CM146: Introduction to Machine Learning **Winter 2023**

Final Exam

March 21, 2023

- This is an open book, open notes exam.
- Everything you need in order to solve the problems is supplied in the body of this exam.
- This exam has a total of **22 pages** including the cover sheet.
- Please WRITE YOUR NAME AND UID ON EACH PAGE OF THE EXAM.
- Electronic devices (laptops, tablets, phones, calculators) SHOULD BE TURNED OFF for the duration of the exam.
- Mark your answers ON THE EXAM ITSELF in the space provided.
- DO NOT put answers on the back of any page of the exam.
- DO NOT detach ANY pages.
- DO NOT attach ANY extra sheets.
- For true/false questions, CIRCLE True OR False.
- Unless otherwise instructed, for multiple-choice questions, CIRCLE ALL CORRECT CHOICES (in some cases, there may be more than one).
- If you think something about a question is open to interpretation, feel free to ask the instructor or make a note on the exam.
- Useful formulas and scratch space are provided at the end of the exam.
- You have **2 hours 45 minutes**.

Good Luck!

Legibly write your name and UID in the space provided below.

Name:

UID:

Name and UID		/1
True/False		/20
Multiple choice		/21
Evaluation		/6
Backprop		/6
Sleep prediction		/14
Kernelized logistic regression		/10
GMM		/12
Total		/90

1 True or False (20 pts)

Choose either True or False for each of the following statements.

1. (2 pts) In AdaBoost, a weak learner can make a negative contribution (β_t in round t) to the final classifier.

   True False

2. (2 pts) The K-means algorithm, run with multiple random initializations, will converge to the global minimum.

   True False

3. (2 pts) The principal components (PCs) computed in PCA will remain the same after multiplying each feature by a constant $c \neq 0$ (you can ignore the sign of the PCs so that $-\boldsymbol{p}$ is considered the same solution as $\boldsymbol{p}$).

   True False

4. (2 pts) You use k-fold cross-validation to tune the hyperparameters of your model. This requires training k models.

   True False

5. (2 pts) You perform multi-class classification for K classes using multinomial logistic regression. In this approach, a class C_k is predicted only if the predicted probability for class k is greater than 0.5.

   True False

6. (2 pts) In the Perceptron learning algorithm, the classifier incorrectly predicts example $\boldsymbol{x}_n$ with true label y_n. This results in the weights of the classifier being updated to $\boldsymbol{w}$. The new weights are guaranteed to correctly classify $\boldsymbol{x}_n$.

True False

7. (2 pts) Consider a binary classification problem with D input features. The VC dimension of the hypothesis space for a SVM is larger than that of a logistic regression (assume that both models work with the original features).

True False

8. (2 pts) Assume that the daily weather (whether it is sunny or raining for each day) follows a Markov process. Having observed the weather today, observations of last week's weather will not improve the accuracy with which you can predict the weather tomorrow.

True False

The next two questions refer to the a binary classification problem on the training dataset shown in Figure 1 where each training data point $x_n \in \mathbb{R}$ and each label $y_n \in \{-1, +1\}$ (black or filled circles refer to the +1 label).

Figure 1: Training data

9. (2 pts) You decide to use a support vector machine (SVM) to solve the classification problem. When using the features x_n as input to the SVM, a soft-margin SVM is a better choice for this problem than a hard-margin SVM.

True False

10. (2 pts) You decide to now apply k-Nearest Neighbors (k-NN) to this problem. k-NN with $k = 3$ will yield zero training error on this problem.

True False

2 Multiple choice (21 pts)

MARK ALL CORRECT CHOICES (in some cases, there may be more than one)

1. (3 pts) Which of the following statement(s) about the ID3 algorithm for learning decision trees are correct ?

   (a) The ID3 algorithm finds the decision tree with minimum depth that can correctly classify all training instances.
   (b) The ID3 algorithm can only be used for binary classification problems.
   (c) The decision boundary of the classifier learned by the ID3 algorithm can be non-linear.
   (d) The ID3 algorithm can only use categorical features.

2. (3 pts) Instead of minimizing the OLS cost function (residual sum of squares), you consider an alternate cost function $J(\theta_0, \theta_1) = \sum_{n=1}^{N} (y_n - (\theta_0 + \theta_1 x_n))$. Which of the following statements are **true** for this problem?

   (a) The minimizer of this cost function is the same as the OLS solution.
   (b) The minimizer of this cost function is the same as the solution to ridge regression.
   (c) The minimizer of this cost function is unbounded.
   (d) The cost function J is convex.

3. (3 pts) If the ridge regression model with regularization parameter $\lambda = 1$ is overfitting, what are the possible steps to reduce overfitting (select all that apply)?

   (a) Collect new data to increase the size of the training dataset
   (b) Increase the size of the training dataset by duplicating each training example.
   (c) Increase the value of λ.
   (d) Remove features from the model.

4. (3 pts) Denote the weight matrix of the first hidden layer of a neural network as $\boldsymbol{W}$, the design matrix as $\boldsymbol{X}$, and the value of the loss function with input $\boldsymbol{X}$ as l. If we were to replace the data matrix $\boldsymbol{X}$ with $\tilde{\boldsymbol{X}} = \frac{1}{2}\boldsymbol{X}$ and $\boldsymbol{W}$ with $\tilde{\boldsymbol{W}} = 2\boldsymbol{W}$ resulting in a loss $\tilde{l}$. Which of the following relationships holds ?

   (a) $\nabla_{\tilde{\boldsymbol{W}}} \tilde{l} = \frac{1}{2} \nabla_{\boldsymbol{W}} l$
   (b) $\nabla_{\tilde{\boldsymbol{W}}} \tilde{l} = \nabla_{\boldsymbol{W}} l$
   (c) $\nabla_{\tilde{\boldsymbol{W}}} \tilde{l} = 2 \nabla_{\boldsymbol{W}} l$
   (d) We cannot determine the relationship between $\nabla_{\tilde{\boldsymbol{W}}} \tilde{l}$ and $\nabla_{\boldsymbol{W}} l$ using the given information.

5. (3 pts) Which of the following loss functions is/are convex ?

   (a) 0/1 loss
   (b) Hinge loss
   (c) Logistic loss
   (d) Exponential loss

6. (3 pts) We want to deploy a machine learning algorithm on a website to make real-time predictions on what ad to show to a visitor to the website based on their view history. To learn this model, we have a training dataset of a million instances and 10 features. Which of the following learning models would be practical for this application ?

   (a) K-Nearest Neighbors
   (b) Logistic regression
   (c) Decision tree
   (d) Perceptron

7. (3 pts) Let $\lambda_1 > \lambda_2 > \ldots > \lambda_D$ be the eigenvalues of a sample covariance matrix $\boldsymbol{C}$ over D features. The solution to the optimization problem $min_{\boldsymbol{x}} \boldsymbol{x}^T \boldsymbol{C} \boldsymbol{x}$ subject to $||\boldsymbol{x}||_2 = 1$.

   (a) λ_1
   (b) λ_D
   (c) 0
   (d) ∞

3 Evaluation (6 pts)

Consider a binary classification dataset with $9,000$ positive and $1,000$ negative examples.

1. (2 pts) What is the recall of a classifier that classifies any example as positive?

2. (2 pts) What is the precision of a classifier that classifies any example as positive?

3. (1 pts) What is the area under the ROC curve (AUROC) of a classifier that classifies an example as positive with probability p and as negative with probability $1 - p$?

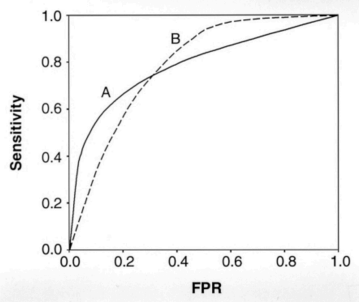

Figure 2: ROC

4. (1 pts) Given two classifiers A and B with ROC curves as shown in Figure 2. For your problem, you require a false negative rate of 10% or less from your classifier. Which classifier should you choose ?

4 Backprop (6 points)

Consider the following neural network which takes a scalar $x \in \mathbb{R}$ as input and computes a real-valued prediction $\hat{y} \in \mathbb{R}$: $a_1 = R(w_1 x_1), \hat{y} = w_2 a_1$. Here R is a differentiable non-linear function and w_1 and w_2 are scalar parameters of this neural network.

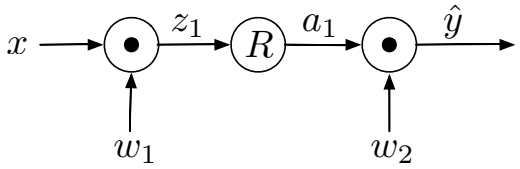

Figure 3: Computation graph

We represent this neural network using the computation graph in Figure 3. We will use the mean squared loss $L(w_1, w_2) = (y - \hat{y})^2$ to train this network.

Questions:

1. (3 pts) Find $\frac{\partial L}{\partial w_2}$. Express your answer in terms of y and any of the variables shown in the computation graph (you may also use R' that denotes the derivative of R).

2. (3 pts) Find $\frac{\partial L}{\partial w_1}$. Express your answer in terms of y and any of the variables shown in the computation graph (you may also use R' that denotes the derivative of R).

5 Sleep prediction (14 points)

You are performing a study to understand the factors that determine the quality of sleep. The quality of sleep in a given day n is summarized as a continuous variable $y_n \in \mathbb{R}$. You also measure D features that could affect the quality of sleep which you represent as a vector $\boldsymbol{x}_n \in \mathbb{R}^D$. You want to build a linear regression model to predict y from $\boldsymbol{x}$.

For each observation n, you assume the following linear model: $y_n = \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{x}_n + \epsilon_n$. Here ϵ_n is a random variable that adds noise to the linear relationship. We assume ϵ_n is drawn independently from a Normal or Gaussian distribution: $\epsilon_n \sim \mathcal{N}(0, \sigma_n^2)$.

You have collected training dataset across $2N$ days where the first N examples $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}$ were collected during summer while the remaining N examples, $\{(\boldsymbol{x}_n, y_n)\}_{n=N+1}^{2N}$ were collected during winter. You also notice that the variance of ϵ_n is twice as large during winter compared to the summer so that $\sigma_n^2 = \sigma^2$ for $n \in \{1, \ldots, N\}$ while $\sigma_n^2 = 2\sigma^2$ for $n \in \{N+1, \ldots, 2N\}$.

Questions:

1. (4 pts) Write the log-likelihood function for the full training dataset. Express the log-likelihood in terms of $\boldsymbol{\theta}$, σ^2, N, and $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^{2N}$. You may use C to represent constant terms that do not depend on $\boldsymbol{\theta}$, σ^2. Please completely expand the density of your normal distribution in your answer.

2. (3 pts) Show that finding the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ is the same as finding the value of $\boldsymbol{\theta}$ that minimizes the cost function:

$$J(\boldsymbol{\theta}) \;=\; A\sum_{n=1}^{N}\left(y_n - \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}_n\right)^2 + B\sum_{n=N+1}^{2N}\left(y_n - \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}_n\right)^2$$

Write down the values of A and B in this cost function.

3. (3 pts) Show that $J(\boldsymbol{\theta})$ can also be written as:

$$J(\boldsymbol{\theta}) \;=\; (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})$$

Here $\boldsymbol{W}$ is a $2N \times 2N$ diagonal matrix. Write the expression for $\boldsymbol{W}$ in terms of A and B from the previous question.

4. (4 pts) Show that the optimal value for $\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{y}$.

6 Kernelized logistic regression (10 pts)

In this problem, we explore how logistic regression can be kernelized.

We are given a set of N training examples, $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ where $\boldsymbol{x}_n \in \mathbb{R}^D$, $y_n \in \{0, 1\}$. We learn a logistic regression model $h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sigma(\boldsymbol{\theta}^T \boldsymbol{x})$ using gradient descent where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function.

In iteration t of gradient descent, we update $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \sum_n \epsilon_n \boldsymbol{x}_n$ where $\epsilon_n = h_{\boldsymbol{\theta}}(\boldsymbol{x}_n) - y_n$ is the error for the n^{th} training sample, and η is the step size or learning rate.

We map $\boldsymbol{x}$ to $\boldsymbol{\phi}(\boldsymbol{x})$ and we would like to learn a logistic regression model $\sigma(\boldsymbol{\theta}^T \boldsymbol{\phi}(\boldsymbol{x}))$ while only working with the inner products $\boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{\phi}(\boldsymbol{x}')$.

1. (4 pts) Assume we initialize $\boldsymbol{\theta}$ to zero in the gradient descent algorithm, *i.e.*, $\boldsymbol{\theta} \leftarrow \mathbf{0}$. Show that at the end of every iteration of gradient descent, $\boldsymbol{\theta}$ is always a linear combination of the training samples: $\boldsymbol{\theta} = \sum_{n=1}^{N} \alpha_n \boldsymbol{\phi}(\boldsymbol{x}_n)$.

2. (4 pts) Using the above result, show how we can write $h_{\boldsymbol{\theta}}(\boldsymbol{x})$ to make a prediction on a new input $\boldsymbol{\phi}(\boldsymbol{x})$ by only using inner products of the form $\boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{\phi}(\boldsymbol{x}')$.

3. (2 pts) The final step in kernelization is to show that we do not need to explicitly store $\boldsymbol{\theta}$. Instead from part (a), we can implicitly update $\boldsymbol{\theta}$ by updating α_n. Show how α_n is initialized and how it is updated.

7 GMMs (12 pts)

You observe the prices of N different bikes without knowing the type of each bike. Each bike could be a road bike or a mountain bike. You decide to model bike prices as a Gaussian mixture model (GMM) with two components ($K = 2$) where one component corresponds to mountain bikes and the other to road bikes. The distribution of prices for a given bike in this model is given by: $p(x|\pi_0, \pi_1, \mu_0, \mu_1, \sigma^2) = \pi_0 \mathcal{N}(x; \mu_0, \sigma^2) + \pi_1 \mathcal{N}(x; \mu_1, \sigma^2)$. Here π_0 and π_1 denote the proportions of road and mountain bikes. μ_k denotes the mean price of a road bike (for $k = 0$) and a mountain bike (for $k = 1$). You assume that the variance of the price is the same for both types of bikes. You observe N i.i.d. samples of bike prices from this GMM: $x_1, \ldots, x_N$.

1. (1 pts) What are constraints on the parameters (π_0, π_1) for this to be a valid probability distribution?

2. (2 pts) Write down the expression for the (incomplete) log likelihood for the full training data: $\mathcal{LL}(\pi_0, \pi_1, \mu_0, \mu_1, \sigma^2)$ (you should expand out the density function of the normal distribution).

3. (1 pts) Since there is no closed-form expression for the values of the parameters that maximize the log likelihood (the MLE), we will use the EM algorithm. Assume we have a cluster membership variable z_n for each data point, $z_n \in \{0, 1\}$ (this variable indicates the bike type for a given bike). Write the probability of $P(z_n = 0)$ as a function of the parameters of the GMM.

4. (1 pts) Write the probability of $P(x_n|z_n = 0)$ as a function of the parameters of the GMM (you should expand out the density function of the normal distribution).

5. (2 pts) Assume that we have an indicator variable γ_{nk} for each data point n and mixture component k. $\gamma_{nk} = 1$ if $z_n = k$ and $\gamma_{nk} = 0$ otherwise. The joint density of the bike prices and the bike type, $\{(x_n, z_n)\}_{n=1}^N$, is the complete likelihood. Fill in the blanks in the derivation of the complete log-likelihood of this dataset is given by:

$$\begin{aligned}
\mathcal{LL}_c(\pi_0, \pi_1, \mu_0, \mu_1, \sigma^2) &= \log P(\{(x_n, z_n)\}_{n=1}^N | (\pi_0, \pi_1, \mu_0, \mu_1, \sigma^2)) \\
&= \sum_{n=1}^{N} \log P(x_n, z_n | (\pi_0, \pi_1, \mu_0, \mu_1, \sigma^2)) \\
&= \sum_{n=1}^{N} \log \prod_{k \in \{0,1\}} P(x_n, z = k | (\pi_0, \pi_1, \mu_0, \mu_1, \sigma^2))^A \\
&= \sum_{n=1}^{N} \sum_{k \in \{0,1\}} A \log P(x_n, z = k | (\pi_0, \pi_1, \mu_0, \mu_1, \sigma^2)) \\
&= \sum_{n=1}^{N} \sum_{k \in \{0,1\}} A[\log P(z = k) + \log P(x_n | z = k)] \\
&= \sum_{n=1}^{N} \sum_{k \in \{0,1\}} A[B + C]
\end{aligned}$$

Write down the expressions for A, B, C in terms of γ_{nk}, x_n, π_k, μ_k, and σ^2.

6. (4 pts) What values of μ_0, μ_1, σ^2 maximize $\mathcal{LL}_c$. This is the M(maximization)-step of the EM algorithm. Express your answer in terms of γ_{nk} and x_n.

7. (1 pts) State in words how you would extend your approach to deal with the case that γ_{nk} is not observed. This is the E(Expectation)-step of the EM algorithm.

Identities

Probability density/mass functions for some distributions

$$\begin{aligned}
\text{Normal} \quad &: \quad P(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\
\text{Multinomial} \quad &: \quad P(\boldsymbol{x};\boldsymbol{\pi}) = \prod_{k=1}^{K} {\pi_k}^{x_k} \\
&\quad \boldsymbol{x} \text{ is a length } K \text{ vector with exactly one entry equal to 1} \\
&\quad \text{and all other entries equal to 0} \\
\text{Poisson} \quad &: \quad P(x;\lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}
\end{aligned}$$

Matrix calculus

Here $\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{b} \in \mathbb{R}^n, \boldsymbol{A} \in \mathbb{R}^{n \times n}$. $\boldsymbol{A}$ is symmetric.

$$\begin{aligned}
\nabla \boldsymbol{x}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{x} &= 2\boldsymbol{A}\boldsymbol{x} \\
\nabla \boldsymbol{b}^{\mathrm{T}} \boldsymbol{x} &= \boldsymbol{b}
\end{aligned}$$

You may use this page for scratch space.

You may use this page for scratch space.