50.007 Machine Learning
2016 Term 6
Midterm

Date. 04 Nov 2016
Time. 2:30 pm
Duration. 2 hours
Total. 50 points

Instructions to Candidates

1. There are 7 questions with 5 printed pages.
   (This title page counts as the first page.)
2. This is a closed book examination.
3. Cheat sheets are not allowed.
4. Answer all the questions.
5. Write your answers in the answer books provided.
6. Do not turn over the title page until you are told.
7. All the best!

Q1. Classification and Regression [5x3=15pt]

For each of the machine learning techniques listed in the table below, fill in the corresponding predictor, learning objective/cost and learning algorithm using the options listed below. Note that each of the options could be used more than once, and each of the cells in the table could contain more than one option. In the answer booklet, you may write your answer in the form, "(1a) P8, C8, A7, A8".

You may assume that the training data is a set of n pairs (x, y) where $x \in \mathbb{R}^d$ is a feature vector and y is either a signed label $y \in \{-1,1\}$ or a real-valued response $y \in \mathbb{R}$, depending on whether the problem-of-interest is classification or regression. The model parameters are $\theta \in \mathbb{R}^d$ and $\theta_0 \in \mathbb{R}$.

	Technique	**Predictor**	**Learning Cost**	**Learning Algorithm**
(1a)	Ridge Regression			
(1b)	Linear Classification using Hinge Loss			
(1c)	Linear Regression			
(1d)	Perceptron (with Offset)			
(1e)	Support Vector Machine with Slack Variables			

Predictor (i.e. Classifier/Regression Function)

P1. $f(x; \theta, \theta_0) = \theta^\top x + \theta_0$
P2. $h(x; \theta, \theta_0) = \text{sign}(\theta^\top x + \theta_0)$

Learning Objective/Cost Function

C1. $\mathcal{L}_n(\theta, \theta_0) = \frac{1}{n} \sum_{\text{data } (x,y)} \frac{1}{2}(y - (\theta^\top x + \theta_0))^2$
C2. $\mathcal{L}_n(\theta, \theta_0) = \frac{1}{n} \sum_{\text{data } (x,y)} \frac{1}{2}(y - (\theta^\top x + \theta_0))^2 + \frac{\lambda}{2}\|\theta\|^2$
C3. $\mathcal{L}_n(\theta, \theta_0) = \frac{1}{n} \sum_{\text{data } (x,y)} \max\{1 - y(\theta^\top x + \theta_0), 0\}$
C4. $\mathcal{L}_n(\theta, \theta_0) = \frac{1}{n} \sum_{\text{data } (x,y)} \max\{1 - y(\theta^\top x + \theta_0), 0\} + \frac{\lambda}{2}\|\theta\|^2$
C5. $\mathcal{L}_n(\theta, \theta_0) = \frac{1}{n} \sum_{\text{data } (x,y)} [\![y(\theta^\top x + \theta_0) \leq 0]\!]$, where $[\![\cdot]\!]$ is the indicator function

Learning Algorithm

A1. Exact Solution
A2. Mistake-Driven Updates
A3. Stochastic (Sub-)Gradient Descent

Q2. Clustering [2+2+2+4=10pt]

The k-means algorithm iteratively computes the set of centroids given a clustering of the data points, and the clustering of the data points given a set of centroids. In this question, you will provide formulas for the iterative steps of the k-means algorithm.

Let the data points be $x^{(1)}, x^{(2)}, \dots, x^{(n)} \in \mathbb{R}^d$.

Let the clusters be subsets $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_k \in \{1, 2, \dots, n\}$ of the indices.

Let the centroids be d-dimensional vectors $z^{(1)}, \dots, z^{(k)} \in \mathbb{R}^d$.

(2a) Suppose that you are given the clusters $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_k \in \{1, 2, \dots, n\}$. Write down the formula for each of the centroids $z^{(1)}, \dots, z^{(k)} \in \mathbb{R}^d$.

(2b) Suppose that you are given the centroids $z^{(1)}, \dots, z^{(k)} \in \mathbb{R}^d$. Write down the quantity that we need to minimize to find the cluster $\mathcal{C}_j$ for a particular data point $x^{(i)}$.

(2c) The cost function in the k-means algorithm is not convex, so it could have local minima that give rise to poor clustering. Briefly describe one strategy for overcoming this issue.

(2d) To find the optimal number k of clusters, a method called *validation* is often used. Describe the steps involved in validation. In particular, state the performance metric used for computing the validation error in k-means clustering.

Q3. Collaborative Filtering [2+4=6pt]

(3a) Name two algorithms that can be used for collaborative filtering.

(3b) Which of the following problems are well-suited for collaborative filtering?

	Problem	**Well-suited**
(3bi)	Predicting missing values in a matrix of sensor readings from a town, where the columns correspond to sensors and the rows correspond to timestamps	Yes/No
(3bii)	Predicting the current water level in a reservoir, given recent rainfall data and historical records of water measurements in the reservoir	Yes/No
(3biii)	Predicting the books that a user would want to read in a library, given historical records of the loans of all the users	Yes/No
(3biv)	Predicting the sentiment for a named object in a new tweet, given a large data set of annotated tweets that may not contain the named object	Yes/No

Q4. Support Vector Machines [4+1=5pt]

(4a) Determine if the following statements about support vector machines (SVMs) are true or false.

(4ai)	If a data point (x, y) is a support vector, then the corresponding multiplier $\alpha_{x,y}$ must be equal to zero.	True/False
(4aii)	The margin of the SVM classifier is given by $\frac{1}{2}\|\theta\|^2$.	True/False
(4aiii)	If the hyperparameter λ in the objective function $$\frac{1}{n}\sum_{\text{data }(x,y)} \max\{1 - y(\theta^\top x + \theta_0),\ 0\} + \frac{\lambda}{2}\|\theta\|^2$$ decreases, then the margin of the classifier increases.	True/False
(4aiv)	Computing the SVM classifier with slack variables involves solving a convex optimization problem.	True/False

(4b) Which of the follow kernel functions should be used for polynomial classification?

a. $K(x, x') = x \cdot x'$
b. $K(x, x') = (x \cdot x')^k + 1$
c. $K(x, x') = (x \cdot x' + 1)^k$
d. $K(x, x') = \exp(-\|x - x'\|^2/2)$

5. Deep Learning [1+4 = 5pt]

(5a) Write down the name of the training algorithm that is based on the chain rule.

(5b) In deep learning, because the learning objective function is highly non-convex, one major issue with gradient descent is getting stuck in a bad local minimum. This issue can be alleviated by carefully initializing the parameters. Describe one strategy for doing this initialization.

6. Generative Methods [5+1+3=9pt]

(6a) The PDF of a Poisson distribution with the real-valued parameter $\lambda \geq 0$ is

$$P(x|\lambda) = \frac{e^{-\lambda}\,\lambda^x}{x!}$$

where $x \geq 0$ is an integer, $x! = 1 \cdot 2 \cdot \cdots \cdot x$ and $0! = 1$. Suppose that the training data consists of independent and identically distributed samples $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$. Derive the maximum likelihood estimate (MLE) of λ given this data set. (Hint: use the log likelihood.)

(6b) Consider the expectation-maximization (EM) algorithm for the mixture of spherical Gaussians. Write down the strategy used to initialize the means $\mu^{(1)}, \ldots, \mu^{(k)}$.

(6c) This question is more challenging. Suppose we have a mixture of multinomial distributions whose PDF is given by

$$P(x|p,q) = \sum_{i=1}^{k} p_i P\left(x\middle|q^{(i)}\right) = \sum_{i=1}^{k} p_i q_x^{(i)}$$

where $x \in \{1, 2, \ldots, d\}$, $p = (p_1, \ldots, p_k) \in \mathbb{R}^k$ and $q^{(1)}, q^{(2)}, \ldots, q^{(k)} \in \mathbb{R}^d$. Here, the entries of each vector $p, q^{(1)}, q^{(2)}, \ldots, q^{(k)}$ are non-negative and sum to one. Write down the formula for the soft labels $p(i|x)$ computed during the expectation step of the EM algorithm.

7. Extra Credit

Your score to this question, if you attempt it, will be given by

$$\max\{\, 4 - 4x_1^2 - 8\alpha_1(1 + x_2) - 8\alpha_2(1 - 4x_1 - x_2) - 8\alpha_3(1 - 5x_1 - x_2),\ \ 1\,\}$$

However, you are only allowed to choose the value of $x = (x_1, x_2) \in \mathbb{R}^2$. I will be choosing the value of $\alpha = (\alpha_1, \alpha_2, \alpha_3) \in \mathbb{R}^3$ where each $\alpha_i \geq 0$. Write down your choice for x.

END OF PAPER