generalized Lagrangian is used to convert constrained optimization problems into forms that are easier to solve using standard techniques.

   [B] Generalized Lagrangian converts the original optimization problem to a dual optimization problem that does not have constraints.

   [C] Generalized Lagrangian introduces new variables to the optimization problem.

9. (Lecture 6 7) For a primal constrained optimization problem, with the objective as $f(w)$, and the constraints as $g_i(w) \leq 0, i = 1, \cdots, k$, $h_j(w) = 0, j = 1, \cdots, l$. The Generalized Lagrangian equation is $\mathcal{L}(w, \alpha, \beta)$ (α_i is the weight for inequality constraints, and β_j is the weight for equality constraints). What is $\max_{\alpha, \beta : \alpha_i, \beta_j \in \mathbb{R}} \mathcal{L}(w, \alpha, \beta)$ if w satisfies all the constraints?

[A] ∞
[B] $f(w)$
[C] $f(w) + \max_i\{g_i(w)\}$
[D] $f(w) + \max_j\{h_j(w)\}$

10. (Lecture 6 7) Same context as in the last question, what is $\max_{\alpha,\beta:\alpha_i\geq 0} \mathcal{L}(w, \alpha, \beta)$ if w satisfies all the inequality constraints but does not satisfy one of the equality constraints?

[A] ∞
[B] $f(w)$
[C] $f(w) + \max_i\{g_i(w)\}$
[D] $f(w) + \max_j\{h_j(w)\}$

11. (Lecture 6 7) Why are we doing all the things to derive the dual problem for SVM? What is main advantage?

[A] The dual problem objective is convex while the original objective is not.
[B] Solving the dual problem typically requires fewer computational resources than solving the primal, especially as the number of features increases.
[C] The dual formulation enables the direct calculation of the decision function without involving the input features explicitly, thus facilitating the use of kernel methods.
[D] The dual problem provides a more robust solution by enforcing stricter constraints on the optimization process.

12. (Lecture 8) Which of the following models is a generative model used in machine learning? (1) Linear Regression (2) Naive Bayes (3) Support Vector Machine (4) VAE (5) KNN (6) Hidden Markov Model

[A] (2)(3)(6)
[B] (2)(4)(6)
[C] (1)(3)(4)(5)
[D] None of above option is correct.

13. (Lecture 8) Generative models model $p(x)$ or $p(x, y)$, while discriminative models model $p(y|x)$ (suppose y is the label and x is the input)

[A] True.
[B] False.

14. (Lecture 8) Suppose x is our data, generative models allow you to compute $p(x_0)$ with close-form expression give x_0.

[A] True.

[B] False.

15. (Lecture 8) Suppose x is our data, generative models typically allow you to draw new samples of x.

    [A] True.

    [B] False.

16. (Lecture 9) In terms of MLE and MAP,

    [A] MLE tries to find $argmax_\theta \prod_{i=1}^{N} p(x^{(i)}|\theta)$, MAP tries to find $argmax_\theta \prod_{i=1}^{N} p(\theta|x^{(i)})$

    [B] MLE tries to find $argmax_\theta \prod_{i=1}^{N} p(\theta|x^{(i)})$, MAP tries to find $argmax_\theta \prod_{i=1}^{N} p(x^{(i)}|\theta)$

    [C] MLE and MAP are same when the prior distribution is Gaussian.

    [D] None of above option is correct.

17. (Lecture 10) In Bias-variance ,

    [A] we can increase model complexity or use more features to reduce variance

    [B] we can simplify the model to reduce model bias

    [C] Cross validation can help identify whether the model has a bias or variance problem.

    [D] All of above option are correct.

18. (Lecture 11) Which of the following is NOT true on kmeans clustering?

    [A] The kmeans algorithm does not guarantee to find the global optimus of its objective function

    [B] kmeans is sensitive to initialization of cluster centroids

    [C] Training the kmeans model on test examples is OK and not cheating in some cases.

    [D] If we work with a large number of cat and dog images and runs kmeans clustering on them, the trained model can directly predict "cat" or "dog" labels for the images.

19. (Lecture 11) Which of the following is NOT true on Gaussian Mixture Model (GMM) and Gaussian Discriminative Analysis (GDA)?

    [A] GMM is a generative model.

    [B] We cannot compute the closed-form equation for the likelihood of GMM

    [C] GDA is not a generative model.

    [D] GDA and GMM have the same graphical structure.

20. (Lecture 12) Which of the following is correct on the relationship between Expectation Maximization (EM) and MLE?

    [A] EM is a new parameter estimation approach that differs from MLE.

    [B] EM is an optimization approach to perform MLE for some models.

[C] EM can always optimizes the objective better than direct gradient descent.

[D] EM may not converge, which depends on the initialization.

21. (Lecture 12) Expectation Maximization monotonically increases ELBO.

    [A] True

    [B] False

22. (Lecture 12) Expectation Maximization monotonically increases the marginal data likelihood $p(x)$.

    [A] True

    [B] False

23. (Lecture 13) Which of the following is True for PCA?

    [A] Only keeping data projections onto principal components with non-zero eigenvalues will lose information of the original data

    [B] PCA is a nonlinear transformation of the original data

    [C] Eigenvalues denote the amount of variability along the corresponding dimension

    [D] All the above are incorrect

24. (Lecture 15) Which of the following PGM notation does not indicate $A \perp B|C$?

    [A] $A \rightarrow C \rightarrow B$

    [B] $A \leftarrow C \leftarrow B$

    [C] $A \leftarrow C \rightarrow B$

    [D] $A \rightarrow C \leftarrow B$

25. (Lecture 16 17) In Hidden Markov Models,

    [A] Since it contains latent variable, we can only do variational inference.

    [B] we can use either forward algorithm or backward algorithm to do evaluation problem (find the probability of observed sequence)

    [C] We cannot do EM on HMMs because of the intractable $p(x)$.

    [D] We can use some technique to find the closed-form likelihood of HMMs and then directly do closed-form MLE on it.

    [Lecture 18 19] Consider a convolutional layer which takes as input an RGB image, meaning that each pixel has three color channels. The input images are 32×32 pixels. The layer has three filters, each operating on windows of 4×4 pixels. Answer the following 4 questions.

26. (Lecture 18 19) What dimensions do the output of this layer have, if we choose a stride of 2 and apply l-pixel padding to the input?

[A] $15 \times 15 \times 3$
[B] $16 \times 16 \times 3$
[C] $7 \times 7 \times 3$
[D] $8 \times 8 \times 3$

27. (Lecture 18 19) How many trainable parameters does this layer have, if the filters do not have a bias term?

[A] 48
[B] 432
[C] 160
[D] 144

28. (Lecture 18 19) Let n denote the number of trainable parameters of the layer from last Question. We double the width and height of the input images, and change nothing else. How many trainable parameters would the adjusted layer have in terms of n?

[A] $2n$
[B] n
[C] $\sqrt{2}n$
[D] n^2

29. (Lecture 18 19) Let n denote the number of trainable parameters of the layer from last last Question. This time, we make the images grayscale (a single channel per pixel) and change nothing else. How many trainable parameters would the adjusted layer have in terms of n?

[A] $n/3$
[B] n
[C] $n/\sqrt{3}$
[D] $n^{1/3}$

30. (Lecture 20 Transformers) In self-attention, there are three basic steps: similarity calculation QK^T, softmax and weighted sum. Supposed n is sequence length, d is embedding dimension, what is the correct time complexity for each component.

[A] similarity calculation takes $O(n^2d)$, softmax takes $O(n)$, weighted sum takes $O(nd^2)$
[B] similarity calculation takes $O(n^2d)$, softmax takes $O(n^2)$, weighted sum takes $O(nd^2)$
[C] similarity calculation takes $O(n^2d)$, softmax takes $O(n)$, weighted sum takes $O(n^2d)$
[D] similarity calculation takes $O(n^2d)$, softmax takes $O(n^2)$, weighted sum takes $O(n^2d)$

31. (Lecture 20 Transformers) In self-attention, we often use scaled dot-product attention, which means $\frac{QK^T}{\sqrt{d_k}}$ and then pass it to softmax, what is correct about this $\sqrt{d_k}$

[A] We can replace $\sqrt{d_k}$ with d_k which makes no difference.

[B] Add a $\sqrt{d_k}$ here to reduce the magnitude of dot-product to avoid gradient vanishing in softmax.

[C] Actually there is no practical meaning for this d_k here.

[D] This $\sqrt{d_k}$ term is useful when d_k is relative small, but it become useless when the dimension becomes larger (i.e. d_k is large)

32. (Lecture 21 VAE) What is wrong about VAE, GMM and AE?

[A] VAE learns the distribution of latent variable z while AE does not, so VAE is able to generate new data.

[B] We can use Expectation maximization for GMM but not VAE.

[C] $P(z|x)$ are both intractable in GMM and VAE.

[D] GMM can be treated as a mixture of k Gaussian distribution and VAE can be treated as a mixture of infinite Gaussian distribution, so VAE is stronger.

33. (Lecture 21 VAE) In VAE, we need to propose a posterior distribution $q(z|x)$ to approximate the actual posterior $p(z|x)$, which of the following statements is true?

[A] Maximizing ELBO is equivalent to minimizing $KL(q(z|x)||p(z|x))$.

[B] Maximizing ELBO is equivalent to maximizing $KL(p(z|x)||q(z))$.

[C] It can be transformed to maximize $KL(q(z|x)||p(z))$.

[D] None of the above.

34. (Lecture 22 GAN) The generative model in GANs is the same as that in VAEs in terms of the graphical notation.

[A] True

[B] False

35. (Lecture 22 GAN) GAN training is MLE.

[A] True

[B] False

36. (Lecture 23) What is the typical order of SFT, RLHF, pretraining in LLM development?

[A] pretraining $\rightarrow$ RLHF $\rightarrow$ SFT

[B] SFT $\rightarrow$ pretraining $\rightarrow$ RLHF

[C] pretraining $\rightarrow$ SFT $\rightarrow$ RLHF

[D] RLHF $\rightarrow$ pretraining $\rightarrow$ SFT