CSCI-GA-2565: Machine Learning (Fall 2024)

Midterm Exam (4:55pm–6:55pm, Oct 22)

- Answer the questions in the spaces provided. If you run out of room for an answer, use the blank page at the end of the test.
- Do not turn the front page until the instructor has signaled the beginning of the exam.
- Stop writing immediately after the time is up, or otherwise your exam will not be collected.
- Do not forget to write your name and student ID below.

Name: ______________________________

NYU NetID: ______________________________

Question	Points	Score
Introduction	4	
Optimization	5	
Regularization	7	
SVM	6	
Probabilistic ML	8	
Total:	30	

1. **Introduction.**

   Supposed that you have chosen a large hypothesis space H_1 that is more flexible in modeling different functions than a rather fixed hypothesis space H_2.

   (a) (1 point) You will likely need

   A. A smaller number of training examples

   B. A larger number of training examples

   C. Equal number of training examples

   (b) (1 point) You will likely have

   A. A higher approximation error

   B. A higher estimation error

   C. A higher optimization error

   D. None of the above

   E. All of the above

   (c) (1 point) Describe a real-world application where regression is useful.

   (d) (1 point) Describe a real-world application where classification is useful.

2. **Optimization.**

(a) (2 points) In what situation would stochastic gradient descent be a good choice for optimization compared to full batch gradient descent? Select all correct choice(s).

A. When there is no closed form solution.

B. When there is a lot of variance in the training distribution.

C. When the training set is too large to fit in the memory.

D. When the example vectors have a large norm.

Explain you answer:

(b) (1 point) Why do you need to decrease your step size for stochastic gradient descent?

(c) (2 points) Which of the following statements is/are true? Select all correct choice(s).

A. Early stopping reduces optimization error and therefore is often applied in practice.

B. Subgradients may not exist for non-convex functions, but pointwise non-differentiability is usually tolerable in practice.

C. Stochastic gradient descent converges at the same rate as gradient descent.

D. Gradient of f points at the direction where the function decreases the fastest.

3. **Regularization.**

(a) (2 points) Write the following optimization into Ivanov form:

$$w^* = \arg\min_w \frac{1}{2}\|Xw - y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2. \tag{1}$$

(b) (1 point) What is the role of λ? What happens to the weight vector when λ grows bigger towards ∞?

(c) (1 point) How do you choose the best λ?

(d) (3 points) Show that the ridge regression estimate is the mean (and mode) of the posterior distribution, under a Gaussian prior $\mathbf{w} \sim \mathcal{N}(0, \tau I)$, and a model $\hat{y} \sim \mathcal{N}(\mathbf{x}^\top \mathbf{w}, \sigma^2)$. Write the regularization parameter λ in terms of the variances τ and σ^2.

4. **Support Vector Machines.** Recall the (soft-margin) SVM primal problem:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i \\ \text{subject to} \quad & -\xi_i \leq 0 \quad \text{for } i = 1, \dots, n \\ & \left(1 - y_i \left[w^T \varphi(x_i) + b\right]\right) - \xi_i \leq 0 \quad \text{for } i = 1, \dots, n \end{aligned}$$

and its dual problem:

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \varphi(x_j)^T \varphi(x_i) \\ \text{subject to} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0 \\ & \alpha_i \in \left[0, \frac{c}{n}\right] \quad \text{for } i = 1, \dots, n \end{aligned}$$

(a) (1 point) When is it more advantageous to optimize the dual objective than the primal objective? Select all conditions that are relevant:

A. When the inner product computation can be simplified.

B. When you have more examples violating the margin constraint.

C. When the dual variables are sparse.

D. When you have high dimensional features but relatively less data points ($d \gg n$)

E. When you apply a polynomial kernel.

(b) (1 point) When $\alpha_i = \frac{c}{n}$, which of the following conditions can occur? Select all correct choice(s).

A. Margin > 1

B. Margin $= 1$

C. $0 \leq$ Margin < 1

D. Margin < 0

(c) (1 point) When $\xi_i > 0$, which of the following conditions can occur? Select all correct choice(s).

A. Margin > 1

B. Margin $= 1$

C. $0 \leq$ Margin < 1

D. Margin < 0

(d) (1 point) Explain the meaning of (functional) margin. What happens when the margin is positive or negative?

(e) (1 point) Explain why when the data is linearly separable, then $\xi_i = 0$ for all i.

(f) (1 point) Which of the following statements about optimization are true? Select all correct choice(s).

A. The dual objective is always convex, even when the primal objective is not convex.

B. Complementary slackness means that either the Lagrange multiplier is zero for the constraint or the constraint function f_i is evaluated to be zero.

C. The primal objective is always greater than or equal to the dual objective.

D. The number of dual variables is always equal to the primal variables.

5. **Probabilistic ML.** Imagine you are the instructor of a college class and you have assigned a homework essay. Some students have used chat bots in producing the writing, while others have not. You wish to use the writings as training data to produce a machine learning model that can predict whether an essay was written by a chat bot. One simple strategy is to look at the word choices that are used by chat bots vs. humans.

   (a) (1 point) If you only look at the words individually and treat them independently, what kind of assumption are you making and why?

   (b) (2 points) Let $y = 1$ denote that the essay is generated by a chat bot and $y = 0$ human. Now let's suppose that the probability of generating word j in an essay, given a label $y = \{1, 0\}$ follows a Bernoulli distribution, i.e. the probabilities are $\theta_{j,0}$ and $\theta_{j,1}$. Given the assumption above, write down the probability $p(x, y)$ of entire dataset of size N in terms of $\theta_{j,0}, \theta_{j,1}$ and $p(y_i)$.

   (c) (3 points) Apply the maximum likelihood principle and derive the optimal value for $\theta_{j,1}$.

(d) (1 point) Explain what is the intuition behind the optimal value? What does it represent?

(e) (1 point) If you follow the Bayesian principle of adding a "prior" distribution on θ_j, which of the following situation can a prior on θ help improve the performance? Select all correct choice(s).

A. When the assumption in part a) does not hold.

B. When there is a lack of training samples to estimate θ_j.

C. When word j often appear in both categories of essays.

D. When the count of word j is zero.

Congratulations! You have reached the end of the exam.

You can use the additional space below.

You can use the additional space below.