CSCI-GA-2565: Machine Learning (Fall 2023)

Midterm Exam (4:55pm–6:55pm, Oct 24)

- 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
Generalization	4	
Optimization	6	
Regularization	5	
SVM	7	
Kernels	4	
MLE	4	
Total:	30	

1. **Generalization.**

   H_1 and H_2 are two hypothesis space and $H_1 \subset H_2$.

   (a) (1 point) Give an example of H_1 and H_2:

   > **Solution:** H_1 is the family of linear functions and H_2 is the family of all polynomials.

   (b) (1 point) Which one is likely to have a higher approximation error?

   > **Solution:** H_1 because it has less functions and cannot approximate non-linear functions well.

   (c) (1 point) Which one is likely to have a higher estimation error?

   > **Solution:** H_2 because it has more parameters and requires more samples to estimate.

   (d) (1 point) In empirical risk minimization, by increasing your number of training samples from 10 to 1000, which of the following error(s) is/are likely to decrease?

   A. optimization error
   B. estimation error
   C. approximation error
   D. all of A, B, and C
   E. none of A, B, and C

2. **Optimization.**

(a) (2 points) When using stochastic (mini-batch) gradient descent, a larger batch size means you should use

A. Smaller step size

B. Larger step size

C. Same step size

Explain you answer:

> **Solution:** Larger batch size means that the variance is smaller and you should use a smaller step size.

(b) (2 points) Which of the following statement is true?

A. Gradient descent may not converge to the global minimum in convex problems, even with a properly chosen step size.

B. The advantage of SGD vs. gradient descent is that SGD can be faster to compute.

C. Gradient descent require a decreasing step size schedule in order to converge.

D. Early stopping refers to stop training when the model's training loss starts to go up.

Explain you answer:

> **Solution:** A is not correct because gradient descent can converge. B is correct because SGD requires less time to compute gradients. C is not correct because gradient descent does not require decreasing step size, but SGD does. D is not correct because early stopping is when the validation loss starts to go up.

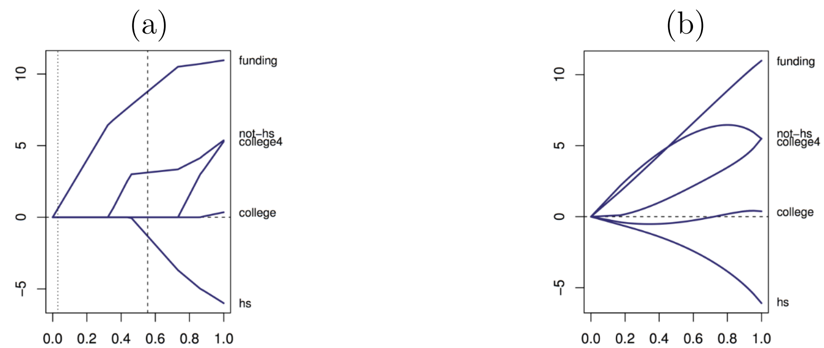
(c) (2 points)

$$L(\theta) = \frac{1}{2}\exp((\boldsymbol{x}^\top \boldsymbol{\theta} - y)^2) - 1$$

Derive the gradient of $\frac{\partial L}{\partial \boldsymbol{\theta}}$:

> **Solution:**
>
> $$\frac{\partial L}{\partial \theta} = \exp((\boldsymbol{x}^\top \boldsymbol{\theta} - y)^2)(\boldsymbol{x}^\top \boldsymbol{\theta} - y)\boldsymbol{x}$$

3. **Regularization.** Compare the following regularization path plots:



(a) (2 points) Which plot corresponds to the lasso and which correponds to ridge regression? Explain your answer.

> **Solution:** a) is Lasso and b) is ridge. Lasso creates piecewise linear optimization path and encourages the weights to zero.

(b) (2 points) Which of the following is a reason to prefer L1 regularization over L2 regularization? Select all correct choices.

A. L1 can be solved with coordinate descent.

B. L1 encourage the weights closer to zero.

C. L1 encourage sparse feature selection, and improve the model's interpretability.

D. L1 is more robust to outliers.

E. L1 can be solve with subgradient descent.

(c) (1 point) Explain how L2 regularization controls the size of the hypothesis space.

> **Solution:** L2 regularization can be converted from Tikhonov to Ivanov form with a hard constrait on the weight norm.

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) What is the meaning of minimizing $||w||^2$? Select all choices that are correct:

   A. maximizing the margin

   B. L2 regularization

   C. balancing the classes

   D. minimizing the margin violation

(b) (1 point) What is the meaning of the constraint $\sum_{i=1}^{n} \alpha_i y_i = 0$? Select all choices that are correct:

   A. maximizing the margin

   B. L2 regularization

   C. balancing the classes

   D. minimizing the margin violation

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

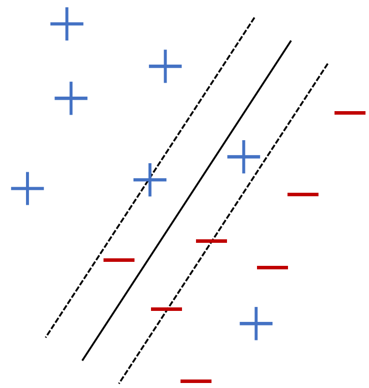
   A. When you have more examples violating the margin constraint.

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

   C. When you apply a linear or polynomial kernel.

   D. When the inner product computation can be simplified.

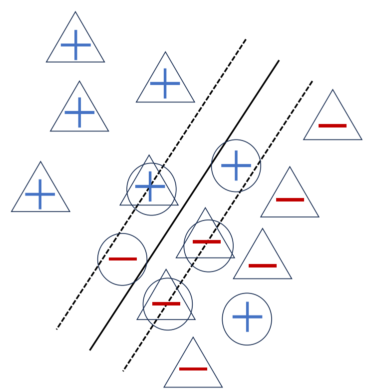
(d) Recall that given the dual solution α_i^*'s, the primal solution is given by $w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$, and the support vectors are defined to be x_i's where $\alpha_i > 0$. The figure below shows a toy dataset and the SVM decision boundary (the solid line) with the corresponding margin (indicated by the two dotted lines).



i. (1 point) Draw a triangle around all points that have the slack variable likely to be zero ($\xi_i = 0$).

ii. (1 point) Draw a circle around all points that are likely support vectors ($\alpha > 0$).

Solution:



(e) Assuming the data is not linearly separable, increasing c is likely to result in (circle the right answer)

i. (1 point) smaller / larger geometric margin

Solution: smaller

ii. (1 point) fewer / more support vectors

Solution: fewer

5. **Kernels.**

Suppose that the dataset is $\{\boldsymbol{x}_i, y_i\}$, and you have a ridge regression model $y = \boldsymbol{w}^\top \boldsymbol{x}$.

(a) (1 point) What does it mean for the solution w^* to be in the span of the data?

> **Solution:** $\boldsymbol{w}^* = \sum_i \alpha_i \boldsymbol{x}_i = X\boldsymbol{\alpha}$.

(b) (1 point) Write the learning objective $L(\boldsymbol{w})$ for ridge regression when you apply feature transformation φ on your input $\boldsymbol{x}$. Let λ be your coefficient of the regularizer.

> **Solution:** $L(\boldsymbol{w}) = \frac{1}{2}||y - \varphi(X)^\top \boldsymbol{w}||^2 + \lambda||\boldsymbol{w}||^2$.

(c) (1 point) Now incorporate part a) and rewrite the objective by using $\varphi(\boldsymbol{x})$, y, and λ only.

> **Solution:** $L(\boldsymbol{w}) = \frac{1}{2}||y - \varphi(X)^\top \varphi X\boldsymbol{\alpha}||^2 + \lambda||\varphi(X)\boldsymbol{\alpha}||^2$.

(d) (1 point) Note that for a given vector $\boldsymbol{a}$, $||\boldsymbol{a}||^2 = \boldsymbol{a}^\top \boldsymbol{a}$, write the learning objective as a function of the kernel matrix K.

> **Solution:**
>
> $$\begin{aligned} L(\boldsymbol{w}) &= \frac{1}{2}||y - K\boldsymbol{\alpha}||^2 + \lambda(\varphi(X)\boldsymbol{\alpha})^\top(\varphi(X)\boldsymbol{\alpha}) \\ &= \frac{1}{2}||y - K\boldsymbol{\alpha}||^2 + \lambda\boldsymbol{\alpha}^\top K\boldsymbol{\alpha}. \end{aligned}$$

6. **MLE.** You are at a casino. On each round of the game, a machine generates a real number $x \in \mathcal{R}$. If the number is positive, you wins x dollars. If the number is negative, you must pay the casino x dollars. So far, you have played 3 times and observed the following dataset:

$$\mathcal{D} = \{-5, 3, -10\}$$

Angela believes the machine is generating its numbers from a normal distribution with mean μ and variance 10:

$$x \sim \mathcal{N}(\mu, 10)$$

For this question, you may find the probability density function of the normal distribution useful:

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

(a) (2 points) Write the log-likelihood function $\ell(\mu) = \log p(\mathcal{D}|\mu)$.

Solution:

$$\begin{aligned}
\log p(\mathcal{D}|\mu) &= \log\{\mathcal{N}(-5|\mu, 10) \times \mathcal{N}(3|\mu, 10) \times \mathcal{N}(-10|\mu, 10)\} \\
&= \log \mathcal{N}(-5|\mu, 10) + \log \mathcal{N}(3|\mu, 10) + \log \mathcal{N}(-10|\mu, 10) \\
&= \left(-\frac{1}{2}\log(20\pi) - \frac{1}{20}(-5-\mu)^2\right) + \left(-\frac{1}{2}\log(20\pi) - \frac{1}{20}(3-\mu)^2\right) + \\
&\quad \left(-\frac{1}{2}\log(20\pi) - \frac{1}{20}(-10-\mu)^2\right) \\
&= -\frac{3}{2}\log(20\pi) - \frac{1}{20}\left((-5-\mu)^2 + (3-\mu)^2 + (-10-\mu)^2\right)
\end{aligned}$$

(b) (2 points) You believe that the casino will make you lose money in the long run. This belief is a prior distribution on μ: $p(\mu) = \mathcal{N}(\mu| -1, 5)$. Find the maximum a posteriori (MAP) estimate of the mean μ under this prior distribution.

Solution: The MAP solution μ_{MAP} satisfies:

$$\mu_{\text{MAP}} = \arg\max_{\mu} \log p(\mu) + \log p(\mathcal{D}|\mu)$$

We have:

$$\begin{aligned}
&\log p(\mu) + \log p(\mathcal{D}|\mu) \\
=& -\frac{1}{10}(\mu+1)^2 - \frac{1}{20}\left((-5-\mu)^2 + (3-\mu)^2 + (-10-\mu)^2\right) + \text{const.}
\end{aligned}$$

Taking the derivative with respect to μ:

$$\begin{aligned}
&\frac{\partial}{\partial \mu}\left\{\log p(\mu) + \log p(\mathcal{D}|\mu)\right\} \\
=& -\frac{1}{5}(\mu+1) + \frac{1}{10}\left((-5-\mu) + (3-\mu) + (-10-\mu)\right) \\
=& \frac{1}{5}(-\mu-1) + \frac{1}{10}\left(-12-3\mu)\right) \\
=& \frac{2}{10}(-\mu-1) + \frac{1}{10}\left(-12-3\mu)\right) \\
=& \frac{1}{10}(-2\mu-2) + \frac{1}{10}\left(-12-3\mu)\right)
\end{aligned}$$

We set the derivative equal to zero and solve for μ:

$$\begin{aligned}
\frac{1}{10}(-2\mu-2) + \frac{1}{10}\left(-12-3\mu)\right) = 0 &\implies -2\mu - 2 - 12 - 3\mu = 0 \\
&\implies -5\mu - 14 = 0 \\
&\implies \mu = -\frac{14}{5}
\end{aligned}$$

Congratulations! You have reached the end of the exam.

You can use the additional space below.

You can use the additional space below.