MCGILL UNIVERSITY

WINTER SEMESTER, 2023
Campus: Montreal - Downtown

COMPUTER SCIENCE

Applied Machine Learning Course

NOTE: Practice Problem Set for Midterm

SECTION A: ML Fundamentals

This part checks your understanding of the basic concepts and algorithms in Machine Learning. Please use text, formula, images (created by yourself) or any combinations of them to answer the questions.

1. Explain each of the following models in a paragraph (use about 200 words or less). Discuss **when** they work (e.g. what kind of data this method is useful for, model complexity, if they can do classification, regression or both, etc.), **how** they work (explain the parameters and discuss what they are learning and how), **why** they work (discuss their inductive bias) and **what** needs to be considered specific to them to make them work (do you need to do data normalization, regularization, etc.).

(a) Nearest Neighbours
(b) Decision Trees
(c) Naive Bayes
(d) Linear Regression
(e) Logistic Regression
(f) Softmax Regression
(g) Multilayer Perceptron (MLP)
(h) Convolutional Neural Networks (CNN)

2. Compare the following models with regards to each other. Use about 50 words or less per comparison. Focus on their key differences and/or similarities in terms of data/task they can be applied to, model complexity and efficiency, loss function, etc.

(a) Linear Regression v.s. Logistic Regression
(b) Logistic Regression v.s. Softmax Regression
(c) Logistic Regression v.s. Naive Bayes Classifier
(d) Logistic Regression v.s. Multilayer Perceptron (MLP)
(e) Multilayer Perceptron (MLP) v.s. Convolutional Neural Networks (CNN)

CONTINUED

3. Explain each of the following concepts discussed in the course with few lines (use about or less than 50 words).

(a) Over-fitting and Under-fitting

(b) Bias and Variance trade-off

(c) Regularization

(d) Generalization

(e) Hyper-parameter

4. Explain the gradient decent approach in a short paragraph (use less than 100 words) and discuss what the Adam (Adaptive Moment Estimation) algorithm is doing to make it work better.

5. Explain Maximum Likelihood Estimation (MLE) in the context of fitting a model to the given data, and how it is different from Bayesian approach, discuss the MAP estimate and how it relates the two, explain if it result in lower or higher variance. (use about or less than 200 words)

SECTION B: ML Practitioner's Knowledge

This part checks your depth of understanding of the different concept with more specific questions.

6. When using gradient descent algorithm, are we guaranteed to find a local minimum? please explain.

7. Can we use gradient descent to solve a linear regression problem? and if so, could it result in multiple local optimum solutions?

8. How will the bias and variance of a trained model change with each of the following? e.g. answer less bias but more variance; or less bias but variance stays the same, etc.

(a) increasing the number of data points the model learns from

(b) increasing k in a k-nearest neighbour model

(c) pruning a decision tree

(d) increasing the regularization parameter (λ) in Ridge regression

(e) adding dropout to an MLP

(f) reducing the batch size when training an MLP with stochastic gradient descent

9. Regularization is more important for a model that have higher or lower expressiveness power? please discuss.

10. With back-propagation, can we learn the globally optimum solution for fully connected feed-forward network with one hidden layer (2 layers MLP)? please explain.

SECTION C: ML Innerworkings Knowledge

This part checks your depth of understanding of the different algorithms with more specific questions.

11. What is the prediction for $x = [1\ 0\ 1]$, when using a Gaussian Naive Bayes model that is trained using maximum likelihood on the following data:

$$x = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 2 & 1 \\ 0 & 1 & 0 \\ 4 & 1 & 1 \\ 6 & 1 & 3 \end{bmatrix}, \quad y = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

12. What is the prediction for input $x = [1\ 0\ 1]$, when using a Softmax regression model with the following weights:

$$w = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 0 & 0 \\ 0 & 2 & 1 \end{bmatrix}$$

13. What is the maximum likelihood estimate for the parameter w when using the following cost function and training data? please write the derivations.

$J(w) = \frac{1}{2}\sum_n (y^{(n)} - wx^{(n)})^2$

$\mathcal{D} = \{(1,1),(1,3),(2,1),(2,5),(2,6),(3,1),(3,8),(3,4),(5,10),(6,10)\}$

14. In the class, we discussed using a linear regression model ($\hat{y} = w^T x + b$) for binary classification, where you set the targets to $\{0,\ 1\}$, is not a good idea since the L2 loss ($L_2(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$) used by the linear regression model may penalize confident correct predictions. Can we fix this by using a modified hinge loss defined as:

$$L(y, \hat{y}) = \begin{cases} max(0, y) & \hat{y} = 0 \\ 1 - min(1, y) & \hat{y} = 1 \end{cases}$$

15. If we change the logistic regression model to use $\tilde{\sigma}(z) = \frac{e^{-z}}{1+e^{-z}}$ instead of the original sigmoid function, $\sigma(z) = \frac{1}{1+e^{-z}}$, and when trained using the same binary cross entropy loss, what would happen to the parameters and predictions of the model? how will they change? please explain and justify your answer.

16. What happens when you increase the momentum in Adam algorithm (i.e. increasing β_1)? What about β_2? Recall that in Adam we have:

$M^{\{t\}} \leftarrow \beta_1 M^{\{t-1\}} + (1 - \beta_1)\nabla J(w^{\{t-1\}})$

$S^{\{t\}} \leftarrow \beta_2 S^{\{t-1\}} + (1 - \beta_2)\nabla J(w^{\{t-1\}})^2$

$w^{\{t\}} \leftarrow w^{\{t-1\}} - \frac{\alpha}{\sqrt{\hat{S}^{\{t\}}+\epsilon}}\hat{M}^{\{t\}}$

17. Consider a two layered multi layered perceptron model given by $u = \sigma(W\, ReLu(Vx))$ where $\sigma(x) = (1 + e^{-x})^{-1}$ and $ReLu = max(0, x)$ when the loss function is set to $L = |\hat{y} - y|$. Consider learning the parameters of this model with stochastic gradient decent (SGD) with learning rate of 0.5, and assume $W^{\{t\}} = [-1 \quad 1]$ and $V^{\{t\}} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$. Now consider a training example with $x = [2 \quad -2]$ and $y = 0$, calculate $V^{\{t+1\}}$ and $W^{\{t+1\}}$.

18. Consider the following input and and convolution filter and when using zero padding of 1 and stride of 2, what is output of $y[1, 1]$ assuming indexing starts at zero?

$$x = \begin{bmatrix} 0 & 1 & 1 & 2 & 2 \\ 1 & 2 & 1 & 3 & 1 \\ 0 & 1 & 0 & 2 & 2 \\ 4 & 1 & 1 & 2 & 5 \\ 6 & 1 & 3 & 2 & 2 \end{bmatrix}, \quad w = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

SECTION D: ML Application Scenarios

This part checks your understanding of how to apply the techniques we discussed in the course. Please use text, formula, images (created by yourself) or any combinations of them to answer the questions.

19. Consider you are applying a linear regression model to estimate corn yield. The data contains many irrelevant feature or measurements as it is not known what actually impacts crop productivity. Will you use Lasso or Ridge regression? please explain and justify your answer.

20. Consider you have a spam detection model (spam = 1, not spam = 0) to filter the messages sent to you. You can control a parameter in the model to increase or decrease the recall. If you set the parameter to have higher recall, you expect to see more or less spam in your inbox? do you expect to see more or less actual emails being forwarded to your spam folder? please explain.

21. Assume you are working with a company that wants to design a fake news detection classifier for news articles. The company is providing you 1000 example articles they have manually labeled as fake or not fake and is asking you to develop a model for them using deep learning (e.g. MLP). Assume that they have a good feature extractor that converts a given text to a vector which serves as the input for your model. Please discuss and justify your answers to the following questions:

(a) how will you split this dataset to design/train/test your model?

(b) What would you do if after training the first model you try, your training loss is too high? could asking more data to be labeled help?

(c) What would you do if your training loss is low but your loss on validation set varies too much between different runs?

(d) How confident you would be on your trained model being able to detect fake news when deployed, assuming you are able to achieve low validation/test loss? please discuss.

22. Assume you are working in a hospital to make a system that helps doctors with cancer diagnosis. You have been given example data of 1000 patients with their different test measurements and diagnosis outcome (no cancer, brain cancer, lung cancer, etc.). Please discuss and justify your answers to the following questions:

(a) What classifier will you use for this task? explain, justify your choice.

(b) Considering some tests are more expensive to run (e.g. performing MRIs are more costly for the hospital compared to blood tests), which means some features are more costly to obtain for a new patient, how would you modify your model to be able to work with less features and ask for more only when needed?

By writing my name below, I confirm that all my exam work will be done entirely by myself, with no help from others. I will not provide any information about the exam's contents and or my solutions to other people until after March 30th 2022.

McGill ID No: _______________ Student name: ____________________________________

This exam is marked out of 180 points and contributes to 30% of your final grade. The grade breakdown is provided below. For multi-part questions, points are equally distributed.

SECTION A: 80 points in total

1. Models, 48 points [equally distributed]:

(a) Nearest Neighbours
(b) Decision Trees
(c) Naive Bayes
(d) Linear Regression
(e) Logistic Regression
(f) Softmax Regression
(g) Multilayer Perceptron (MLP)
(h) Convolutional Neural Networks (CNN)

2. Model Comparisons, 10 points:

(a) Linear Regression v.s. Logistic Regression
(b) Logistic Regression v.s. Softmax Regression
(c) Logistic Regression v.s. Naive Bayes Classifier
(d) Logistic Regression v.s. Multilayer Perceptron (MLP)
(e) Multilayer Perceptron (MLP) v.s. Convolutional Neural Networks (CNN)

3. Core Concepts, 10 points:

(a) Over-fitting and Under-fitting
(b) Regularization
(c) Bias and Variance trade-off
(d) Generalization
(e) Hyper-parameter

4. Gradient Decent, 6 points:

5. MLE & MAP, 6 points:

SECTION B: 20 points in total

6. Gradient descent guarantee, 2 points:

7. Gradient descent & linear regression, 2 points:

8. bias and variance, 12 points:

(a) increasing the number of data points

(b) increasing k in a k-nearest neighbour

(c) pruning a decision tree

(d) increasing λ in Ridge regression

(e) adding dropout to an MLP

(f) reducing the batch size in SGD-MLP

9. Regularization & expressiveness, 2 points:

10. Back-propagation, 2 points:

SECTION C: 45 points in total

11. Gaussian Naive Bayes prediction, 5 points:

12. Softmax regression prediction, 5 points:

13. Maximum likelihood estimate, 5 points:

14. Linear regression as classifier with modified loss, 5 points:

15. Logistic regression with modified activation, 5 points:

16. Adam parameters, 5 points:

17. MLP training, 10 points:

18. CNN layer, 5 points:

SECTION D: 35 points in total

19. Yield estimation, 5 points:

20. Spam Detection, 5 points:

21. Fake News Detection, 13 points:

(a) how will you split this dataset to design/train/test your model?

(b) What would you do if after training the first model you try, your training loss is too high? could asking more data to be labeled help?

(c) What would you do if your training loss is low but your loss on validation set varies too much between different runs?

(d) How confident you would be on your trained model being able to detect fake news when deployed, assuming you are able to achieve low validation/test loss?

22. Cancer diagnosis, 12 points:

(a) What classifier will you use for this task? explain, justify your choice.

(b) Considering some tests are more expensive to run (e.g. performing MRIs are more costly for the hospital compared to blood tests), which means some features are more costly to have for a new patient, how would you modify your model to be able to work with less features and ask for more if needed?